文章编号:1671-6833(2025)02-0043-08

基于改进 GAN 的人机交互手势行为识别方法

张富强[1,2], 白筠妍[1,2], 穆 慧[3]

(1.长安大学 道路施工技术与装备教育部重点实验室,陕西 西安 710064;2.长安大学 智能制造系统研究所,陕西 西安 710064;3.济南职业学院 机械制造学院,山东 济南 250002)

摘 要:为改善现有手势识别算法需要大量训练数据的现状,针对识别准确率不高、识别过程复杂等问题,基于生成式对抗网络(GAN)和变分自编码器,引入标签信息,提出一种基于改进 GAN 模型的人机交互手势行为识别方法。首先,在编码器和解码器中分别添加改进 InceptionV2 和 InceptionV2-trans 结构增强模型的特征还原能力;其次,在各组成网络中进行条件批量归一化(CBN)处理改善过拟合,以 Mish 激活函数代替 ReLU 函数提升网络性能;最后,通过实验证明该方法能够以较少的样本获得 100%的分类准确率,且收敛时间短,验证了该方法的可靠性。

关键词:人机交互; 生成对抗网络; 变分自编码器; 手势识别; 条件批量归一化

中图分类号: TP391.41 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.02.012

新一代信息技术、人工智能与先进制造技术的融合发展使得人机交互成为开发智能装备及智能产品的重要组成部分。近年来,语音、手势、姿势、表情等自然人机交互技术使计算系统具有强感知能力、多通道能力及自然性等特点,替代鼠标、键盘,成为人与机器相互交流和沟通、决策与执行的交互媒介。在人机交互场景中,融入手势识别技术建立人与机器的对话机制将带来更加灵活高效的体验。

深度学习是机器学习的一个重要分支,具备极强的自适应能力,并以其优越的性能和无须手工工程的便利性成为目前手势识别研究的主要工具。Gadekallu 等[1]采用基于乌鸦搜索的卷积神经网络模型实现了高精度的手势识别。张富强等[2]提出了一种基于三维卷积神经网络的多模态加工作业手势识别方法。Alawwad 等[3]开发了基于 faster R-CNN 的手势识别系统。Zhou 等[4]结合注意力机制和多路径特征融合方法,提出一种轻量级静态手势识别网络。范晶晶等[5]引入重影通道映射对基于 YOLOv4-tiny 的手势识别算法进行了改进。Chen 等[6]通过添加层聚合网络及 CBAM 注意力机制,提出一种改进 YOLOv5 手势识别方法,增强了模型鲁棒性。上述深度学习算法均具有优异的特征表示和学习能力,但对数据集的规模和质量要求较高,且在准确度上依然有待提升。生成式对抗网络(generative adversarial network, GAN)是 Goodfellow 等[7]于 2014 年提出的一种由博弈论中零和博弈概念演化而来的无监督学习模型。虽然 GAN 更广泛应用于图像生成、数据增强领域,但因其可以在对抗训练过程中生成样本以扩充数据集,并提高判别器训练的效果,因而在图像分类领域也表现出优越性。彭冲等[8]提出基于条件生成对抗网络的手语骨架缺失关节点修复方法,极大提升了手语识别的准确率。钱园园等[9]结合 GAN 和 VGGNet-16 设计了一个针对遥感图像的半监督分类方法,减轻了分类问题对标签信息的依赖。Jiang 等[10]提出一种基于 WiFi 的手势识别系统 WiGAN,通过 GAN 提取卷积特征,借助支持向量机完成手势的识别。Meng 等[11]在多个数据集上验证了 GAN 处理图像分类任务的优越性。郝博等[12]在 YOLOv5 的基础上加入 GAN 和 Swin Transformer 模块,融入 SENet 注意力机制,提高了手势识别模型准确度。

本文针对手势识别准确度不高、识别过程复杂等缺陷,提出一种基于改进 GAN 的人机交互手势行为识别方法,在网络结构中添加 InceptionV2 及 In-

收稿日期:2024-10-06;**修订日期**:2024-12-30

基金项目:国家重点研发计划项目(2021YFB3301702)

作者简介:张富强(1984—),男,山西运城人,长安大学副教授,博士,主要从事面向人机交互的智能制造的研究,E-mail:fqzhang@chd.edu.cn。

引用本文:张富强,白筠妍,穆慧.基于改进 GAN 的人机交互手势行为识别方法[J].郑州大学学报(工学版),2025,46(2):43-50.(ZHANG F Q,BAI J Y,MU H. Human-machine interaction oriented gesture recognition method based on improved GAN[J]. Journal of Zhengzhou University(Engineering Science),2025,46(2):43-50.)

ceptionV2-trans 改进模块,引入条件批量归一化操作和 Mish 激活函数对模型进行优化。使用 PyTorch 框架搭建模型,用公开手势图片数据集训练并测试模型,最后以识别准确率和训练速度为标准验证了该方法的可行性。

