arXiv:2305.03343v1 [cs.CV] 5 May 2023

LOGO-FORMER：用于动态面部表情识别的局部-全局时空Transformer

Fuyan Ma、*Bin Sun*、*Shutao Li*

学院　湖南大学电气与信息工程学院，中国长沙　a

摘要

以往针对野外动态面部表情识别（DFER）的方法主要基于卷积神经网络（CNN），其局部运算忽略了视频中的长程依赖关系。基于Transformer的方法能够取得更好的DFER性能，但会带来更多浮点运算次数和计算开销。为解决这些问题，本文提出局部-全局时空Transformer（LOGO-Former），在平衡复杂度的同时，捕捉每帧中的判别性特征并建模帧间的上下文关系。基于面部肌肉局部运动、面部表情逐渐变化的先验知识，我们首先将空间注意力和时间注意力都限制在局部窗口内，以捕捉特征标记之间的局部交互。此外，我们通过迭代地使用来自各局部窗口的特征查询一个词元，实现全局注意力，从而获取整个视频序列的长程信息。此外，我们提出紧凑损失正则项，进一步促使学习到的特征具有最小的类内距离和最大的类间距离。在两个野外动态面部表情数据集（即DFEW和FERV39K）上的实验表明，我们的方法为利用空间和时间依赖关系进行DFER提供了一种有效途径。

索引词— 动态面部表情识别，Transformer，时空依赖，损失正则化

1. 引言

面部表情识别（FER）一直是一个新兴课题，这得益于其在驾驶员安全监控、人机情感交互、老年人医疗保健等领域的重要现实应用。此前已有大量研究（如 [1,2]）针对静态面部表情识别（SFER）展开。然而，面部表情是一个动态过程，由不同面部区域中的各种面部肌肉运动组成。

本研究得到中国国家自然科学基金（62171183、62221002）和湖南省自然科学基金（2022JJ20017）的资助，并部分得到CAAI-Huawei MindSpore 开放基金的赞助。通讯作者：Bin Sun。

以往的动态面部表情识别（DFER）方法主要分为两类（即基于静态帧的方法和基于动态序列的方法）[3]。大多数基于静态帧的方法利用 Gabor 小波 [4]和卷积特征[5, 6] 选择视频中的峰值（顶点）帧，并进一步对这些帧进行面部表情识别。尽管这些方法通过选择峰值帧取得了良好效果，但它们忽略了面部帧之间的时间动态变化和相关性。与基于静态帧的方法不同，基于动态序列的方法通常使用 3D 卷积神经网络（3DCNN）[7], 和长短期记忆网络（LSTM） [8]学习时空关系，从而对长期依赖关系进行建模并提升 DFER 的性能。由于遮挡、头部姿态变化、光照不佳以及真实场景中的其他意外情况，这些方法的性能仍远未令人满意。

近年来，Transformer 在计算机视觉任务中的蓬勃发展，极大加深了我们对判别性特征表示和上下文信息建模的理解。例如，Li 等人 [9] 利用双向 Transformer 捕获帧间的时间信息。将标准 Transformer[10, 11] 直接扩展到动态面部表情识别，需要在所有 S 空间位置和 T 时间位置上联合执行多头自注意力。也就是说，完整的时空注意力复杂度为 $O(T^2S^2)$，这使得标准 Transformer 框架在高效动态面部表情识别方面面临沉重的计算负担。降低完整时空注意力计算成本的一种简单方法，是先计算仅针对空间的注意力，再进行时间平均，其复杂度为 $O(TS^2)$。最近，人们提出了分解式时空注意力 [12] 和时空混合注意力 [13]，与完整时空注意力相比，两者都能显著降低计算开销，分别将复杂度降至 $O(TS^2 + T^2S)$和 $O(TS^2)$。

不同的面部肌肉在局部面部区域内活动，而视频中相邻帧之间的面部表情会逐渐变化。我们的目标是在利用视频中时空信息的同时，尽可能降低 Transformer 的计算成本，从而高效地进行动态面部表情识别。为此，我们提出了局部-全局时空Transformer（LOGO-Former）

