基于生成式对抗网络的鲁棒人脸表情识别

姚乃明 [1,2] 郭清沛 [1,2] 乔逢春 [1,2] 陈 辉 [1,2] 王宏安 [1,2,3]

摘 要 人们在自然情感交流中经常伴随着头部旋转和肢体动作, 它们往往导致较大范围的人脸遮挡, 使得人脸图像损失部分表情信息. 现有的表情识别方法大多基于通用的人脸特征和识别算法, 未考虑表情和身份的差异, 导致对新用户的识别不够鲁棒. 本文提出了一种对人脸局部遮挡图像进行用户无关表情识别的方法. 该方法包括一个基于 Wasserstein 生成式对抗网络 (Wasserstein generative adversarial net, WGAN) 的人脸图像生成网络, 能够为图像中的遮挡区域生成上下文一致的补全图像; 以及一个表情识别网络, 能够通过在表情识别任务和身份识别任务之间建立对抗关系来提取用户无关的表情特征并推断表情类别. 实验结果表明, 我们的方法在由 CK+, Multi-PIE 和 JAFFE 构成的混合数据集上用户无关的平均识别准确率超过了 90%. 在 CK+ 上用户无关的识别准确率达到了 96%, 其中 4.5% 的性能提升得益于本文提出的对抗式表情特征提取方法. 此外, 在 45° 头部旋转范围内, 本文方法还能够用于提高非正面表情的识别准确率.

关键词 人脸补全, 用户无关, 人脸表情识别, 生成式对抗网络, 卷积神经网络

引用格式 姚乃明, 郭清沛, 乔逢春, 陈辉, 王宏安. 基于生成式对抗网络的鲁棒人脸表情识别. 自动化学报, 2018, **44**(5): 865−877

DOI 10.16383/j.aas.2018.c170477

Robust Facial Expression Recognition With Generative Adversarial Networks

YAO Nai-Ming[1,2] GUO Qing-Pei[1,2] QIAO Feng-Chun[1,2] CHEN Hui[1,2] WANG Hong-An[1,2,3]

Abstract In natural communication, people would express their expressions with head rotation and body movement, which may result in partial occlusion of face and a consequent information loss regarding facial expression. Also, most of the existing approaches to facial expression recognition are not robust enough to unseen users because they rely on general facial features or algorithms without considering differences between facial expression and facial identity. In this paper, we propose a person-independent recognition method for partially-occluded facial expressions. Based on Wasserstein generative adversarial net (WGAN), a generative network of facial image is trained to perform context-consistent image completion for partially-occluded facial expression images. With an adversarial learning strategy, furthermore, a facial expression recognition network and a facial identity recognition network are established to improve the accuracy and robustness of facial expression recognition via inhibition of intra-class variation. Extensive experimental results demonstrate that 90% average recognition accuracy of facial expression has been reached on a mixed dataset composed of CK+, Multi-PIE, and JAFFE. Moreover, our method achieves 96% accuracy of user-independent recognition on CK+. A 4.5% performance gain is achieved with the novel identity-inhibited expression feature. Our method is also capable of improving recognition accuracy for non-frontal facial expressions within a range of 45-degree head rotation.

Key words Face completion, person-independent, facial expression recognition, generative adversarial net (GAN), convolutional neural network (CNN)

Citation Yao Nai-Ming, Guo Qing-Pei, Qiao Feng-Chun, Chen Hui, Wang Hong-An. Robust facial expression recognition with generative adversarial networks. *Acta Automatica Sinica*, 2018, **44**(5): 865−877

收稿日期 2017-08-30 录用日期 2018-02-07
Manuscript received August 30, 2017; accepted February 7, 2018
国家自然科学基金 (61661146002, 61572479), 国家重点研发计划 (2017YFB1002805), 中国科学院前沿科学重点研究计划 (QYZDY-SSW-JSC041) 资助
Supported by National Natural Science Foundation of China (61661146002, 61572479), National Fundamental Research Grant of Science and Technology (2017YFB1002805), and Frontier Science Key Program of Chinese Academy of Sciences (QYZDY-SSW-JSC041)
本文责任编委 左旺孟
Recommended by Associate Editor ZUO Wang-Meng
1. 中国科学院软件研究所人机交互北京市重点实验室 北京 100190 2. 中国科学院大学 北京 100049 3. 中国科学院软件研究所计算机科学国家重点实验室 北京 100190
1. Beijing Key Laboratory of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences, Beijing 100190 2. University of Chinese Academy of Sciences, Beijing 100049 3. State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190

赋予机器感知人类情绪的能力, 使得机器能够识别人的情绪状态, 已经成为提高人机交互系统自动化水平的关键. 在过去的十年中, 人脸表情的识别方法得到了深入研究[1−4], 并逐渐成为分析用户情绪的一种强效技术. 其中, 识别自然的人脸表情是一个重要的研究方向. 在自然交流中, 人的情绪表达往往伴随着丰富的头部姿态和肢体动作, 使得提取有效的表情特征非常困难. 许多方法要求或假设在表达

情绪时, 用户头部始终位于正面或近正面, 并且没有受到人脸局部遮挡的影响. 然而, 这样的限制条件显著降低了表情识别算法的鲁棒性. 此外, 一些方法直接对用户施加约束, 学习用户相关的表情特征. 这种特征对用户身份信息非常敏感, 因而对未知用户的鲁棒性较差. 一个可靠的表情识别系统应当对人脸局部遮挡和用户身份具有较强的鲁棒性, 即能够对存在遮挡的人脸图像进行用户无关的表情识别.