1 问题描述

在人机交互过程中,设备通过传感器、物联网等技术实时采集设备的运行数据和状态数据,经解析后传输给孪生设备,驱动其保持与物理设备的互联互通和实时映射。工业相机采集操作人员手势图像并实时传输给手势识别模块,经由基于改进 GAN 算法的手势识别模块,生成对应控制指令,驱动孪生设备执行相应操作,而孪生环境将通过状态行为控制模块对物理实体实施物理行为控制。机器的状态数据将输送给行为控制策略模块进行需求分析和调整,得出交互策略,通过孪生场景反馈给操作人员,便于操作人员利用手势指令控制孪生设备,进而驱动机器对其工作状态做出调整,形成闭环。人机交互逻辑图如图 1 所示。

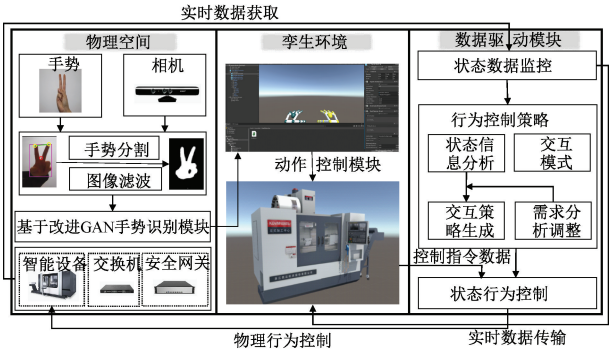


图 1 人机交互手势识别逻辑图

Figure 1 Human-machine interaction gesture recognition logic diagram

2 改进 GAN 算法介绍

2.1 生成对抗网络

GAN 由生成器 G 和判别器 D 组成。G 以随机噪声 z 为输入,尝试学习真实样本的分布以生成足以欺骗 D 的虚假样本。D 的输入分别为真实样本 x 和 G 的生成样本 $G(z)$,其目标是尽可能准确判别真假,输出一个 0~1 的概率值。两个网络在连续的对抗中共同进步,直至达到纳什均衡。得到目标函数:

$$\min_G \max_D V(D,G) = E_{x\sim P_{\text{data}}(x)}[\log D(x)] + E_{z\sim P_z(z)}[\log(1 - D(G(z)))]。 \quad (1)$$

式中:$P_{\text{data}}(x)$ 为真实数据的分布;$P_z(z)$ 为噪声的分布。G 的目的是使 $D(G(z))$ 趋近于 1;D 的目的是使得 $D(x)$ 趋近于 1,$D(G(z))$ 趋近于 0。

Mirza 等[13]将标签信息分别输入 G 和 D 中,提出了一种条件生成对抗网络(conditional generative adversarial nets,CGAN),由标签信息引导图像向指定类别方向优化更新,加速模型训练。训练过程中,标签 c 先与随机噪声 z 拼接起来输入 G 中,获得生成数据 $G(z|c)$,再与真实数据 x 和生成数据 $G(z|c)$ 分别拼接作为 D 的输入,其目标函数可表示为

$$\min_G \max_D V(D,G) = E_{x\sim P_{\text{data}}(x)}[\log(D(x|c))] + E_{z\sim P_z(z)}[\log(1 - D(G(x|c)))]。 \quad (2)$$

2.2 变分自编码器

变分自编码器(variational autoencoder,VAE)由 Kingma 等[14]提出,从隐变量分布入手展开学习,而非盲目尝试还原隐变量具体取值,从而具备了规则性和对噪声的鲁棒性,改善了过拟合。编码器 Enc 将原始数据 x 编码为潜在表示 z,$z \sim \text{Enc}(x) = q(z|x;\phi)$;解码器 Dec 将潜在表示 z 重构为 $\bar{x}$,$\bar{x} \sim \text{Dec}(z) = p(x|z;\theta)$。通过梯度的反向传播更新待优化网络参数 ϕ 和 θ,优化函数如下:

$$L(\phi,\theta;x) = E_{z\sim q(z|x;\phi)}[\log p(x|z;\theta)] - D_{\text{KL}}(q(z|x;\phi) \parallel p(z;\theta))。 \quad (3)$$

式中:$E_{z\sim q(z|x;\phi)}[\log p(x|z;\theta)]$ 为 x 与 $\bar{x}$ 之间的重构误差;$D_{\text{KL}}(q(z|x;\phi) \parallel p(z;\theta))$ 为近似后验分布 $q(z|x;\phi)$ 与假设先验分布 $p(z;\theta)$ 的相似度量。

2.3 改进 GAN

VAE 的解码编码操作相当于重现一张被压缩的图片,经此处理的图片必然面临清晰度降低问题。CGAN 训练时易发生梯度消失现象,导致训练不稳定,这是因为训练初期生成器输入为随机噪声,生成分布可能与真实分布差距过大或是毫无重叠,使得 JS 散度变为常数,优化梯度消失为零。改进 GAN 融合了 VAE 和 CGAN,前半部分可以看作 VAE,后半部分看作 CGAN,VAE 中的解码器在 CGAN 中充当生成器的角色[15]。对于 VAE 来说,由于判别器的存在,博弈对抗可以提高解码器生成图片的质量,提升清晰度。而对于 CGAN 来说,将编码器的输出作为输入替代高斯噪声,可以大大提升训练初期生成图片的质量,使其快速摆脱随机性和无序性,避免优化梯度消失。其网络结构如图 2 所示。