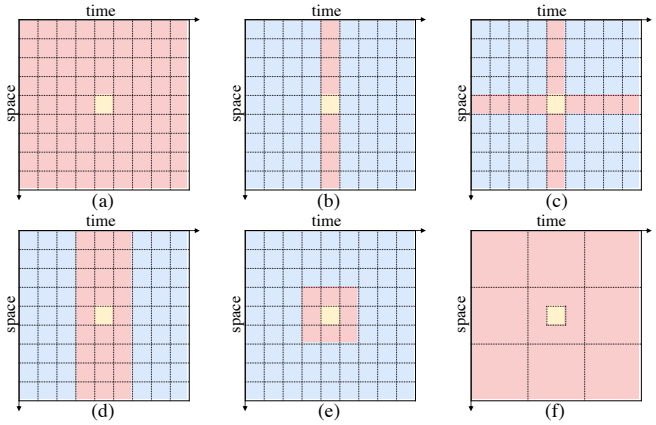

图1。不同类型时空自注意力方案的可视化示意。为了便于说明，我们用浅黄色表示查询标记，用玫瑰色表示键词元，用浅蓝色表示非注意力词元。(a) 完整时空注意力 [10]: $O(T^2S^2)$。(b) 仅空间注意力: $O(TS^2)$。(c) 分离式时空注意力 [12]: $O(TS^2+T^2S)$。(d) 时空混合注意力[13]: $O(TS^2)$。(e) 和 (f): 我们提出的局部和全局时空注意力。

，用于捕捉短程和长程依赖，同时降低 Transformer 的计算成本。我们在非重叠窗口内计算自注意力，以捕捉词元之间的局部交互。这种局部时空注意力无法捕捉全局信息。因此，我们采用全局时空注意力，使查询标记关注窗口级词元，如图 1 所示。为进一步提升模型的判别能力，我们提出紧凑损失正则化项，以缩小类内距离并增大类间距离。定量结果和可视化结果证明了我们的方法在自然场景动态面部表情识别中的有效性。

2. 所提出的方法

2.1. 输入嵌入生成

所提出的框架如图 2 所示，由三个部分组成，即输入嵌入生成、LOGO-Former 和紧凑损失监督。给定一个图像序列$\boldsymbol{X} \in \mathbb{R}^{F\times H_0\times W_0\times 3}$，其中包含从视频中采样得到的 F 个大小为 $H_0 \times W_0$ 的 RGB面部帧，我们利用 CNN骨干网络提取帧级特征。采用标准的 CNN骨干网络（例如 ResNet18）为每一帧生成大小为 $H \times W$ 的高级特征图。通过拼接所有帧的特征图，得到片段级特征 $\boldsymbol{f}_0 \in \mathbb{R}^{F\times H\times W\times C}$。

随后，我们将视频片段级特征图 $\boldsymbol{f}_0$ 的空间维度展平，并通过 1×1 卷积对其进行投影，得到新的特征序列 $\boldsymbol{f}_1 \in \mathbb{R}^{F\times (H\times W)\times d}$。

需要指出的是， $\boldsymbol{f}_1$ 的时间顺序与输入 X一致。为补充特征序列的时空位置信息，我们将可学习位置嵌入与 $\boldsymbol{f}_1$ 结合。我们还在时间维度上将分类标记 $[CLS]$ 添加到序列开头，用于建模序列的全局状态，并进一步用于识别。同样，也添加了时间位置嵌入。最终得到时空 Transformer的输入嵌入 $\boldsymbol{X}^0$ 。

2.2. LOGO-Former

LOGO-Former 由 N 个模块组成，每个模块由多头局部注意力和多头全局注意力构成，通过迭代学习上下文判别性时空特征表示。

多头局部注意力：鉴于相邻词元之间的关联强于非局部词元之间的关联，我们在互不重叠的窗口内执行多头局部注意力，以建模这些词元之间的局部交互。

以大小为 $F\times H\times W$ 的输入特征图作为输入（为简洁起见，此处省略 $[CLS]$ 词元），我们将其均匀划分为多个大小为 $f\times hw$ 的窗口，得到 $\frac{F}{f}\times\frac{HW}{hw}$ 个窗口，如图 1 所示。我们将窗口内的这些词元展平（ $i\ j$ $\boldsymbol{X}_{i\,j}\in$, ），$\mathbb{R}^{(fhw)\times d}$ 记作 , 。第 - 个模块的多头局部注意力可表示为 k

$$\boldsymbol{Y}^k = \boldsymbol{X}_{i,j}^{k-1} + \mathrm{MSA}(\mathrm{LN}(\boldsymbol{X}_{i,j}^{k-1})), \tag{1}$$