人脸表情识别算法通常需要直接从图像中提取可用于推断表情类别的特征, 然而当人脸局部遮挡存在时, 大多数表情特征的有效性和准确性会因遮挡而降低. 通过图像合成方法还原遮挡图像, 然后再进行表情识别, 是缓解该问题的一类有效方法. 一些研究者为已对齐的人脸图像建立稀疏编码, 通过图像重构来实现遮挡还原[5]. 这种方法假设在相同图像中能够找到相似的模式, 然而对于人脸表情图像, 图像重构不能为遮挡区域恢复充分的表情信息, 因为人脸中的不同部分均含有反映用户身份和表情状态的独特模式, 难以简单地通过组合其它图像部分来得到[6]. 另一方面, 随着卷积神经网络 (Convolutional neural networks, CNN) 在图像分类任务上的突破[7], 许多基于 CNN 的表情识别方法被提出, 弥补了传统方法在鲁棒性方面的不足[8−11]. 对于局部遮挡问题, 一些研究者提出使用 CNN 建立无监督学习模型, 通过编解码网络从遮挡图像中学习特征编码, 在完成保留身份和表情特征的图像合成或变换之后, 再进行表情识别[12−15]. 另外一些研究工作使用生成式对抗网络 (Generative adversarial net, GAN)[16−17], 先局部或完整地生成保持上下文一致性的人脸图像, 然后再对其进行识别[6]. 在基于 GAN 的方法中, 生成器网络 (Generator) 尽可能生成真实的人脸图像, 判别器网络 (Discriminator) 尽可能辨别面部遮挡区域被补全后的图像真实性.

提取表情的本质特征是表情识别算法有效性的关键. 表情特征应对表情变化丰富的区域具有较高的响应, 对身份相关性高而表情相关性低的区域具有较低的响应. 用户相关的表情识别算法能够比较准确地识别在训练时出现过的用户的表情, 然而实际当中的用户身份是难以限定的. 由于对未知用户的泛化能力较差, 这种方法很少被单独使用. 与此不同, 用户无关的表情识别方法对用户身份不敏感, 它通过稀疏编码[18], 差分图像[19] 以及图像融合[20] 等方法对表情图像中的用户身份特征进行抑制, 然后再识别表情. 随着 VGG[21], GoogLeNet[22] 和 ResNet[23] 等 CNN 模型的广泛应用, 表情识别算法能够以数据驱动的方式从表观信息中提取用户无关的表情特征. 尽管如此, 直接使用 CNN 对表情图像数据进行特征提取的方法仍然受到类内差异的限制, 从而难以获得期望的性能. 在同种表情的图像样本之间, 用户身份和图像采集条件等表观差异带来了表情的类内差异, 容易导致表情特征的可辨别性不够鲁棒. 为此, 展示了一种通过抑制类内差异信息来突出表情特征的学习方法, 能够使用 CNN 自动地提取用户无关的表情特征.

本文提出了一种鲁棒的人脸表情识别方法, 能够以用户无关方式识别具有局部遮挡的人脸表情. 基于 Wasserstein GAN (WGAN), 训练了一个稳定的人脸图像生成网络, 然后使用遮挡图像集优化网络的输入隐变量, 对遮挡区域进行保持上下文一致性的人脸图像补全. 对无遮挡图像和遮挡补全图像, 在表情识别任务和身份识别任务之间建立了一种对抗关系, 通过在表情特征提取过程中抑制由身份信息导致的类内差异来提升表情识别的准确性和鲁棒性.

本文的主要贡献: 1) 提出了一种基于 WGAN 的人脸图像补全算法, 能够以生成方式近似还原被遮挡的人脸图像, 缓解因局部表情信息缺失带来的影响, 提高识别算法的鲁棒性. 2) 提出了一种新颖的表情特征学习方法, 通过在表情信息和身份信息之间建立对抗关系来抑制身份特征对表情特征的影响. 该方法能够有效地消除类内差异带来的影响, 从而提高表情识别的准确性和鲁棒性. 3) 展示了一种联合的表情识别算法框架, 在多个基准表情数据集上取得了准确的表情识别结果, 并且能够对 45° 头部旋转范围以内的非正面人脸图像进行用户无关的表情识别.

1 相关工作概述

1.1 生成式对抗网络

生成式对抗网络 (GAN) 是一种无监督的概率分布学习方法, 能够学习真实数据的分布并生成具有较高相似性的新数据集. 设置隐变量 $\boldsymbol{z}$, 生成器网络能够将它映射为新的图像集合, 然后由判别器网络度量真实图像分布与生成图像分布之间的相似性. 判别器网络通过调整自身参数使其分类面远离生成图像分布, 直到最终输出随机判别结果, 即无法区分生成图像和真实图像. 当真实分布和生成分布之间没有交集时, 使用 Jensen-Shannon (JS) 散度度量概率分布距离的经典 GAN 模型, 由于不能获得稳定的回传梯度信息而难以训练. Radford 等[24] 提出了使用具有卷积和反卷积对称结构的 DCGAN 模型, 加强了 GAN 训练的稳定性, 但仍然使用 JS 散度作为概率分布的距离度量. 与此不同, Arjovsky 等[25] 提出了 Wasserstein GAN 模型, 采用 Wasserstein 距离来度量两个概率分布之间的相似性, 缓解

了 GAN 训练过程中梯度消失的问题. WGAN 模型的损失函数值为生成的图像质量提供了量化标准, 更小的损失值意味着生成的图像更加真实. 此外, 在训练 WGAN 时, 不用小心地平衡生成器网络和判别器网络的训练进程, 而是可以采用先优化判别器网络直到收敛, 然后再更新生成器网络的方法, 以使整个网络更快收敛. 为了能够将生成的补全图像直接用于人脸表情识别, 本文基于 WGAN 建立人脸图像补全网络.

1.2　人脸图像补全

局部遮挡使得人脸图像损失了一部分表情信息, 妨碍了识别算法对表情的推断. 通过对遮挡区域中的图像信息进行估计, 能够尽可能还原缺失的表情信息. 从图像编辑的角度, Ding 等[26] 和 Li 等[27] 使用人脸对称位置上的像素对遮挡部分进行填充, 但补全后的图像不够自然. Zhu 等[28] 使用人脸对称位置上的像素梯度对缺失部分进行泊松编辑, 可以令补全部分的肤色和光照更加自然. 从图像生成的角度, 人脸图像补全可以被形式化为概率分布的学习问题. 每一个像素的取值都可以被认为是在图像概率空间中的一次抽样, 而生成图像的过程则是从所有像素的联合概率分布中进行一次采样. 由于邻近的像素之间存在较强的上下文语义关联, 补全图像需要保持与真实图像一致的身份和表情上下文. Pathak 等[29] 提出了一种基于 CNN 的图像上下文信息编解码网络, 能够联合图像遮挡部分和未遮挡部分来补全图像. Yeh 等[30] 提出了一种针对大范围图像补全问题的 GAN 模型. 通过向生成器网络中增加未遮挡部分的上下文损失和服从训练集分布的先验损失, 该方法能够补全不同遮挡区域中的图像内容. Li 等[6] 提出了一种基于自编码器的生成式人脸补全算法, 通过增加人脸语义对象 (例如五官) 的损失来增强生成图像的真实性. 本文通过优化图像真实性, 上下文相似性和平滑性目标, 控制图像生成网络估计遮挡区域内的像素分布, 从而补全缺失的图像信息.