3 基于改进 GAN 的手势识别

为了更好地胜任小样本下的手势识别问题,提高识别准确率,缩短训练时间,提出基于改进 GAN

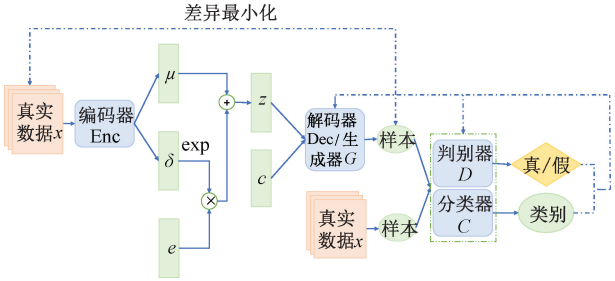


图 2　改进 GAN 网络结构图

Figure 2　Structure of improved GAN

模型的手势识别方法。

3.1　条件批量归一化

批量归一化(batch normalization,BN)是一种常见的网络训练手段。模型训练时一个层的参数调整会导致后续层输入数据的分布发生变化,这种微小变化累加起来将造成近输出层输出的剧烈波动。BN 层的加入可以通过规范化处理上一层的输出减轻这种不稳定性,加快网络收敛。但在 GAN 中使用 BN 会导致生成图片出现一定程度的同质化问题,这是因为 BN 对一个批量里不同类别的训练数据使用了统一的放缩和偏置进行归一化处理,而显然不同类别的数据分布具有不同的均值和方差,简单的统一处理是不合理的[16]。

由此条件批量归一化(conditional batch normalization,CBN)[17]技术应运而生。CBN 处理数据时对各类数据特征图分别进行归一化处理,根据类别标签确定放缩和偏置,保留每一类数据之间的差异性。CBN 的批处理过程如下:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i; \tag{4}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B)^2; \tag{5}$$

$$x_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}; \tag{6}$$

$$y_i = \gamma(c)x_i + \beta(c)。 \tag{7}$$

式(4)、式(5)为数据均值和方差的计算步骤;式(6)、式(7)表示数据的标准化、平移缩放过程。

在 GAN 网络的各卷积层后使用 CBN 层代替 BN 层进行归一化处理,能增强模型处理多类数据的适应能力,从而加速收敛。

3.2　激活函数的改进

原模型中使用的 ReLU 激活函数将负区输入置零,导致训练过程中部分神经元的权重参数不断更新为 0,失去学习和表征能力,造成“神经元坏死”。为缓解该问题,提出以 Mish 激活函数替换 ReLU 激活函数。Mish 激活函数是 ReLU 函数的拓展和延伸,该函数在负区仍然具有非零梯度,因而在极大程度上缓解了神经元坏死问题,且曲线更加平滑,具有连续可导性和非单调性,能更好地捕捉数据特征,从而加速模型收敛[18]。它的定义如式(8)所示,函数图像如图 3 所示。

$$\mathrm{Mish}(x) = x \cdot \mathrm{Tanh}(\mathrm{Softplus}(x))。 \tag{8}$$

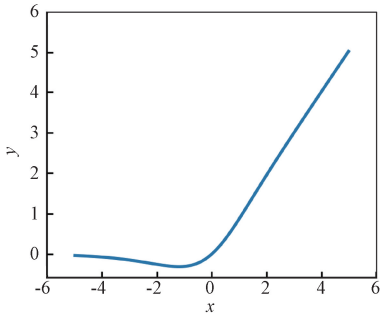


图 3　Mish 激活函数图像

Figure 3　Image of Mish activation function

3.3　增强对抗网络生成图像的能力

考虑到 GAN 系列网络的训练中生成器往往处于劣势地位,有必要提升生成器的特征提取和还原能力。加深或拓宽网络都能达到该效果,但也会导致网络参数过多,大大增加计算复杂度,造成过拟合。对此,谷歌团队提出 Inception 结构,利用多个大小不一的卷积核提取图像的全局和局部信息并进行通道上的拼接融合,在保留图像特征的基础上保持了网络结构的稀疏性,借助聚类形成的密集矩阵提升了网络性能。InceptionV2 是在 Inception 基础上增加了 BN 层计算的改进版本,减少信息损失的同时降低计算复杂度[19]。分别用 CBN 层和 Mish 激活函数代替 InceptionV2 中的 BN 层和 ReLU 激活函数进行改进,并将改进 InceptionV2 结构中的卷积层替换为反卷积层,将池化层替换为反池化层,构成 InceptionV2-trans。在编码器和解码器网络中分别添加改进后的 InceptionV2 和 InceptionV2-trans 结构,如图 4 和图 5 所示,使得编码器利用不同大小的卷积核获取图像不同维度的特征,将其输送给解码器,再通过不同大小的反卷积层尽可能完整地还原特征。

3.4　改进网络的实现

改进网络由变分自编码器(编码器、解码器)、判别器、分类器组成,其网络结构如图 6 所示。

编码器包含 4 层 3×3、步长为 2 的卷积层,两层 InceptionV2 结构以及一层全连接层,为各卷积层添加 CBN 层和 Mish 激活函数规范化输出,增强网络表达能力。全连接层将数据维度调整至 1×1×100,分别输出均值和方差用以合成采样变量。