其中，LN 和 MSA 分别表示层归一化和多头自注意力。每个注意力模块之后都采用残差捷径，以避免梯度消失。由于我们将注意力计算限制在局部窗口内，多头局部注意力的复杂度为 $O(FHWfhw)$，其计算成本为

$$O(\mathrm{MHLA}) = (fhw)^2 \times \frac{FHW}{fhw} = FHWfhw. \tag{2}$$

与完整的时空注意力 $O(T^2S^2)$ 相比（它也可以改写为 $O(F^2H^2W^2)$），我们的局部注意力效率高得多。

多头全局注意力：尽管我们的局部注意力计算效率很高，但缺乏捕获全局信息的能力。为了进一步捕获帧序列中的全局关联，我们采用多头全局注意力作为补充，以学习长程依赖关系。

与标准的多头自注意力机制类似，我们的全局注意力也以查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 作为输入。为降低计算开销，我们提出使用窗口级池化对输入 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 进行下采样。具体而言，我们通过卷积操作将特征图划分并池化

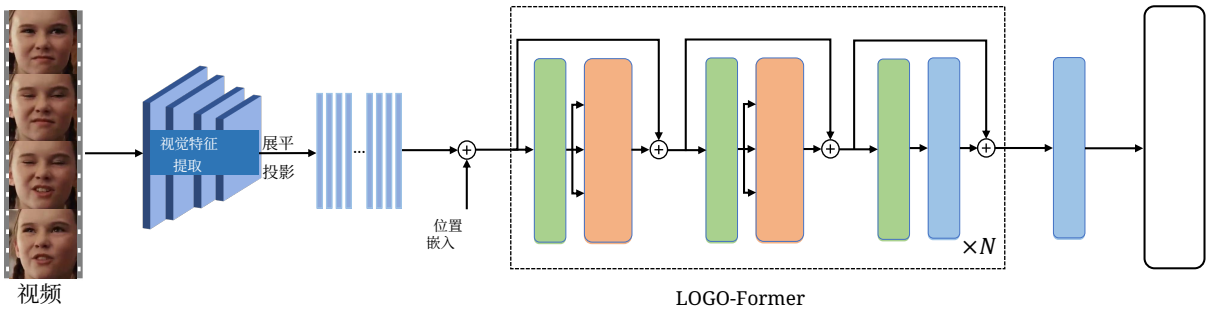

图 2。我们提出的动态面部表情识别方法概述。从视频中采样的面部帧经 CNN骨干网络处理，得到帧级特征序列。LOGO-Former 联合计算空间注意力和时间注意力，以捕获具有判别性的特征标记。

为不重叠区域，其中每个区域都是特征图的时空抽象。每个区域用于向各个查询标记传递全局上下文信息。多头全局注意力的形式化表示为

$$\boldsymbol{Q}^k = W_Q^k \mathrm{LN}(\boldsymbol{Y}^k), \tag{3}$$

$$\boldsymbol{K}^k = W_K^k \mathrm{LN}(\mathrm{Pool}(\boldsymbol{Y}^k)), \tag{4}$$

$$\boldsymbol{V}^k = W_V^k \mathrm{LN}(\mathrm{Pool}(\boldsymbol{Y}^k)), \tag{5}$$

$$\mathrm{MHGA}(\boldsymbol{Y}^k) = \boldsymbol{Y}^k + \mathrm{softmax}(\frac{\boldsymbol{Q}^k(\boldsymbol{K}^k)^T}{\sqrt{D_h}})\boldsymbol{V}^k, \tag{6}$$

$$\boldsymbol{X}^k = \mathrm{MHGA}(\boldsymbol{Y}^k) + \mathrm{MLP}(\mathrm{LN}(\mathrm{MHGA}(\boldsymbol{Y}^k))), \tag{7}$$