1.3　用户无关的表情识别

在同类表情的不同用户数据之间往往存在着较大差异, 提取不受这些差异影响的表情特征关系到识别算法的鲁棒性. 一些工作通过对二维图像或三维头部模型进行融合来获得用户无关的表情表示. Chen 等[20] 将身份不同但表情相同的图像进行融合, 得到一种用户无关的表情表示, 弱化了身份特征, 增强了表情特征. Zhu 等[28] 将三维头部模型分解为中性模型, 身份模型和表情模型, 将身份和表情的类内差异通过两种形变模型进行分离, 但没有考虑表情与身份之间的关联. 另一些工作尝试通过稀疏表示来提取用户无关的表情特征. Zafeiriou 等[19] 通过待识别表情图像和相同身份的中性图像之间的差分图像来抑制身份特征, 但只限于能够预先获取当前身份中性图像的情况. Lee 等[31] 为每类表情的图像构造与待识别表情图像具有相似身份的图像, 然后通过在二者之间进行差分来抑制类内差异. 基于稀疏表示的方法对训练表情数据有较高的要求, 并且在数据量较大的训练集上难以直接求解. 受到以上工作和对抗网络的启发, 本文以多任务学习的方式, 在表情识别任务和身份识别任务之间建立一种对抗关系, 使其能够区分表情特征和身份特征, 从而提取到更本质的表情特征.

2　基于 WGAN 的人脸图像补全

本文提出的鲁棒人脸表情识别方法由人脸遮挡图像补全和表情识别两个阶段组成, 如图 1 所示. 1) 训练一个基于 WGAN 的人脸图像生成网络, 对输入图像中由二值掩码矩阵标记的遮挡部分进行补全, 如图 1 中上半部分所示; 2) 训练一个基于 VGG 16[21] 的卷积神经网络对补全图像进行人脸特征提取, 然后采用对抗学习策略, 提取用户无关的表情特征并推断表情类别, 如图 1 中下半部分所示. 本节介绍人脸图像的补全方法, 下一节介绍用户身份抑制的表情识别方法.

2.1　人脸图像补全网络

补全局部遮挡的人脸图像可以转化为保持上下文一致性的图像生成. 首先建立一个能够产生人脸图像的 GAN 网络, 然后使用该模型生成与遮挡图像最相似的图像, 再用它填充遮挡区域. 生成器网络产生与真实图像集最相似的人脸图像, 然后由判别器网络通过 Wasserstein 距离度量生成图像集的真实性. 生成器网络使用核大小为 5 像素 × 5 像素的卷积层对隐变量 $\boldsymbol{z}$ 进行上采样, 将输出通道数逐层缩减为前一层的一半, 同时 feature map 的尺寸扩大为原来的 2 倍. 除第一层卷积外, 在其余各卷积层后增加 Batch Normalization (BN) 层[32] 防止协变量漂移 (Covariate shift). 使用 ReLU[33] 作为各卷积层的激活函数. 判别器网络与生成器网络保持对称结构, 以加快模型参数在对抗训练过程中的收敛速度.

Wasserstein 距离的定义为

$$W(p_r, p_g) = \inf_{\gamma \sim \Pi(p_r, p_g)} \mathrm{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \gamma} \|\boldsymbol{x} - \boldsymbol{y}\|$$

其中, $\Pi(p_r, p_g)$ 是以 p_r 和 p_g 为边缘分布的所有可能的联合概率分布 γ 的集合. $W(p_r, p_g)$ 为 $\gamma(\boldsymbol{x}, \boldsymbol{y})$ 期望的下确界, 表示为了将 p_r 移到 p_g 需要将 $\boldsymbol{x}$ 移动到 $\boldsymbol{y}$ 的距离. 与 JS 散度相比, 即使两个分布之间

