基于多尺度Transformer网络的多种生理信号情绪识别

Tu Vu†, Van Thong Huynh†, Soo-Hyung Kim全南大学人工智能融合系韩国光州{tu369,vthuynh,shkim}@jnu.ac.kr

arXiv:2305.00769v2 [eess.SP] 8 May 2023

摘要——**本文提出一种高效的基于多尺度Transformer的方法，用于从生理数据中进行情绪识别。由于现代传感器和机器学习技术能够从这些信号中提取大量信息，这一研究方向已受到学界广泛关注。我们采用多模态技术并结合数据缩放，以建立身体内部信号与人类情绪之间的关系。此外，我们运用Transformer和高斯变换技术，提升信号编码效果和整体性能。我们的模型在EPiC竞赛的CASE数据集上取得了不错的结果，均方根误差得分为1.45**。

索引词——**生理信号、深度学习、Transformer、多尺度**

I. 引言

情绪识别是人类交流的一个基本方面，准确检测情绪状态的能力对从医疗保健到人机交互等各种应用都有着重要影响。情绪通常会反映在生理信号 [1]，面部表情 [2]，和语音 [3]中。近年来，由于生理信号有望实时提供情绪状态的客观测量 [4]，利用生理信号进行情感计算受到了广泛关注。

近年来，人们越来越关注利用生理信号 [5]–[9]开发用于情感计算的机器学习算法。这些算法可用于对情绪状态进行分类、预测情绪状态随时间的变化，或识别生理信号中最有助于检测情绪状态的特定特征。人们也开始关注开发可穿戴传感器，以便在工作场所或社交场合等真实环境中采集生理信号 [10]。

将端到端深度学习架构用于生理信号，有望简化情绪识别系统的开发与部署 [4]。通过省去预处理步骤，这些架构能够降低系统开发的复杂度和所需时间，同时提高系统的可扩展性和准确性。此外，端到端架构还能支持开发可同时处理多种生理信号的系统，例如心率、呼吸和皮肤电活动，从而更全面、更准确地测量情绪状态。

尽管端到端深度学习架构在情感计算方面具有潜在优势，但仍有一些挑战亟待解决。其中一个挑战是开发能够处理含噪且非平稳生理信号的架构，因为这些信号可能受到运动伪影、信号漂移及其他噪声源的影响。另一个挑战是确保学习特征具有可解释性和意义，这有助于提升系统的透明度和可解释性。

本文提出了一种用于基于生理信号进行连续情绪回归的端到端多尺度架构。我们使用 EPiC 2023竞赛基准评估所提架构的性能，该基准包含在实验室环境下开展实验所收集的数据。

II. 相关工作

A. 基于多模态生理信号的连续情绪识别

生理信号被广泛认为是情感科学和情感计算中最可靠的数据形式之一。尽管个体能够操控面部表情或言语等外在信号，但有意识地控制自身内部状态却相当困难。因此，分析人体信号是全面识别和理解个体情绪状态的一种可靠且稳健的方法 [1], [11]。这种可靠性在医疗应用中尤为重要，例如心理健康治疗或精神疾病诊断。

从生理数据中识别情感仍是一项重大挑战，不仅在数据采集过程中如此，在情绪评估方面也是如此。由于能够控制实验变量，基于实验室的研究在情感科学领域占据主导地位。研究人员可以

本研究得到韩国政府（MSIT）资助的韩国国家研究基金会（NRF）项目（NRF- 2020R1A4A1019191）支持。本研究还得到韩国政府（MSIT）资助的信息通信技术规划与评估研究院（IITP）项目（No.2021-0-02068, 人工智能创新中心）支持。（通讯作者：*Soo-Hyung Kim*。）

†贡献相同。

精心挑选并准备情绪刺激材料，并使用各种传感器设备追踪和记录受试者的情绪状态，同时尽量减少意外事件和干扰 [4]。然而，这些研究大多依赖测验、问卷或离散情绪类别等间接的离散方法进行情绪评估，忽视了人类情绪体验随时间变化的特性。Sharma 等人 [12]提出了一种新颖的基于操纵杆的情绪报告界面（JERI），以克服情绪评估方面的一项局限。JERI 支持同时标注效价和唤醒度，从而实现对情绪的逐时评估。通过 JERI 采集的连续标注情绪信号（CASE）数据集，为研究人员识别情绪触发因素出现的时间提供了更多信息。借助该数据集，情绪生理与体验协作（EPiC）研讨会举办了一项竞赛，旨在根据生理信号对效价和唤醒度进行逐时连续评分建模。该竞赛为评估情绪识别模型的精确度和时间动态特性提供了宝贵的平台。因此，我们采用 EPiC竞赛基准作为评估所提出情绪识别模型的主要基准。