其中，Pool 表示窗口级池化方法（例如平均池化、卷积），softmax 表示 softmax 函数，D_h 是每个头的隐藏维度。此外，MLP 表示用于非线性变换的多层感知机，此处还使用了残差捷径。在不失一般性的前提下，假设我们将特征图 $\boldsymbol{Y}_k$ 池化为 $\frac{FWH}{fwh}$ 个词元，则多头全局注意力的复杂度为

$$O(\mathrm{MHGA}) = (\frac{FWH}{fwh})FWH = \frac{(FHW)^2}{fwh}. \tag{8}$$

因此，LOGO-Former 的整体注意力复杂度可计算为

$$\begin{aligned} O(\mathrm{Ours}) &= FHWfhw + \frac{(FHW)^2}{fwh} \\ &< \underbrace{(HW)^2F + HWF^2}_{\mathrm{TimeSformer}} < \underbrace{(FHW)^2}_{\mathrm{MSA}}. \end{aligned} \tag{9}$$

最后，我们对最终块的分类标记 $\boldsymbol{X}_{(0,0)}^N$ 应用一个全连接（FC）层：

$$\boldsymbol{p} = FC(\boldsymbol{X}_{(0,0)}^N), \tag{10}$$

其中，$\boldsymbol{p}$ 是 C 面部表情类别的预测分布。

紧凑损失正则化： 要学习适用于野外动态面部表情识别的判别性时空特征，损失函数必须具备拉大不同类别间特征距离的能力。为此，我们提出使用对称 Kullback-Leibler（KL）散度 $\{\mathcal{D}(\boldsymbol{u}'||\boldsymbol{p}') + \mathcal{D}(\boldsymbol{p}'||\boldsymbol{u}')\}$ 来衡量分布 $\boldsymbol{u}'$ 与 $\boldsymbol{p}'$ 之间的差异，并对预测分布 $\boldsymbol{p}'$施加约束。其中，$\boldsymbol{u}'$ 是关于 $C-1$ 的均匀分布，$\boldsymbol{p}'$ 是预测分布，但不包括对应目标 y的概率。$\boldsymbol{u}'$ 由 Softmax 函数计算得出：

$$\boldsymbol{p}' = \mathrm{softmax}(\{\hat{y}_c^{logits}\}_{c\neq y}), \tag{11}$$

其中，$\{\hat{y}_c^{logits}\}_{c\neq y}$ 表示非目标预测 logits。因此，正则化项的形式为

$$\begin{aligned} \mathcal{L}_{term} &= \mathcal{D}(\boldsymbol{u}'||\boldsymbol{p}') + \mathcal{D}(\boldsymbol{p}'||\boldsymbol{u}') \\ &= \sum_{c\neq y} \frac{1}{C-1} \log\left(\frac{1}{(C-1)\boldsymbol{p}'}\right) + \sum_{c\neq y} \log\left(\frac{1}{(C-1)\boldsymbol{p}'}\right). \end{aligned} \tag{12}$$

借助正则化项 $\mathcal{L}_{term}$，交叉熵损失引导模型提高非目标预测 logits 的均匀性。

3. 实验

我们使用两个野外动态面部表情识别数据集（即 DFEW [14] 和 FERV39K [15]）评估所提出的方法。对于 DFEW 和 FERV39K，处理后的面部区域图像均由官方检测和对齐，并公开提供。我们的模型在 DFEW 和 FERV39K 上以批量大小 32 训练 100 个周期，使用两块 NVIDIA GTX 1080Ti GPU。我们使用初始学习率为 0.001 的 SGD 优化器 [16]和尖锐度感知最小化 [17] 来优化所提出的模型。我们使用在 MS-Celeb-1M [18]上预训练的 ResNet18 作为 CNN骨干网络。时空 Transformer 层数 N 和注意力头数经实验分别设为 4 和 8。非加权平均召回率（UAR）和加权平均召回率（WAR）作为评估指标。

表1。在 DFEW 上通过 5折交叉验证与其他当前最先进的方法进行比较。最佳结果以粗体标示，下划线表示次优结果。TI 表示时间插值，DS 表示动态采样。评估指标包括非加权平均召回率（UAR）和加权平均召回率（WAR）。