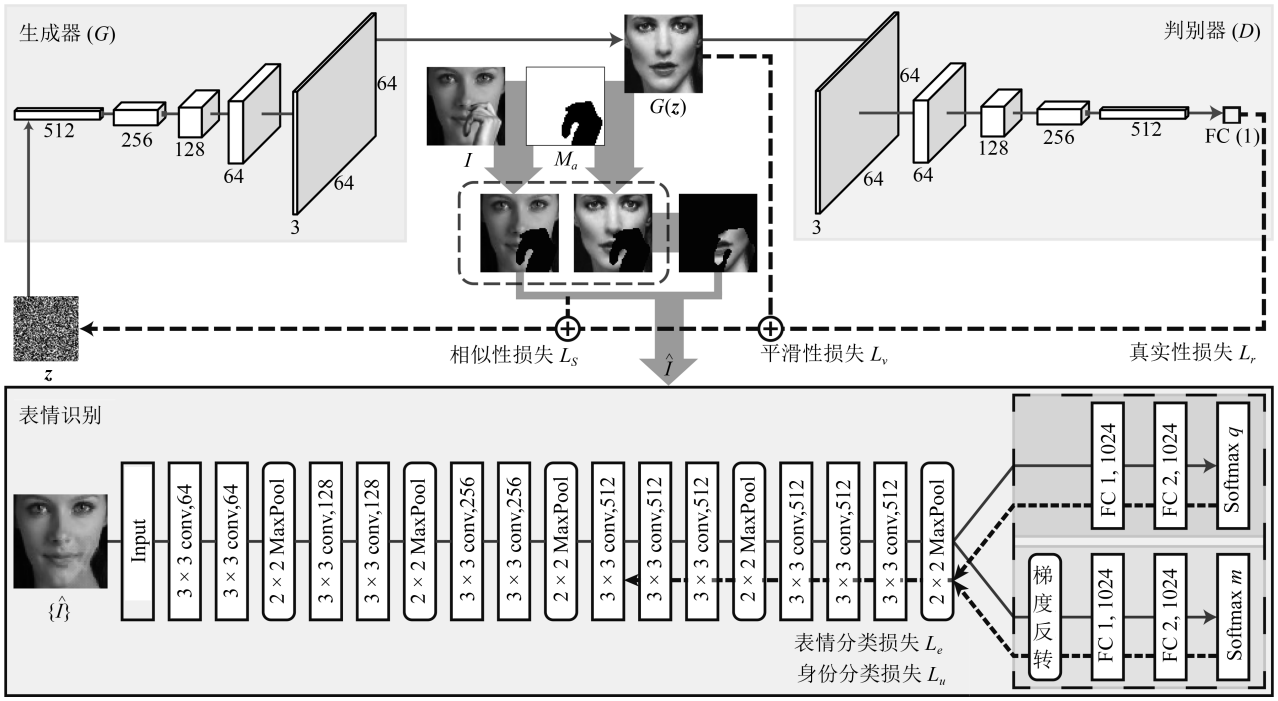


图 1 鲁棒人脸表情识别的算法框架

Fig. 1 Framework of our robust facial expression recognition algorithm

没有交集, Wasserstein 距离也能反映它们之间的相似度, 进而产生有意义的梯度. 直接计算任意分布之间的 Wasserstein 距离比较困难, 故考虑其 Kantorovich-Rubinstein 对偶形式.

$$W(p_r, p_g) = \frac{1}{K} \sup_{\|f\|_L \le K} \mathrm{E}_{\boldsymbol{x}\sim p_r}[f(\boldsymbol{x})] - \mathrm{E}_{\boldsymbol{x}\sim p_g}[f(\boldsymbol{x})]$$

其中, $f(\boldsymbol{x})$ 为任意满足 K-Lipschitz 连续的函数, 即 $f(\boldsymbol{x})$ 导函数的绝对值存在上界. 进一步地, 将 f 定义为由判别器网络参数 θ_D 确定的函数 f_{θ_D}, 并将 $\boldsymbol{x}_{\boldsymbol{x}\sim p_g}$ 写成由生成器网络表示的形式 $g_{\theta_G}(\boldsymbol{z})_{\boldsymbol{z}\sim p_z}$, 可得到判别器网络的优化目标:

$$\max_{\theta_D} \mathrm{E}_{\boldsymbol{x}\sim p_r}\left[f_{\theta_D}(\boldsymbol{x})\right] - \mathrm{E}_{\boldsymbol{z}\sim p_z}\left[f_{\theta_D}(g_{\theta_G}(\boldsymbol{z}))\right] \quad (1)$$

和生成器网络的优化目标:

$$\min_{\theta_G} -\mathrm{E}_{\boldsymbol{z}\sim p_z}\left[f_{\theta_D}(g_{\theta_G}(\boldsymbol{z}))\right] \quad (2)$$

其中, $\boldsymbol{z}$ 表示隐变量, 采样自分布 p_z.

基于 WGAN 训练算法[25] 训练人脸图像补全网络. 生成器网络 g_{θ_G} 和判别器网络 f_{θ_D} 的结构如图 1 所示. 在训练判别器网络时, 首先分别从正态分布 $\mathrm{N}(0,1)$ 和训练集中各采样 b 个样本作为一个批次的训练数据, 然后根据式 (1) 计算判别器网络的损失和梯度更新方向, 使用 RMSProp[34] 优化算法更新梯度. 为了使判别器网络近似满足 Lipschitz 连续性条件, 在判别器网络的参数更新完成之后, 对其梯度进行剪裁, 使之落入一个较小的区间 $[-c, c]$ 中. 在训练生成器网络时, 首先固定判别器网络的参数, 从正态分布 $\mathrm{N}(0,1)$ 中采样 b 个样本作为一个批次的训练数据输入判别器网络, 然后根据式 (2) 计算生成器网络的损失, 同样采用 RMSProp 算法更新它的参数. 由于更好的判别器网络可以反向传播给生成器网络更准确的梯度信息, 因此从训练开始, 在每一次更新生成器网络之前, 均更新判别器网络 K 次, 以使判别器网络更快收敛. 完整的训练过程如算法 1 所示.

算法 1. 人脸图像补全网络的训练算法

输入. $\boldsymbol{z}$: 隐变量; T: 训练数据集; b: 批次大小; η: 学习率; c: 判别器网络的梯度剪裁常数; K: 生成器网络优化过程中的判别器网络更新次数.

输出. θ_D: 判别器网络的参数; θ_G: 生成器网络的参数.

随机初始化 θ_D 和 θ_G
repeat
  for $t = 0, \cdots, K$ **do**
    从分布 $\mathrm{N}(0,1)$ 中采样 b 个样本 $\boldsymbol{z}_i$
    从训练集 T 中采样 b 个样本 $\boldsymbol{x}_i$
    $L_D \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_{\theta_D}(\boldsymbol{x}_i) - \frac{1}{b}\sum_{i=1}^{b} f_{\theta_D}[g_{\theta_G}(\boldsymbol{z}_i)]$
    $\theta_D \leftarrow \theta_D + \eta \times RMSProp(\theta_D, \nabla_{\theta_D} L_D)$
    剪裁 θ_D, 将其限制在 $[-c, c]$ 范围内
  end for
  从分布 $\mathrm{N}(0,1)$ 中采样 b 个样本 $\boldsymbol{z}_i$
  $L_G \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_{\theta_D}[g_{\theta_G}(\boldsymbol{z}_i)]$
  $\theta_G \leftarrow \theta_G - \eta \cdot RMSProp(\theta_G, -\nabla_{\theta_G} L_G)$
until 判别器网络收敛.

2.2　人脸补全算法

建立一个与局部遮挡图像等大的二值掩码矩阵 M_a, 其元素值为 0 表示像素被遮挡, 否则为 1. 为不失一般性, 假设比较准确的矩阵 M_a 能够在图像补全之前被确定. 人脸补全算法通过优化图像真实性, 上下文相似性和平滑性目标来更新图像补全网络的输入隐变量 $\boldsymbol{z}$, 对输入图像中被 M_a 标记的遮挡区域进行图像补全, 如算法 2 所示.

图像真实性约束使得补全人脸能够尽可能接近真实人脸. 将补全图像的真实性损失 L_r 定义为

$$L_r = D(G(\boldsymbol{z};\theta_G);\theta_D)$$

其中, $\boldsymbol{z}\sim \mathrm{N}(0,1)$ 是输入生成器网络的隐变量, $G(\boldsymbol{z};\theta_G)$ 表示由 θ_G 参数化的生成器网络的输出图像, $D(\cdot;\theta_D)$ 表示由 θ_D 参数化的判别器网络的输出, 度量了补全图像与真实图像之间的概率分布距离. 随着判别器网络损失的逐渐降低, 生成图像将逐渐接近训练集中的真实人脸.