此外，有研究指出，单一生理信号相对难以准确反映人类的情绪变化。因此，近年来，许多研究开始关注利用多模态生理信号检测人类情绪。这些研究使用了多种生理信号。有些研究记录心脏相关信号，例如心电图（ECG） [13]–[15],血容量脉搏（BVP） [16], [17], ；另一些则使用大脑的电活动（脑电图/EEG） [18], [19]或肌肉的电反应（肌电图/EMG） [20],[21]。此外，有些研究还采用皮肤温度（SKT）[21], 、皮肤汗腺活动（EDA） [18], [20]以及呼吸的深度和频率（呼吸/RSP） [20]。基于 EPiC 竞赛提供的信号模态，我们的方法使用了 8 种不同类型的信号，包括 ECG、BVP、EMG CORU（皱眉肌）、EMG TRAP（斜方肌）、EMG ZYGO（颧大肌）、GRS、RSP 和 SKT。

B. 基于 *Transformer* 的多模态生理信号情绪识别方法

与其他涉及物理信号的情绪识别问题类似，生理数据中的情感计算已广泛采用机器学习技术，尤其是深度学习方法。在 Dominguez 等人的工作中 [5] ，他们采用了多种传统机器学习技术，包括高斯朴素贝叶斯、k 近邻和支持向量机，用于效价-唤醒度估计。然而，这些方法高度依赖于手工特征选择和特征提取过程的质量。为克服这一挑战，其他研究 [6]–[8] 提出采用深度学习技术实现端到端方法，使模型能够自动学习提取特征，而无需预先设计特征描述符。

随着深度学习的发展，各种最先进的技术已被用于基于生理信号的情感分析。Santamaria 等人 [7] 使用带有一维卷积层的卷积神经网络（CNN）进行情绪检测，而 Harper 等人 [8] 将 CNN 与常用的循环神经网络（RNN）相结合，用于基于心电图信号的情绪识别。自 2016年以来，Transformer [22]已成为深度学习领域首选的模型。Transformer 在自然语言处理中的稳健表现展现了其应用于时间序列信号的潜力；自然语言处理所处理的数据与时间序列数据具有一些相似特征。因此，近年来时间序列领域的研究已将 Transformer 用作模型架构的核心模块 [23]–[25]。针对生理信号，一些研究提出使用 Transformer 及其变体进行情绪检测 [6],[26]–[28]。Vazquez 等人的工作 [6], [26], 专注于将预训练 Transformer 应用于多模态信号处理。然而，这仍然是对 Transformer 模块非常基础的应用。Wu 等人 [27] 和 Yang 等人 [28] 提出使用更先进的基于 Transformer 的模型技术，包括用于单模态和多模态信号处理的自监督 Transformer 以及卷积增强型 Transformer。尽管这些研究已证明 Transformer 对生理信号的有效性，但它们通常将固定原始尺寸的信号输入模型，这可能导致全局特征信息的丢失。为解决这一问题，我们提出了一种新的多尺度 Transformer 架构，用于多模态情绪识别。

III. 所提出的方法

A. 问题定义

多模态生理信号情绪识别问题的输入是从人类受试者在诱发情绪的刺激过程中提取的 8 种生理信号，即 ECG、血容量脉搏、肌电图 CORU、肌电图 TRAP、肌电图 ZYGO、GRS、呼吸和皮肤温度。这些信号构成长度为 L 的 8 个序列。在情感计算领域，情绪识别问题的目标因所采用的情绪模型而异。本研究采用 CASE 数据集的 SAM（自我评估人偶） [29] 模型，因此问题的目标是估计效价-唤醒度（V-A）值。V-A 得分由两个连续浮点数构成，取值范围为 0.5 至 9.5。0.5 表示效价最负或唤醒度最低，5 表示效价或唤醒度处于中性水平，9.5 表示效价最正或唤醒度最高。