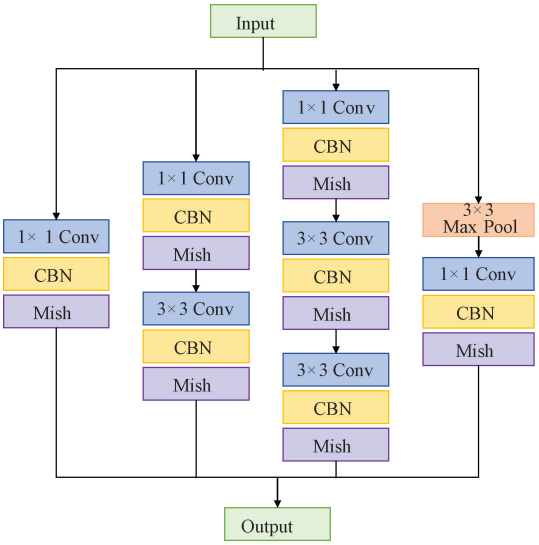


图 4　改进 InceptionV2 结构

Figure 4　Structure of improved InceptionV2

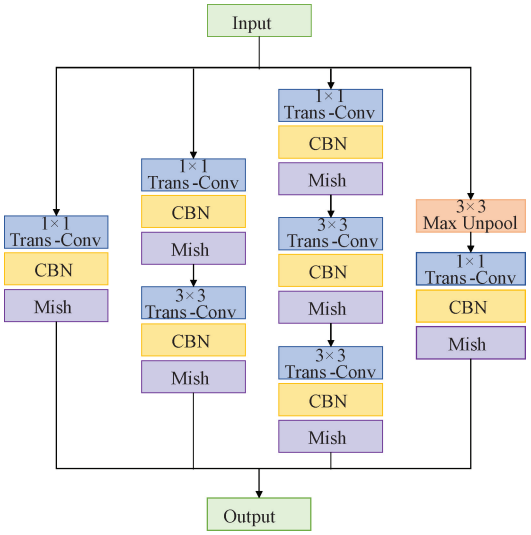


图 5　改进 InceptionV2-trans 结构

Figure 5　Structure of improved InceptionV2-trans

解码器由一层全连接层、4 层转置卷积网络以及两层 InceptionV2-trans 网络构成，除最后一个转置卷积层使用 Tanh 激活函数外，其余均使用 Mish 激活函数，并在使用激活函数前进行 CBN 处理防止过拟合。

判别器由 4 层卷积层和两层全连接层组成，此处用步长为 2 的卷积层替换原 GAN 网络的最大池化层，减小信息丢失。卷积层激活函数使用 Mish，全连接层使用 Sigmoid，除输入层外其余卷积层均采用 CBN 进行条件批量归一化处理。