方法	样本策略	各情绪准确率（%）							指标（%）	
		快乐	悲伤	中性	愤怒	惊讶	厌恶	Fear	UAR	WAR
R(2+1)D18	TI	79.67	39.07	57.66	50.39	48.26	3.45	21.06	42.79	53.22
ResNet18+长短期记忆网络	TI	78.00	40.65	53.77	56.83	45.00	**4.14**	21.62	42.86	53.08
3D R.18+中心损失	TI	78.49	44.30	54.89	58.40	52.35	0.69	25.28	44.91	55.48
EC-STFL	TI	79.18	49.05	57.85	60.98	46.15	2.76	21.51	45.35	56.51
3D ResNet18	DS	76.32	50.21	64.18	62.85	47.52	0.00	24.56	46.52	58.27
ResNet18+长短期记忆网络	DS	83.56	61.56	68.27	65.29	51.26	0.00	29.34	51.32	63.85
Resnet18+GRU	DS	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02
Former-DFER	DS	84.05	62.57	67.52	70.03	**56.43**	3.45	31.78	53.69	65.70
LOGO-Former	DS	**85.39**	**66.52**	**68.94**	**71.33**	54.59	0.00	**32.71**	**54.21**	**66.98**

表 2。与 FERV39K 上其他方法的比较。最佳结果以粗体显示。

方法	浮点运算次数	指标 (%)	
		非加权平均召回率	加权平均召回率
3D ResNet18	-	26.67	37.57
ResNet18+长短期记忆网络	-	30.92	42.59
VGG13+长短期记忆网络	-	32.79	44.54
Former-DFER	9.11	37.20	46.85
基线	25.16	37.41	47.63
带正则化的基线	25.16	37.76	48.11
无正则化的 LOGO-Former	10.27	37.60	47.85
LOGO-Former	10.27	**38.22**	**48.13**

3.1. 与最先进方法的比较

我们在 DFEW 和 FERV39K 上根据 UAR 和 WAR 将所提方法与其他方法进行比较。DFEW 和 FERV39K 的比较结果分别见表1和表2。我们的方法在这两项指标上均取得了最佳结果。具体而言，Former-DFER 是此前的最先进方法，其 UAR 为 53.69%，WAR 为 66.65%。与 Former-DFER 相比，我们的方法在 UAR 和 WAR 上分别提升了 0.52% 和 1.28%。此外，我们的方法在 UAR 和 WAR 上分别比 EC-STFL 提高了 8.86% 和 10.47%；EC-STFL 同样旨在增强类内相关性并增大类间距离。如表1所示，与其他方法相比，我们的方法在类别级准确率方面也取得了更好的结果。

我们还在 FERV39K 上进行了进一步评估，并在表2中列出了浮点运算次数。我们通过在 [11] 中使用标准的空间注意力和时间注意力，构建了一个强基线方法。
表2中的结果表明，与我们的基线方法相比，我们的模型不仅能够实现更高的识别性能，同时还具有显著更高的效率。图3中的可视化分析表明，我们的紧凑损失

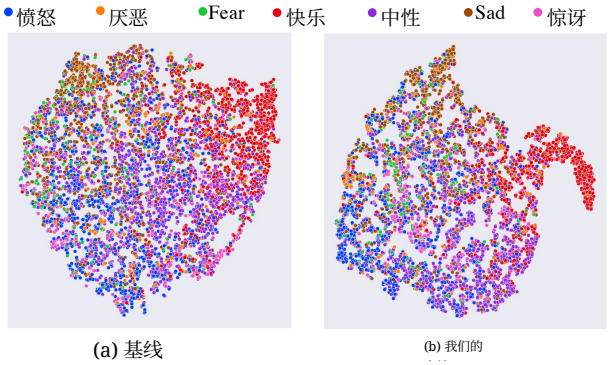

图3。t-SNE [19] 在 FERV39K 上生成的特征分布可视化。(a) 表示采用标准交叉熵损失的模型。(b) 表示采用紧凑损失正则化项的模型。

正则化项使学习到的特征具有更好的聚合效果，并在不同表情之间呈现更清晰的类间边界。

4. 结论

本文提出了一种简单而有效的局部-全局 Transformer（LOGO-Former）以及紧凑损失正则化项，用于野外动态面部表情识别（DFER）。我们在每个模块中联合应用局部注意力和全局注意力，以迭代方式学习时空表示。为进一步增强模型的判别能力，我们通过紧凑损失正则化项约束预测分布，从而增强类内相关性并增大类间距离。实验结果和可视化结果表明，我们的方法能够学习具有判别性的时空特征表示，并提高分类间隔。