B. 方法论

我们构建了一种新型多尺度架构，用于根据 8 路生理信号估计效价和唤醒度。该架构由两个核心模块组成：特征编码模块和多尺度融合模块。整体架构如图 1 所示。

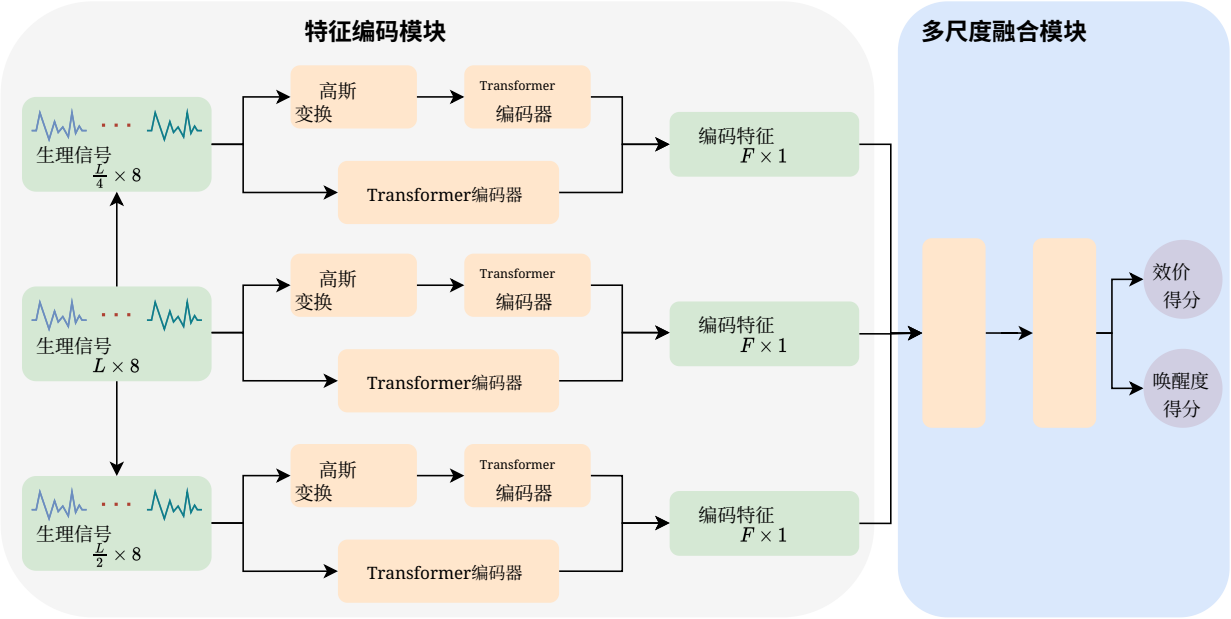


图1. 所提出架构的概览。

1）特征编码：为使特征编码模块能够为估计器提取全局特征，并消除输入信号中的噪声和干扰信息，我们采用一维平均池化将 8 路输入信号缩放为三种不同长度： L、$L/2$ 和 $L/4$。这一过程有助于增强模型提取有用信息以及消除不必要噪声和干扰的能力。

接下来，我们同时采用两种特征编码器：高斯变换 [30] 和Transformer编码器[22]。Transformer编码器模块以多头自注意力作为其核心机制。给定输入序列信号 $S \in R^{L \times C}$，其中 L 表示信号序列的长度，$C = 8$ 表示通道（信号模态）的数量，我们应用位置编码和嵌入层，将原始输入转换为标记序列。随后，将这些标记输入由多头自注意力（MSA） [22]、层归一化（LN）和多层感知机（MLP）模块组成的Transformer层。各个元素形式化表示如下：

$$y^i = MSA(LN(x^i)) + x^i \tag{1}$$

$$x^{l+i} = MLP(LN(y^i)) + y^i \tag{2}$$

这里，i 表示标记的索引，x^i 表示生成的特征标记。值得注意的是，由于多头自注意力机制允许并行处理多个序列，因此所有8个信号通道都会同时输入Transformer编码器。

高斯变换 [30] 传统上用于通过非线性变换输入特征，将线性模型核化：先通过单层进行变换，再基于变换后的特征训练线性模型。然而，在深度学习架构中，随机特征也可以加以利用，因为它们能够通过随机投影实现降维或近似某些函数。作为一种非参数技术，这种变换会将输入数据映射为更紧凑的表示，排除噪声信息，同时仍能实现计算高效的处理。这种技术可作为Transformer编码器架构的有益补充，弥补可能缺失的信息。