分类器网络与判别器同构，用于判定图片类别，但最后全连接层的输出由 1 位改为 10 位，其中 10 是待识别的手势种类数。

编码器通过将输入的真实数据 x_r 压缩映射到隐

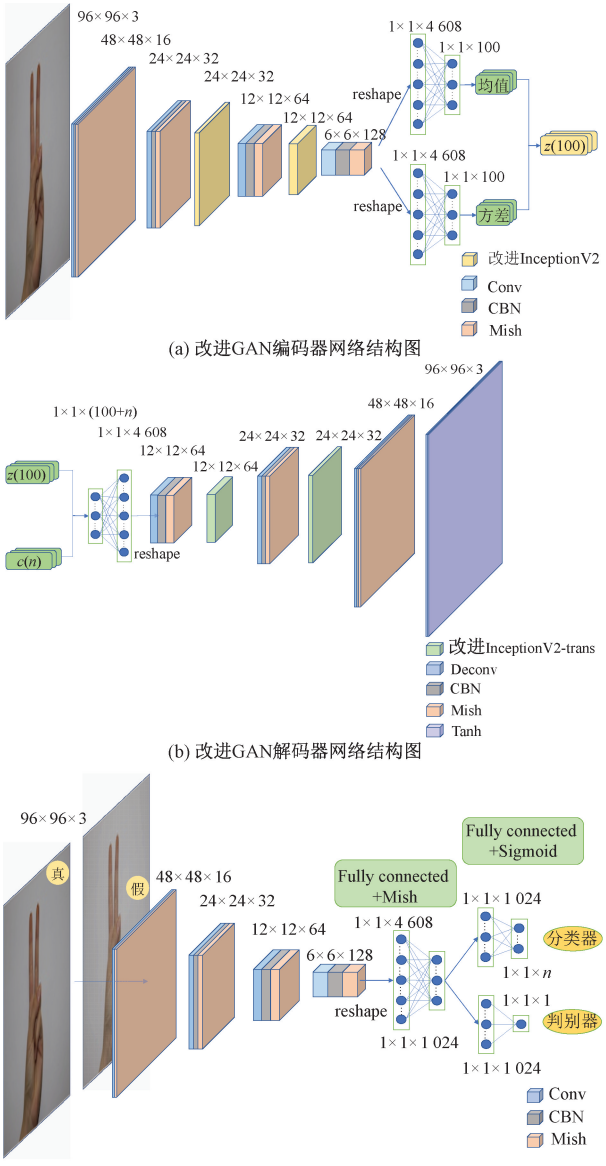


图 6　改进 GAN 网络结构图

Figure 6　Structure of improved GAN

空间提取图像的原始特征 $z, x_r \sim P_r, z \sim \mathrm{Enc}(x_r) = q(z|x_r)$。解码器将真实图像编码 z 或噪声 $z_p(z_p \sim P_z)$ 与标签的 one-hot 编码拼接后作为输入，以期生成能够让判别器不能正确判别的图片 x_f 和 $x_p, x_f \sim \mathrm{Dec}(z) = p(x_r|z), x_p \sim \mathrm{Dec}(z_p)$。因而 VAE 的损失函数由 3 部分组成，首先，编码器输出的正态分布内部表示经解码器的输出 x_f 应尽可能接近原图 x_r；其次，解码器生成的图片 x_f 应尽可能让判别器输出 1；最后，x_f 应尽可能让判别器判断出对应类别。进一步得到损失函数：

$$L_{\mathrm{VAE}} = E_{z\sim q(z|x_r)}[\log p(x_r|z)] + E_{z\sim E(x_r,c)}[\log D(\mathrm{Dec}(z))] - E[\log P(c|\mathrm{Dec}(z))] - D_{\mathrm{KL}}(q(z|x_r,c) \| p(z))。 \quad (9)$$

将真实图片 x_r 和生成图片 x_f、x_p 输入判别器中进行判别,结果经 Sigmoid 层输出,为 0~1 任意概率值。对 x_r 的判决值应尽量趋近 1,对 x_f 和 x_p 的判决值应尽量趋近 0。其损失函数可表示为

$$L_D = -E_{x_r\sim P_r}[\log D(x_r)] - E_{z\sim E(x_r,c)}[\log(1 - D(\mathrm{Dec}(z)))] - E_{z_p\sim P_z}[\log(1 - D(\mathrm{Dec}(z_p)))]。\quad (10)$$

分类器用于分辨真实图片 x_r 类别,其输出代表每一类别的概率,由 Argmax 函数取其输出最大值的索引,即可确定对应分类。分类器的目标是使输出的类别信息与原标签尽量一致。其损失函数为

$$L_C = -E_{x_r\sim P_r}[\log P(c\mid x_r)]。\quad (11)$$

如图 7 所示,训练时首先训练分类器,将梯度置零,反向计算梯度、更新参数,再依次训练判别器和变分自编码器,循环交替训练直至达到规定迭代次数。

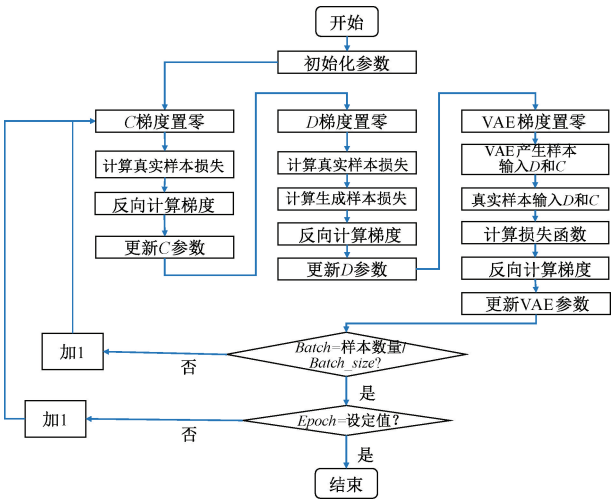


图 7　改进 GAN 算法流程图

Figure 7　Flow chart of improved GAN

4　仿真验证

本文对 Mavi 手势数据集[20]展开实验。计算机环境为 Win11,配置 Intel i5 处理器、内存为 4 GB 的 GPU(NVIDIA,GeForce RTX2050),基于 Python 语言在机器学习框架 PyTorch 中搭建模型。

4.1　数据预处理

Mavi 手势数据集包含样本均衡的 0~9 手势数字图片,共计 2 062 张。图像数据维度为[3,100,100],即长宽皆为 100 的三通道图像。采用 5∶1的比例按类别随机为训练集和测试集分配样本。

在训练与测试之前,为保证输入数据的一致性,同时保证每一次卷积均能得到整除的尺寸,便于操作和计算,使用 Resize 函数将图片统一裁剪至[3,96,96],再使用 Normalize 函数将图像值的范围从[0,255]归一化至[0,1]。

4.2　训练网络

在 PyTorch 框架中搭建网络并进行训练,训练过程中使用 Adam 优化器,各组成网络均取学习率 $lr=0.000\,1$,优化器超参数 $b_1=0.5$、$b_2=0.999$。

训练过程中,使用可视化工具 TensorBoard 记录观察损失曲线、梯度更新等信息,监督网络训练过程,使其保持良好的优化方向。

4.3　实验结果与分析

设定最大迭代次数为 300 次,分别在 CNN(所构建 CNN 模型与本文提出模型的分类器结构一致)、RNN、CGAN、VAE-GAN、CVAE-GAN 和所提出的改进 GAN 模型上进行实验。

训练结束后需对模型性能做出评估,在此将多分类手势识别问题转化为二分类问题进行研究。表 1 列出了真实情况与预测结果间的 4 种组合,即真正类(*TP*)、假反类(*FN*)、假正类(*FP*)和真反类(*TN*)。