图像上下文相似性约束迫使图像补全网络在生成图像空间中搜索与遮挡图像中无遮挡部分最相似的样本来优化输入隐变量, 保持无遮挡部分与补全部分之间的上下文一致性, 最大程度保留身份和表情信息. 将遮挡图像和生成图像中的无遮挡部分之间的相似性损失 L_s 定义为

$$L_s = \delta\left(G(\boldsymbol{z};\theta_G)\odot M_a, I\odot M_a\right)$$

其中, $\delta(\cdot)$ 表示度量矩阵间相似度的函数, 本文取为 L2 范数; I 表示遮挡图像, M_a 是对应的掩码矩阵, $\odot$ 表示元素级乘法运算.

为了使补全图像尽可能平滑, 引入图像的全变差损失 L_v, 其定义如下:

$$L_v = \sum_{(x,y)\in G(\boldsymbol{z};\theta_G)} \left(\nabla_x p_{x,y} + \nabla_y p_{x,y}\right)$$

其中, $p_{x,y}$ 是生成图像 $G(\boldsymbol{z};\theta_G)$ 中 (x,y) 处的像素值, ∇_x 和 ∇_y 是沿 x 方向和 y 方向的梯度.

综上, 总体的优化目标为

$$\min_{\boldsymbol{z}\sim \mathrm{N}(0,1)} (L_s + \lambda_r L_r + \lambda_v L_v) \tag{3}$$

其中, λ_r 和 λ_v 分别表示真实性损失权重和平滑性损失权重.

经过充分训练的生成器网络能够将隐变量空间映射到人脸图像空间. 在补全图像时, 固定生成器网络和判别器网络的参数, 使用遮挡图像数据集按式 (3) 优化隐变量 $\boldsymbol{z}$, 使得生成图像能够尽可能地接近遮挡图像. 最终, 补全人脸图像 $\hat{I}$ 由遮挡图像 I 中的无遮挡部分和生成图像中与原遮挡区域相对应的部分组成, 即

$$\hat{I} = I\odot M_a + G(\hat{\boldsymbol{z}};\theta_G)\odot(1-M_a)$$

其中, $\hat{\boldsymbol{z}}$ 表示经过优化的隐变量 $\boldsymbol{z}$.

算法 2. 人脸图像补全算法

输入. $\boldsymbol{z}$: 隐变量; P: 输入的遮挡图像集; M: 与 P 对应的掩码矩阵集; θ_D: 已训练的判别器网络参数; θ_G: 已训练的生成器网络参数; λ_r: 真实性权重; λ_v: 平滑性权重; η: 学习率; K: 优化隐变量 $\boldsymbol{z}$ 的更新次数.

输出. $\hat{P}$: 补全图像集.

从分布 $\mathrm{N}(0,1)$ 中采样 $|P|$ 个样本 $\boldsymbol{z}_i$
for $k = 0,\cdots,K$ **do**
　$L_{\boldsymbol{z}} \leftarrow L_s + \lambda_r L_r + \lambda_v L_v$
　$\boldsymbol{z}_i \leftarrow \boldsymbol{z}_i - \eta \times RMSProp(\boldsymbol{z}_i, \nabla_{\boldsymbol{z}} L_{\boldsymbol{z}})$
end for
for all $I\in P$, $M_a\in M$ **do**
　$\hat{I} \leftarrow I\odot M_a + G(\hat{\boldsymbol{z}}_i;\theta_G)\odot(1-M_a)$
　$\hat{P} \leftarrow \hat{P}\cup\{\hat{I}\}$
end for.

3　对抗式类内差异抑制的人脸表情识别

理想的情绪识别系统应当是用户无关的. 使用某些用户的表情图像训练得到的识别算法也应该能够很好地用于识别另一些用户的表情. 对于同一类表情, 希望在提取表情特征的过程中减弱用户身份等类内差异, 使同种表情的特征尽可能在分布上接近. 为此, 提出了表情特征和身份特征互相对抗的人脸表情识别模型, 如图 1 下半部分所示.

3.1　识别模型

形式化地, 将表情训练集表示为

$$\{(\boldsymbol{x},\boldsymbol{y}_{exp},\boldsymbol{y}_{id})|\boldsymbol{x}\in X^l, \boldsymbol{y}_{exp}\in E^q, \boldsymbol{y}_{id}\in U^m\}$$

其中, X 表示 l 维输入图像空间, E 表示 q 维表情类别空间, U 表示 m 维身份类别空间. 取一个批次的训练数据 $\{(\boldsymbol{x}_i, y_{exp_i}, y_{id_i})\}_{i=1}^b$, 表情特征提取函数 N_f 从中学习特征向量:

$$\boldsymbol{feat}_i = N_f(\boldsymbol{x}_i;\theta_f)\in F^d$$

其中, b 是批次的大小, F 对应 d 维表情特征空间, θ_f 是网络 N_f 的参数. 令 $T(\boldsymbol{x},\cdot)$ 表示训练集中的图像和表情类别在 $F\otimes E$ 空间上的分布, 则训练集中某一类表情的特征分布为

$$S(\boldsymbol{feat},k) = \{N_f(\boldsymbol{x};\theta_f)|\boldsymbol{x}\sim T(\boldsymbol{x},k), k\in E\}$$

其中, $T(\boldsymbol{x},k)$ 表示第 k 类表情图像的分布, $S(\boldsymbol{feat}, k)$ 表示第 k 类表情在特征空间中的分布.

为了抑制表情特征中的用户身份, 需要使同类表情的特征分布 $S(\boldsymbol{feat},k)$ 在空间 F 中更加集中,

这意味着根据 $S(\boldsymbol{feat}, k)$ 难以区分用户身份. 由于分布 $S(\boldsymbol{feat}, k)$ 未知, 并且随着训练的进行, 样本在 F 空间上的分布不断变化, 因此难以直接衡量分布之间的相似度. 为此, 建立一个表情识别网络和一个身份识别网络, 通过在它们之间进行对抗学习来近似逼近该分布.