2）多尺度融合：我们将从特征编码器模块提取的不同尺度特征通过拼接操作进行融合。随后，将拼接后的特征输入一系列全连接层（FCN），以估计效价和唤醒度这两个得分。选择修正线性单元（ReLU）作为激活函数，是因为它能够引入非线性，从而提高分数估计的准确性。这种方法的优势在于能够高效估计所需得分，同时保持架构简洁明了。

IV. 实验与结果

A. 数据集

在情绪生理学与体验协作（EPiC）研讨会之后，我们使用连续标注情绪信号（CASE）数据集[12]评估我们的模型。该数据集包含来自多个生理传感器的数据以及情绪连续标注。这些数据采集自30名受试者，他们在观看多个视频刺激的同时，使用JERI报告自己的情绪体验。所使用的设备包括用于记录心电图（ECG）、血容量脉搏（BVP）、皮肤电反应（GSR）、呼吸（RSP）、皮肤温度（SKT）和肌电图（EMG）的传感器。这些传感器返回8种

表 I 场景级测试数据结果。

场景类型	唤醒度		效价	
	RMSE	STD	RMSE	STD
跨时间场景	1.503	1.083	1.639	1.2246
跨受试者场景	1.336	0.2276	1.345	0.0576
跨诱发因素场景	1.509	0.6771	1.514	0.6054
跨版本场景	1.369	0.2525	1.352	0.1919

生理信号：ECG、BVP、EMG CORU、EMG TRAP、EMG ZYGO、GRS、RSP和SKT。情绪刺激由11段视频组成，时长范围为120至197秒。情绪标注和生理数据的采样率分别为 20 Hz和1000 Hz。初始效价-唤醒度得分范围设定为 [-26225, 26225]。

在 EPiC竞赛中，组织者重新整理数据，并将标注值缩放至 [0.5, 9.5] 范围。重构后的数据集包含 4 种场景，对应 4 种不同的评估方法：跨时间场景、跨受试者场景、跨诱发因素场景和跨版本场景。

- 跨时间场景：每个样本代表一个人观看一段视频，训练集和测试集按时间划分。具体而言，视频前段用于训练，后段则留作测试。
- 跨受试者场景：参与者被随机分配到不同组别，同一组中的所有样本根据折的划分归入训练集或测试集。
- 跨诱发因素场景：在唤醒度-效价空间的每个象限中，每位受试者都有两个样本（视频）。每一折都会排除与某个象限相关的两个样本，因此共分为四折，每折排除一个象限。
- 跨版本场景：在唤醒度-效价空间的每个象限中，每位受试者都有两个样本。在此场景下，一个样本用于训练模型，另一个样本用于测试，最终得到两折。

B. 实验设置

我们的网络使用 TensorFlow 框架实现。我们使用 AdamW 优化器 [31]训练模型，学习率设为 0.001，并使用余弦退火热重启调度器 [32] 训练 10 个训练轮次。我们使用均方误差损失函数优化网络，并使用均方根误差作为评估阶段的指标。序列长度设为 2048。Transformer编码器包含 4 个 Transformer层，每个注意力模块包含 4 个注意力头。Transformer 的隐藏维度设为 1024。所有训练和测试过程均在 GTX 3090 GPU 上进行。

C. 结果

表 I 展示了我们的模型在测试集上的场景级评估结果。总体而言，最终的均方根误差在效价和唤醒度估计任务中，我们获得的得分均为 1.45。我们的模型在跨受试者场景中表现最佳，唤醒度得分为 1.336，效价得分为 1.345，二者的标准差相对较低，分别为 0.2276 和 0.0576。这些结果表明，我们的模型能够有效泛化至新的受试者，并在完整观看视频后准确捕捉情绪变化。与此同时，在跨诱发因素场景中，模型表现相对较差，唤醒度和效价得分分别为 1.509 和 1.514。这表明，基于先前学到的特定情绪状态，模型无法很好地推断训练期间未见过的情绪状态。

V. 结论

本文提出了一种用于生理信号多模态情绪识别的新型多尺度架构。我们的方法通过在多个尺度上使用 Transformer编码器对信号进行编码，以捕获全局和局部特征，并获得信息更丰富的表征。我们的方法在 EPiC竞赛的测试数据上取得了不错的结果。