表 1　真实情况和预测结果组合

Table 1　Combination of real situation and predicted results

真实情况	预测结果	
	正类	反类
正类	*TP*	*FN*
反类	*FP*	*TN*

根据分类结果混淆矩阵可得到以下评价指标。

(1)准确率(*ACC*):表示模型预测结果与真实情况相符的样本数量在所有样本中的占比。

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}。\quad (12)$$

(2)*F*1:*F*1 是精确度 *P* 和召回率 *R* 的调和平均。

$$P = \frac{TP}{TP + FP};\quad (13)$$

$$R = \frac{TP}{TP + FN};\quad (14)$$

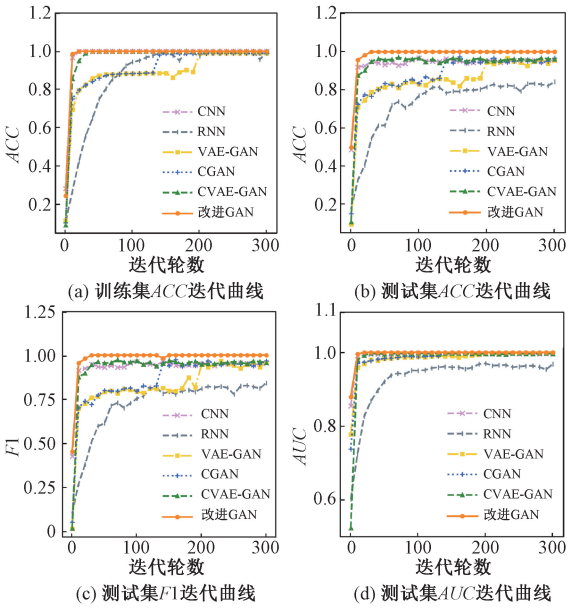
$$F1 = \frac{2 \times P \times R}{P + R}。\quad (15)$$

(3)*AUC*:*AUC* 为 ROC 曲线下的面积,表示经分类器预测后正类得分大于反类的概率。

对于多分类的手势识别问题,以上评估指标的取值为各手势类别评估指标加和后的平均值。

各网络指标迭代曲线如图 8 所示。不同网络各指标首次最高值及对应的迭代次数如表 2 所示。

由图 8 和表 2 得出,在 Mavi 数据集上迭代 300



(a) 训练集*ACC*迭代曲线 (b) 测试集*ACC*迭代曲线

(c) 测试集*F*1迭代曲线 (d) 测试集*AUC*迭代曲线

图 8 前 300 轮各网络指标迭代曲线

Figure 8 Iteration curves of each network index in the first 300 epochs

轮后,各网络均表现出优异的效果。从识别精度上来看,传统判别式模型 CNN 与生成式模型 CGAN、VAE-GAN、CVAE-GAN 在训练集上的准确率均可到达较高水平,但由于数据量有限,在测试集上往往表现出过拟合现象,导致其识别精度不够高。而所提出的改进 GAN 网络在测试集上的各项评价指标均可达 100%,明显高于其他模型,证明所提出方法的有效性。CNN 虽使用了本文所提出的改进方法,但由于缺少对抗生成的过程,识别精度较所提模型仍有较大差异。从收敛速度上来看,生成式模型总体训练速度逊于 CNN,CVAE-GAN 的训练速度较 CGAN 和 VAE-GAN 有了明显提升,但较 CNN 无较大差异。本文所提模型训练集指标于第 19 次迭代达到峰值,测试集指标于第 31 次迭代达到峰值,收敛速度明显快于其他所有模型,且由迭代曲线可以看出其训练过程最为平稳。综上可知,本文所提方法在精度和收敛速度上均存在较大优势,在小样本数据集上表现出优异的性能。

表 2 不同网络各指标首次最高值和对应迭代次数

Table 2 The maximum value of each index in different networks and the corresponding epoch

数据集	模型	最高值(迭代次数)			
		ACC(训练)	*ACC*(测试)	*F*1	*AUC*
Mavi[20]	CNN	0.996 4(28)	0.975 3(209)	0.973 5(209)	0.997 5(101)
	RNN	0.995 8(263)	0.869 2(207)	0.862 3(207)	0.970 0(201)
	VAE-GAN	0.986 8(274)	0.969 2(257)	0.969 2(257)	0.999 0(228)
	CGAN	0.986 2(297)	0.974 9(161)	0.974 8(161)	0.999 5(204)
	CVAE-GAN	0.996 4(48)	0.980 8(229)	0.981 0(229)	0.998 1(117)
	改进 GAN	**0.996 4(19)**	**1.000 0(31)**	**1.000 0(31)**	**1.000 0(31)**
HaGRID[21]	CNN	0.982 8(33)	0.980 8(135)	0.981 0(135)	0.995 4(185)
	RNN	0.982 8(225)	0.905 3(211)	0.902 2(211)	0.978 6(235)
	VAE-GAN	0.988 9(176)	0.979 9(197)	0.979 9(197)	0.999 7(155)
	CGAN	0.975 8(230)	0.973 2(264)	0.973 5(264)	0.999 2(227)
	CVAE-GAN	0.999 0(37)	0.990 6(79)	0.991 0(79)	0.999 5(66)
	改进 GAN	**0.999 0(25)**	**0.999 8(36)**	**0.999 8(36)**	**0.999 9(36)**