具体地, 在表情识别任务中, 表情分类网络 N_e 将特征向量映射为表情类别, 即

$$\hat{y}_{exp_i} = N_e(\boldsymbol{feat}_i; \theta_e)$$

其中, $\hat{y}_{exp_i}$ 表示推断的表情类别, θ_e 是网络 N_e 的参数.

类似地, 身份分类网络 N_u 将特征向量映射为身份类别, 即

$$\hat{y}_{id_i} = N_u(\boldsymbol{feat}_i; \theta_u)$$

其中, $\hat{y}_{id_i}$ 表示推断的身份类别, θ_u 是网络 N_u 的参数.

交叉熵损失衡量预测类别和真实类别之间的距离, 其定义为

$$L_y = -\sum_{i=1}^{N}\Big[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\Big]$$

其中, y_i 和 $\hat{y}_i$ 分别表示真实类别和预测类别进行 one-hot 编码后的第 i 位. 记表情识别的交叉熵损失为 L_e, 通过最小化该损失来优化参数 θ_f 和 θ_e, 使得提取的特征能够有效地识别表情. 该过程可表示为

$$\min_{\theta_f}\min_{\theta_e} L_e(\theta_f, \theta_e)$$

对于身份识别任务, 记身份识别的交叉熵损失为 L_u, 通过调整 θ_u 和 θ_f 与表情识别任务形成对抗关系, 迫使网络 N_u 在分布 $S(\boldsymbol{feat}, k)$ 上难以识别用户身份. 该过程可表示为

$$\max_{\theta_f}\min_{\theta_u} L_u(\theta_f, \theta_u)$$

联合表情识别任务, 为模型设置多任务目标函数, 并使用随机梯度下降 (Stochastic gradient descent, SGD) 算法优化它的参数. 具体地, 目标函数 J 的定义为

$$J(\theta_f, \theta_e, \theta_u) = \sum_{i=1}^{b} L_e(\theta_f, \theta_e) - \lambda \sum_{i=1}^{b} L_u(\theta_f, \theta_u)$$

其中, b 为批次的大小, λ 是平衡表情识别任务和身份识别任务的权重. $J(\cdot)$ 中的第 1 项反映了表情识别任务的损失, 可直接通过 SGD 算法优化; 第 2 项反映了身份识别任务的损失, 它的优化方向与第 1 项相反. 对 $J(\cdot)$ 分别计算关于 θ_f, θ_e 和 θ_u 的梯度, 有以下更新规则.

$$\begin{aligned}
\theta_f &= \theta_f - \eta\frac{\partial L_e}{\partial \theta_f} + \eta\lambda\frac{\partial L_u}{\partial \theta_f} = \\
&\quad \theta_f - \eta\left(\frac{\partial L_e}{\partial \theta_f} - \lambda\frac{\partial L_u}{\partial \theta_f}\right) \\
\theta_e &= \theta_e - \eta\frac{\partial L_e}{\partial \theta_e} \\
\theta_u &= \theta_u - \eta\frac{\partial L_u}{\partial \theta_u}
\end{aligned} \tag{4}$$

其中, η 表示学习率, 控制每次更新的步长. 式 (4) 表示模型以表情识别为目标的同时, 能够尽可能地对身份特征进行抑制.

3.2 模型细节

采用 VGG 16 作为表情特征提取网络 N_f, 使用在 ImageNet[35] 上预训练的参数初始化其前三组卷积层并固定, 保留 VGG 16 对低层视觉特征的感知能力, 然后在训练过程中调优其余参数. 对于表情识别任务, 建立一个能够区分 q 类表情的多层感知器网络 N_e; 对于身份识别任务, 使用另一个能够区分 m 类用户身份的多层感知器网络 N_u; 如图 1 下半部分中的虚线框所示.

在训练时, 由于式 (4) 中存在 θ_f 梯度的负系数, SGD 算法不能直接对表情识别模型进行训练. 为此, 在网络 N_f 和 N_u 之间增加了梯度翻转层 (Gradient reversal layer, GRL)[36]. 对于任意的输入 X 和输出 Y, GRL 对其进行等值前向传播, 即

$$Y = IX$$

和反向缩放传播, 即

$$\frac{\partial L}{\partial X} = -\alpha\frac{\partial L}{\partial Y}$$

其中, L 表示损失函数, I 为单位矩阵, α 为梯度缩放系数.

此外, 在模型训练初期, 网络 N_u 的身份识别能力较弱, 较强的反向传播容易加大噪声. 为此, 在训练过程中逐渐提高身份识别损失 L_u 的权重.

$$\lambda = \frac{2}{1+\exp(-10p)} - 1$$

其中, p 为当前训练轮数占最大训练轮数的比例. 随着训练的进行, λ 的取值逐渐从 0 递增到 1.

4 实验

4.1 实验设置

为了使人脸图像补全网络尽可能学习到真实人脸的图像分布, 在 CelebA[37] 数据集上进行训练. CelebA 数据集收录了 10 177 人的 202 599 张人脸图像, 其训练集、验证集和测试集分别包含 162 770, 19 867 和 19 962 张图像. 为了与表情图像的对齐方法保持一致, 对 CelebA 训练集中的图像使用 Chen 等[20] 提出的基于三维头部模型重建的方法进行人脸对齐, 剪裁出的人脸部分, 缩放到 64 像素 × 64 像素大小, 再将像素值归一化到 $[0,1]$ 之间. 隐变量 $\boldsymbol{z}$ 的维数为 100, 批大小设置为 64, 初始学习率 η 为 0.01, 真实性损失权重 λ_r 为 1.0, 平滑性损失权重 λ_v 为 0.5, 梯度剪裁常数 c 为 0.01.

在训练表情识别网络之前, 首先使用 Chen 等[20] 提出的三维头部模型重建方法和 Zhu 等[28] 提出的人脸归一化方法对表情图像进行预处理. 由于三维刚性变换不能为人脸自遮挡提供额外的图像信息, 非正面表情图像在归一化后可能出现像素值为 0 的自遮挡区域. 使用掩码矩阵对遮挡区域进行标记, 然后对这些区域进行图像补全. 此外, 对训练图像进行数据扩增: 在图像周围 2 像素范围内补 0 并随机裁剪为 64 像素 × 64 像素分辨率大小, 以 0.5 的概率对图像进行随机的左右翻转, 并且随机调整图像的对比度和亮度. 在训练表情识别网络时, 为每个批次的表情图像减去 ImageNet 数据集中所有图像的 RGB 通道均值, 并将输入图像的像素值归一化到 $[0,1]$ 之间. 初始学习率 η 设置为 10^{-4}, 并且每当训练进行到 1 000 和 4 000 轮时, 将学习率衰减为当前值的 0.1 倍; 批大小设置为 128, L2 权值衰减系数设置为 2×10^{-4}, BN 层的滑动平均估计系数设置为 0.9997.