伦理影响声明

本研究采用深度学习方法，根据生理信号进行情绪估计，因此涉及潜在偏差和泛化能力方面的伦理影响。用于模型训练的数据采集自特定人群，这可能引入偏差，并限制模型对其他人群的泛化能力。个体之间的生理信号可能存在差异，从而影响情绪估计的准确性，进一步降低泛化能力。我们承认这些局限性，并建议未来研究解决这些问题，制定伦理准则，以确保模型得到恰当使用。

参考文献

[1] L. Shu、J. Xie、M. Yang、Z. Li、Z. Li、D. Liao、X. Xu 和 X. Yang，《使用生理信号进行情绪识别的综述》，*Sensors*，第 18 卷，第 7 期，文章编号 2074，2018年。[2] S. Li 和 W. Deng，《深度面部表情识别：综述》，*IEEE*情感计算汇刊，第 13 卷，第 3 期，第 1195–1215 页，2020年。[3] B. W. Schuller，《语音情绪识别：二十年回顾、基准与发展趋势》，*ACM* 通讯，第 61 卷，第 5 期，第 90–99 页，2018年。[4] S. Saganowski、B. Perz、A. Polak 和 P. Kazienko，《利用可穿戴设备的生理信号进行日常情绪识别：系统性文献综述》，*IEEE*情感计算汇刊，2022年。[5] J. A. Dom´ınguez-Jim´enez、K. C. Campo-Landines、J. C. Mart´ınez- Santos、E. J. Delahoz 和 S. H. Contreras-Ortiz，《一种基于机器学习的生理信号情绪识别模型》，《生物医学信号处理与控制》，第 55 卷，文章编号 101646，2020年。[6] J. Vazquez-Rodriguez、G. Lefebvre、J. Cumin 和 J. L. Crowley，《使用预训练变换器和多模态信号进行情绪识别》，载于*2022*年第*10*届国际情感计算与智能交互会议（*ACII*），第 1–8 页，IEEE，2022年。[7] L. Santamaria-Granados、M. Munoz-Organero、G. Ramirez-Gonzalez、E. Abdulhay 和 N. Arunkumar，《利用深度卷积神经网络在生理信号数据集（AMIGOS）上进行情绪检测》，*IEEE Access*，第 7 卷，第 57–67 页，2018年。