5. 参考文献

[1] Zengqun Zhao、Qingshan Liu 和 Shanmin Wang，《学习用于野外面部表情识别的深度全局多尺度与局部注意力特征》，*IEEE* 图像处理汇刊，第 30 卷，第 6544–6556 页，2021年。[2] Fuyan Ma、Bin Sun 和 Shutao Li，《结合视觉变换器与注意力选择性融合的面部表情识别》，*IEEE* 情感计算汇刊，2021年，doi：10.1109/TAFFC.2021.3122146。[3] Shan Li 和 Weihong Deng，《深度面部表情识别：综述》，*IEEE* 情感计算汇刊，2020年，doi：10.1109/TAFFC.2020.2981446。[4] Seung Ho Lee、Wissam J Baddar 和 Yong Man Ro，《利用峰值表情和类内变化人脸图像进行协同表情表示，以实现视频中实用的主体无关情绪识别》，模式识别，第 54 卷，第 52–67 页，2016年。[5] Huiyuan Yang、Umur Ciftci 和 Lijun Yin，《通过去表情残差学习进行面部表情识别》，载于*IEEE* 计算机视觉与模式识别会议论文集，2018年，第 2168–2177 页。[6] Yuanyuan Liu、Xiaohui Yuan、Xi Gong、Zhong Xie、Fang Fang 和 Zhongwen Luo，《条件卷积神经网络增强的面部表情识别随机森林》，模式识别，第 84 卷，第 251–261 页，2018年。[7] Th´eo Ayral、Marco Pedersoli、Simon Bacon 和 Eric Granger，《用于三维卷积神经网络的时间随机 softmax：在面部表情识别中的应用》，载于*IEEE/CVF* 冬季计算机视觉应用会议论文集，2021年，第 3029–3038 页。[8] Valentin Vielzeuf、St´ephane Pateux 和 Fr´ed´eric Jurie，《用于野外视频情绪分类的时间多模态融合》，载于第 *19* 届 *ACM* 多模态交互国际会议论文集，2017 年，第 569–576 页。[9] Shuyan Li、Xiu Li、Jiwen Lu 和 Jie Zhou，《通过双向变换器实现自监督视频哈希》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2021年，第 13549–13558 页。[10] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等，《一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer》，*arXiv* 预印本 *arXiv:2010 11929*，2020。。[11] Fuyan Ma、Bin Sun 和 Shutao Li，《用于野外动态面部表情识别的时空 Transformer》，*arXiv* 预印本 *arXiv:2205.04749*，2022年。[12] Gedas Bertasius、Heng Wang 和 Lorenzo Torresani，《视频理解只需要时空注意力吗？》，*arXiv* 预印本 *arXiv:2102.05095*，第 2 卷，第 3 期，第 4 页，2021年。[13] Adrian Bulat、Juan Manuel Perez Rua、Swathikiran Sudhakaran、Brais Martinez 和 Georgios Tzi miropoulos，《用于视频变换器的时空混合注意力》，神经信息处理系统进展，第 34 卷，第 19594–19607 页，2021年。[14] Xingxun Jiang、Yuan Zong、Wenming Zheng、Chuangao Tang、Wanchuang Xia、Cheng Lu 和 Jiateng Liu，《DFEW：用于识别野外动态面部表情的大型数据库》，载于第*28*届 *ACM* 多媒体国际会议论文集，2020年，第2881–2889页。[15] Yan Wang、Yixuan Sun、Yiwen Huang、Zhongying Liu、Shuyong Gao、Wei Zhang、Weifeng Ge 和 Wenqiang Zhang，《FERV39K：用于视频面部表情识别的大型多场景数据集》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022年，第20922–20931页。[16] Sebastian Ruder，《梯度下降优化算法综述》，*arXiv* 预印本 *arXiv:1609.04747*，2016年。[17] Pierre Foret、Ariel Kleiner、Hossein Mobahi 和 Behnam Neyshabur，《用于高效提升泛化能力的尖锐度感知最小化》，*arXiv* 预印本 *arXiv:2010 01412*，2020年。。[18] Yandong Guo、Lei Zhang、Yuxiao Hu、Xiaodong He 和 Jianfeng Gao，《MS-Celeb-1M：大型人脸识别数据集与基准》，载于欧洲计算机视觉会议。Springer，2016年，第87–102页。[19] Laurens Van der Maaten 和 Geoffrey Hinton，《使用 t-SNE 可视化数据》，机器学习研究期刊，第9卷，第11期，2008年。