4.2 人脸表情识别

为了评价本文提出的表情识别方法, 基于 CK +[38], Multi-PIE[39] 和 JAFFE[40] 表情数据库构建了一个混合数据集, 增加了表情特征和身份特征的复杂度. CK + 数据库包含非洲、亚洲、欧洲和北美洲的 123 名用户表演的 327 个连续表情片段. 由于连续帧之间的相关性较强, CK + 包含大量时间冗余的数据, 因此将 CK + 中有标注的用户表情视频拆分为 neutral 表情和其他各类表情, 选取峰值强度最大的 2 帧, 排除强度相对较弱的 fear 表情, 得到包含 neutral 表情在内的 7 种表情, 共计 654 张表情图像. Multi-PIE 数据库包含欧美和亚洲的 337 名用户表演的 smile, surprise, squint, disgust 和 scream 表情, 提供了在 13 个间隔 15° 的视角和 19 种不同光照条件下的表情图像. 去除 scream 表情, 选取 $-15°\sim15°$ 之间的 935 张表情图像. JAFFE 数据库提供了包括 neutral 表情在内的 10 名日本女性用户的表情数据, 选取除 fear 表情之外的所有图像. 综上, 混合数据集总共包含 474 名用户的 1 802 张表情图像. 此外, 由于源数据库中的表情类别不一致, 根据表情表达的相似程度, 对混合数据集进行了统一分类, 如表 1 所示.

表 1　混合数据集中的统一表情分类
Table 1　Unified expression categories in the mixed dataset

CK +	Multi-PIE	JAFFE	统一分类
neutral	neutral	NE	NE (中性)
anger	–	AN	AN (愤怒)
contempt, disgust	squint, disgust	DI	DI (厌恶)
happy	smile	HA	HA (高兴)
sadness	–	SA	SA (悲伤)
surprise	surprise	SU	SU (惊讶)
fear	–	FE	–
–	scream	–	–

1) 在混合数据集上对所提表情识别算法进行用户相关 (Person dependent, PD) 和用户无关 (Person independent, PI) 的表情识别实验. 训练集和测试集按照 90 % 和 10 % 的比例进行划分. 在用户无关的实验中, 测试集不包含训练集的用户数据. 每项实验均按照十折交叉验证的方法重复 10 次, 按数据集分别统计最优模型取得的平均识别准确率和各类表情的识别准确率. 表 2 对实验结果进行了汇总. 观察到识别算法在混合数据集和三个源数据库上的平均识别准确率均超过 90 %, 表明算法可以有效地对数据集中的人脸图像进行表情识别. 与用户相关实验相比, 用户无关实验的识别准确率在混合数据集, CK +, Multi-PIE 和 JAFFE 上分别降低了 6.45 %, 6.7 %, 2.97 % 和 5.09 %. 较小的降低幅度表明识别算法对身份信息不敏感, 意味着对用户身份具有较好的鲁棒性. 从各类表情的识别结果中, 发现悲伤表情的识别准确率在 PD 和 PI 实验之间的差距超过了 10 %: 在混合数据集, CK + 和 JAFFE 上分别相差 17.43 %, 23.97 % 和 12.86 %. 一种可能的原因是混合数据集中用户表达的悲伤表情不一致, 并且与中性, 愤怒和厌恶表情非常相似, 导致识别算法在这三种表情之间出现混淆. 计算相应的混淆矩阵 (见表 3), 结果显示悲伤表情容易被识别算法错分为愤怒表情, 而中性表情则容易与其他几种表情混淆.

2) 从原始的 Multi-PIE 数据库中选取了 273 名用户在 $-45°\sim45°$ 范围内标准光照下的 8 714 张表

表 2 在混合数据集上的人脸表情识别准确率 (%)

Table 2 Results of recognition accuracy (%) on the mixed dataset

数据库名称	设置	NE	AN	DI	HA	SA	SU	平均识别率
混合数据集	PD	95.41	98.85	94.12	96.55	100.00	97.54	97.07
	PI	87.78	95.52	93.15	93.91	82.57	90.82	90.62
CK +	PD	100.00	97.64	100.00	100.00	97.37	98.35	99.05
	PI	98.60	96.43	94.37	100.00	73.40	100.00	92.35
Multi-PIE	PD	87.50	–	91.43	95.45	–	95.35	93.10
	PI	85.00	–	86.36	94.34	–	93.94	90.13
JAFFE	PD	97.87	100.00	98.73	99.13	100.00	97.80	98.84
	PI	90.91	100.00	96.77	100.00	87.14	87.88	93.75

表 3 混合数据集上人脸表情识别的混淆矩阵

Table 3 Confusion matrix for the mixed dataset

	NE	AN	DI	HA	SA	SU
NE	**87.77**	1.06	5.32	3.19	0.53	2.13
AN	0	**95.52**	3.14	0	1.35	0
DI	3.65	0	**93.15**	0.04	3.16	0
HA	4.78	0.43	0.43	**93.91**	0.43	0
SA	0.92	11.93	4.59	0	**82.57**	0
SU	2.55	2.04	2.04	1.02	1.53	**90.82**

情图像. 随机选择 90 % 的图像用于训练, 其余用于测试. 对这些图像进行人脸归一化预处理, 然后分别对未补全的图像和补全图像进行表情识别. 两组实验的平均识别准确率如表 4 所示. 从实验结果中可以观察到, 补全人脸图像使得头部旋转角度在 15° 和 30° 条件下非正面表情识别的平均准确率分别提高了 2.89 % 和 2.85 %, 表明人脸图像补全网络能够增强 30° 范围之内的表情识别性能. 随着头部旋转角度逐渐加大, 补全图像对于表情识别准确率的增强作用逐渐减弱 (从 2.89 % 到 −1.13 %), 这可能是因为遮挡范围增大且累积了过多的噪声.