[8] R. Harper 和 J. Southern，《一种基于贝叶斯深度学习、从心跳端到端预测情绪的框架》，IEEE情感计算汇刊，第 13 卷，第 2 期，第 985–991 页，2020年。[9] M. Algarni、F. Saeed、T. Al-Hadhrami、F. Ghabban 和 M. Al-Sarem，《一种利用脑电图（EEG）信号进行情绪识别的深度学习方法：基于双向长短期记忆网络（双向长短期记忆网络）》，Sensors，第 22 卷，第 8 期，文章编号 2976，2022年。[10] S. Saganowski、M. Behnke、J. Komoszy´nska、D. Kunc、B. Perz 和 P. Kazienko，《一种利用可穿戴设备在日常生活中采集带有情绪标注的生理信号的系统》，载于2021年第9届国际情感计算与智能交互研讨会及演示（ACIIW），第 1–3 页，IEEE，2021年。[11] Z. Ahmad 和 N. Khan，《基于生理信号的情绪识别综述》，《生物工程》，第 9 卷，第 11 期，文章编号 688，2022年。[12] K. Sharma、C. Castellini、E. L. van den Broek、A. Albu-Schaeffer 和 F. Schwenker，《一个包含连续情感标注和生理信号、用于情绪分析的数据集》，《科学数据》，第 6 卷，第 1 期，文章编号 196，2019年。[13] L. Hu、J. Yang、M. Chen、Y. Qian 和 J. J. Rodrigues，《SCAI-SVSC：用于可持续生命体征采集的智能服装，实现有效交互》，《未来一代计算机系统》，第 86 卷，第 329–338 页，2018年。[14] K. Rattanyu、M. Ohkura 和 M. Mizukawa，《在生活空间中利用生理信号为服务机器人进行情绪监测》，载于 ICCAS 2010，第 580–583 页，IEEE，2010年。[15] K. Rattanyu 和 M. Mizukawa，《在智能空间中利用生物信号进行情绪识别》，载于《人机交互：面向移动和智能交互环境》，第14届国际会议 HCI International 2011，2011年，9 14，586 592，2011年，美国佛罗里达州奥兰多，7月，论文集第三部分，第 – 页，Springer。[16] B. Zhao、Z. Wang、Z. Yu 和 B. Guo，《EmotionSense：基于可穿戴腕带的情绪识别》，载于2018年 IEEE SmartWorld、普适智能与计算、高级与可信计算、可扩展计算与通信、云计算与大数据计算、人联网及智慧城市创新大会（SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI），第 346–355 页，IEEE，2018年。[17] M. Ragot、N. Martin、S. Em、N. Pallamin 和 J.-M. Diverrez，《利用生理信号进行情绪识别：实验室传感器与可穿戴传感器的比较》，载于《可穿戴技术与游戏设计中的人因进展：AHFE 2017人因与可穿戴技术进展国际会议论文集》，2017年7月17日至21日，美国加利福尼亚州洛杉矶威斯汀博纳旺蒂尔酒店，第 8 卷，第 15–22 页，Springer，2018年。[18] T. Martens、M. Niemann 和 U. Dick，《情感倾向的传感器测量》，《心理学前沿》，第 11 卷，文章编号 379，2020年。[19] B. Nakisa、M. N. Rastgoo、A. Rakotonirainy、F. Maire 和 V. Chandran，《基于神经网络的情绪识别框架中的长短期记忆网络超参数优化》，IEEE Access，第 6 卷，第 49325–49338 页，2018年。[20] P. Schmidt、R. D¨urichen、A. Reiss、K. Van Laerhoven 和 T. Pl¨otz，《真实环境中的多目标情感检测：一项探索性研究》，载于2019年 ACM可穿戴计算机国际研讨会论文集，第 211–219 页，2019年。[21] L. Romeo、A. Cavallo、L. Pepa、N. Bianchi-Berthouze 和 M. Pontil，《利用生理信号进行情绪识别的多示例学习》，IEEE情感计算汇刊，第 13 卷，第 1 期，第 389–407 页，2019年。[22] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，《注意力就是你所需要的一切》，《神经信息处理系统进展》，第 30 卷，2017年。[23] S. Li、X. Jin、Y. Xuan、X. Zhou、W. Chen、Y.-X. Wang 和 X. Yan，《增强Transformer的局部性并打破其在时间序列预测中的内存瓶颈》，《神经信息处理系统进展》，第 32 卷，2019年。[24] N. Wu、B. Green、X. Ben 和 S. O'Banion，《用于时间序列预测的深度Transformer模型：流感患病率案例》，arXiv预印本 arXiv:2001.08317，2020年。[25] Y. Li、X. Peng、J. Zhang、Z. Li 和 M. Wen，《DCT-GAN：基于扩张卷积Transformer的时间序列异常检测生成对抗网络》，IEEE知识与数据工程汇刊，2021年。[26] J. Vazquez-Rodriguez、G. Lefebvre、J. Cumin 和 J. L. Crowley，《基于Transformer的情绪识别自监督学习》，载于2022年第26届国际模式识别会议（ICPR），第 2605–2612 页，IEEE，2022年。[27] Y. Wu、M. Daoudi 和 A. Amad，《基于Transformer的自监督多模态表征学习：面向可穿戴情绪识别》，*IEEE*情感计算汇刊，2023年。[28] K. Yang、B. Tag、Y. Gu、C. Wang、T. Dingler、G. Wadley 和 J. Goncalves，《利用卷积增强型Transformer和多种生理信号进行移动情绪识别》，载于*2022*年570，2022 562国际多媒体检索会议论文集，第 – 页。[29] M. M. Bradley 和 P. J. Lang，《情绪测量：自我评估人形量表与语义差异量表》，《行为疗法与实验精神病学杂志》，第25卷第1期，第49–59页，1994年。[30] A. Rahimi 和 B. Recht，《用于大规模核机器的随机特征》，《神经信息处理系统进展》，第20卷，2007年。[31] I. Loshchilov 和 F. Hutter，《解耦权重衰减正则化》，*arXiv*预印本 *arXiv:1711.05101*，2017年。[32] I. Loshchilov 和 F. Hutter，《SGDR：采用热重启的随机梯度下降》，*arXiv*预印本 *arXiv:1608.03983*，2016年。