在 HaGRID 数据集[21]上验证模型的泛化性,该数据集是最新开源的超大型手势图像公开数据集,包含 50 万个 1 920×1 080 像素的 RGB 图像,涉及 18 类常见手势,且在背景、灯光等方面均有所区别。对数据样本进行精简,选取 10 类手势,每类照片选取 500 张图片,总共 5 000 张。为适应本文所提出模型,统一将样本图片的分辨率等比例调整到 100×100 像素。表 2 中的结果表明本文所提算法在大型数据集上依然表现出优越的性能,具有良好的泛化性。

为了证明模型各部分改进功能模块的有效性,设计 4 组细粒度消融实验,依次去掉所需验证的功能模块,基于 Mavi 数据集进行模型的训练和测试,比较测试集上的最高准确率 *ACC*,每组实验的迭代次数均设置为 300 轮。实验数据如表 3 所示。消融实验 *ACC* 迭代曲线如图 9 所示。

表 3 消融实验性能对比表

Table 3 Performance comparison of ablation experiments

改进方案	CBN	InceptionV2	Mish	*ACC*
0				0.980 8
1	√			0.988 0
2	√	√		0.996 4
3	√	√	√	1.000 0

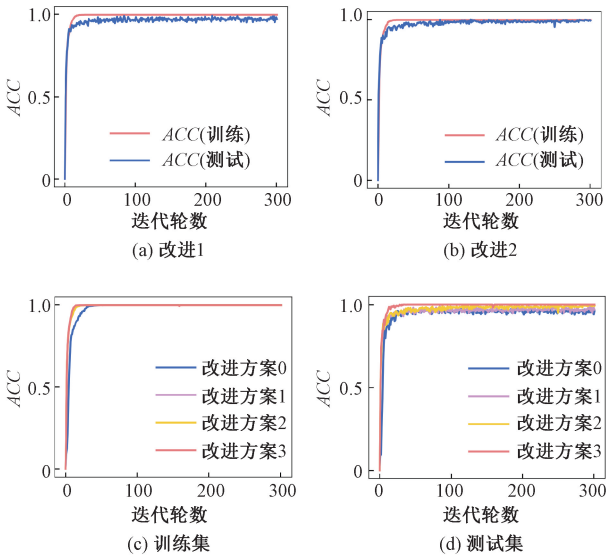


图 9　消融实验 *ACC* 迭代曲线

Figure 9　*ACC* iteration curves of ablation experiments

观察表 3 和图 9 的实验结果可知,将 BN 替换为 CBN 后测试集的准确率由 0.980 8 提升至 0.988 0,且前期测试集的准确率提升加快;改进方案 2 在此基础上添加 InceptionV2 模块,增强了生成器生成图片的能力,通过博弈进一步将分类器准确率提升至 0.996 4;最后,改进方案 3 基于上述改进将 ReLU 激活函数替换为 Mish 激活函数,使得测试集准确率达到 1.000 0,证明了本文所提出 3 项改进方案的有效性。

5　结论

(1)针对人机交互中的手势识别问题,提出了一种基于改进 GAN 网络的手势行为识别方法。以生成对抗网络为基础,通过变分自编码器进行特征提取,利用标签信息指导指定类别图像的生成,提高了 GAN 初期生成图片的质量,缩短了训练时间。

(2)提出在网络中加入改进 InceptionV2 和改进 InceptionV2-trans 结构,使用多个大小不一的卷积核提取并还原数据各尺度信息,进行拼接融合,在保持网络稀疏性的前提下加强了模型生成图片的能力,使训练过程更加稳定;使用 CBN 根据标签信息实施归一化,并以 Mish 激活函数提升模型对数据的表达能力,从而改善过拟合并缓解梯度弥散,提高了模型准确率。

(3)与 CNN、RNN、CGAN、VAE-GAN、CVAE-GAN 算法的识别结果对比表明,提出的改进 GAN 算法能以较小的数据集训练得到较好的分类效果,且具备良好的快速性,验证了所提算法解决手势识别问题的优越性。

参考文献:

[1] GADEKALLU T R, ALAZAB M, KALURI R, et al. Hand gesture classification using a novel CNN-crow search algorithm[J]. Complex & Intelligent Systems, 2021, 7(4): 1855-1868.

[2] 张富强, 曾夏, 白筠妍, 等. 多模态数据融合的加工作业动态手势识别方法[J]. 郑州大学学报(工学版), 2024, 45(5): 30-36.
ZHANG F Q, ZENG X, BAI J Y, et al. Dynamic gesture recognition method for machining operations based on multi-modal data fusion[J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(5): 30-36.

[3] ALAWWAD R A, BCHIR O, MAHER M. Arabic sign language recognition using faster R-CNN[J]. International Journal of Advanced Computer Science and Applications, 2021, 12(3): 692-700.

[4] ZHOU W N, LI X L. PEA-YOLO: a lightweight network for static gesture recognition combining multiscale and attention mechanisms[J]. Signal, Image and Video Processing, 2024, 18(1): 597-605.

[5] 范晶晶, 薛皓玮, 吴欣鸿, 等. 引入重影特征映射和通道注意力机制的手势识别算法[J]. 计算机辅助设计与图形学学报, 2022, 34(3): 403-414.
FAN J J, XUE H W, WU X H, et al. Gesture recognition algorithm introducing ghost feature mapping and channel attention mechanism[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(3): 403-414.

[6] CHEN R X, TIAN X. Gesture detection and recognition based on object detection in complex background[J]. Applied Sciences, 2023, 13(7): 4480.

[7] GOODFELLOW I J, POUGHT-ABADIE J, MIRZA M, et al. Generative adversarial networks [EB/OL]. (2014-06-10)[2024-09-15]. https://doi.org/10.48550/arXiv.1406.2661.

[8] 彭冲, 张金艺, 楼亮亮. 基于条件生成对抗网络的手语样本骨架缺失关节点修复[J]. 计算机辅助设计与图形学学报, 2023, 35(3): 423-433.
PENG C, ZHANG J Y, LOU L L. Missing joint point repair of sign language sample skeleton based on conditional generation adversarial networks[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(3): 423-433.

[9] 钱园园, 刘进锋, 朱东辉. 一种生成对抗网络半监督遥感图像分类方法[J]. 遥感信息, 2022, 37(4): 36-42.
QIAN Y Y, LIU J F, ZHU D H. A semi-supervised remote sensing image classification method on generative adversarial network[J]. Remote Sensing Information, 2022, 37(4): 36-42.

[10] JIANG D H, LI M Q, XU C L. WiGAN: a WiFi based gesture recognition system with GANs[J]. Sensors, 2020, 20(17): 4757.

[11] MENG H, GUO F R. Image classification and generation based on GAN model[C]//2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). Piscataway: IEEE, 2021: 180-183.

[12] 郝博, 尹兴超, 闫俊伟, 等. 基于Gan-St-YOLOv5的复杂环境下的手势识别[J]. 东北大学学报(自然科学版), 2023, 44(7): 953-963.
HAO B, YIN X C, YAN J W, et al. Gesture recognition in the complex environment based on Gan-St-YOLOv5 [J]. Journal of Northeastern University (Natural Science), 2023, 44(7): 953-963.

[13] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-06)[2024-09-15]. ttps://doi.org/10.48550/arXiv.1411.1784.

[14] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB/OL]. (2022-12-10)[2024-09-15]. https://doi.org/10.48550/arXiv.1312.6114.

[15] BAO J M, CHEN D, WEN F, et al. CVAE-GAN: fine-grained image generation through asymmetric training[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2764-2773.

[16] 王爱丽, 薛冬, 吴海滨, 等. 基于条件生成对抗网络的手写数字识别[J]. 液晶与显示, 2020, 35(12): 1284-1290.
WANG A L, XUE D, WU H B, et al. Handwritten digit recognition based on conditional generative adversarial network[J]. Chinese Journal of Liquid Crystals and Displays, 2020, 35(12): 1284-1290.

[17] MIYATO T, KOYAMA M. cGANs with projection discriminator[EB/OL]. (2018-02-15)[2024-09-15]. https://doi.org/10.48550/arXiv.1802.05637.

[18] ALAFTEKIN M, PACAL I, CICEK K. Real-time sign language recognition based on YOLO algorithm[J]. Neural Computing and Applications, 2024, 36(14): 7609-7624.

[19] BOSE S R, KUMAR V S. Efficient inceptionV2 based deep convolutional neural network for real-time hand action recognition[J]. IET Image Processing, 2020, 14(4): 688-696.

[20] MAVI A. A new dataset and proposed convolutional neural network architecture for classification of American sign language digits[EB/OL]. (2020-11-16)[2024-09-15]. https://doi.org/10.48550/arXiv.2011.08927.

[21] ALEXANDER K, KARINA K, ALEXANDER N, et al. HaGRID-HAnd gesture recognition image dataset[C]// 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2024: 4560-4569.

Human-machine Interaction Oriented Gesture Recognition Method Based on Improved GAN

ZHANG Fuqiang[1,2], BAI Junyan[1,2], MU Hui[3]

(1. Key Laboratory of Road Construction Technology and Equipment of Ministry of Education, Chang'an University, Xi'an 710064, China; 2. Institute of Smart Manufacturing Systems Engineering, Chang'an University, Xi'an 710064, China; 3. School of Mechanical Manufacturing, Jinan Vocational College, Jinan 250002, China)

Abstract: In order to improve the current situation that the existing gesture recognition algorithms required a large amount of training data, aiming at the drawbacks of low accuracy and complex recognition process, a gesture recognition method for human-machine interaction based on improved GAN model was proposed through taking the generative confrontation networks combined with the variational self-encoder and the label information. Firstly, the improved InceptionV2 and InceptionV2-trans structures were added to the encoder and decoder respectively to enhance the feature recovery ability of the model. Secondly, conditional batch normalization (CBN) was carried out in each component network to improve overfitting, and Mish activation function was used to replace ReLU to improve the network performance. Finally, the experimental results indicated that the proposed method could obtain 100% classification accuracy with fewer samples and short convergence time, which verified the reliability of the method.

Keywords: human-machine interaction; generative adversarial networks; variational autoencoder; gesture recognition; conditional batch normalization