表 4 在 Multi-PIE 上针对不同头部姿态和图像补全设置的人脸表情识别准确率 (%)

Table 4 Comparisons of recognition accuracy (%) between settings of face completion for different head poses on Multi-PIE

	0	±15°	±30°	±45°
w/o 人脸补全	93.24	88.42	86.85	83.33
w/ 人脸补全	93.24	**91.31**	**89.70**	82.20

3) 在 CK + 数据库上识别 7 类基本表情, 并与其他同样使用空间表观信息的识别方法进行对比. 对 CK + 中 118 人的 327 段标注视频按照用户分组, 再按照用户编号由小到大排序, 然后取步长为 10, 采样 10 个测试集和 10 个训练集 (采样剩余图片), 再进行十折交叉验证. 表 5 对采用相同实验协议取得的识别准确率进行了汇总, 结果表明在抑制类内差异的条件下, 我们的识别算法在单帧图像上的识别准确率达到 96.00 %, 超过了除 GCNET 以外的方法. 在不抑制用户身份时, 算法的识别准确率大约降低了 4.5 %, 表明抑制方法对表情识别任务的增益程度.

表 5 比较不同方法在 CK + 表情数据集上的识别准确率

Table 5 Comparisons of average recognition accuracy among different methods on CK +

识别方法	准确率 (%)
LBP-TOP[41]	88.99
MSR[42]	91.40
3DCNN-DAP[43]	92.40
DTAN[44]	91.44
MSCNN[45]	95.54
GCNET[46]	97.93
本文方法 (w/o 类内差异抑制)	**91.48**
本文方法 (w/ 类内差异抑制)	**96.00**

4.3 人脸图像补全的作用

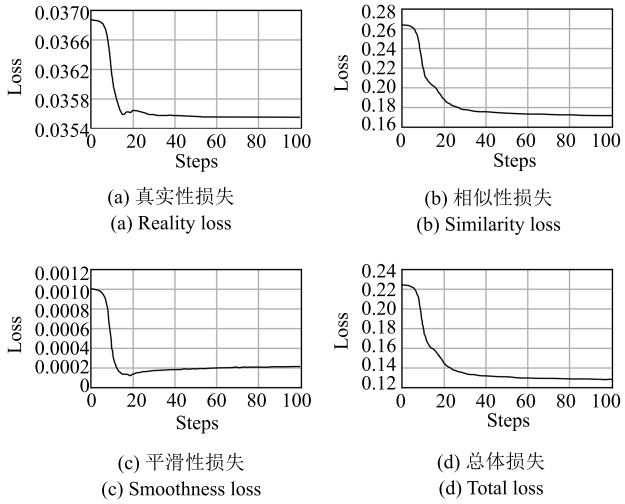
在人工遮挡图像集和自然遮挡图像集上, 对人脸图像补全网络进行了实验评估.

1) 以人工方式对 CelebA 测试集中的人脸图像施加了尺寸相同但位置随机的矩形遮挡. 图 2 展示了人脸图像补全网络在迭代 1, 50, 75 和 100 步之后的输出结果, 各项损失的变化如图 3 所示. 结果表明, 随着优化步数的持续增加, 各项损失均呈现稳定

(a) 输入图像 (a) Input images　(b) 优化 1 次 (b) 1 step　(c) 优化 50 次 (c) 50 steps　(d) 优化 75 次 (d) 75 steps　(e) 优化 100 次 (e) 100 steps　(f) 真实图像 (f) Real images

图 2　人脸补全网络在训练 1, 50, 75 和 100 轮之后的输出图像

Fig. 2　Outputs of the proposed face completion networks after training 1, 50, 75, and 100 steps



(a) 真实性损失 (a) Reality loss　(b) 相似性损失 (b) Similarity loss

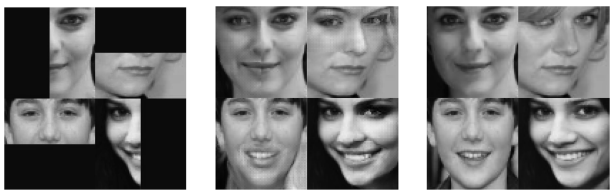
(c) 平滑性损失 (c) Smoothness loss　(d) 总体损失 (d) Total loss

图 3　人脸补全网络的训练损失曲线

Fig. 3　Training loss curves of the proposed face completion networks

下降, 生成的图像质量也在不断提高.

2) 将矩形遮挡的范围扩大至图像的一半, 评估图像补全网络在遮挡范围较大时的补全结果. 较大范围的局部遮挡可能导致五官等人脸物体缺失, 因此增加了图像补全的难度. 如图 4 所示, 被遮挡的一半人脸得到了比较自然的补全, 但同时补全网络也“创造”了一些与原始图像不相似的图像内容.

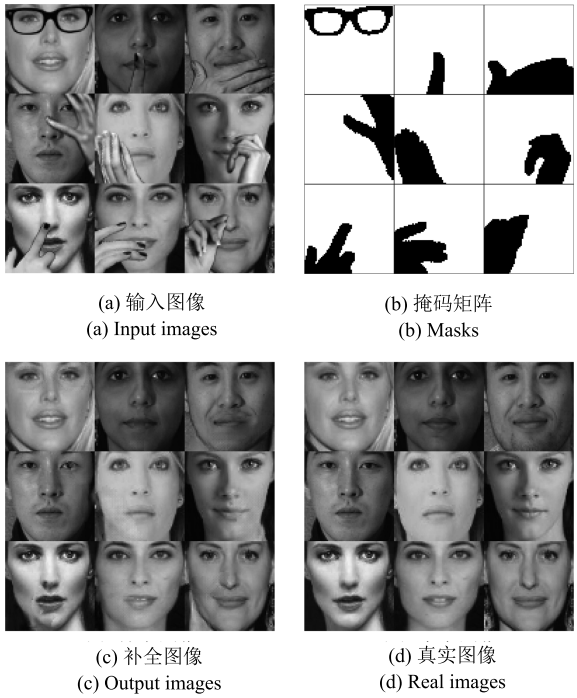


(a) 输入图像 (a) Input images　(b) 补全图像 (b) Output images　(c) 真实图像 (c) Real images

图 4　较大范围局部遮挡图像的补全结果

Fig. 4　Image completion results on images with large-scale occlusion

3) 将一些在用户表达情绪过程中可能出现的肢体动作和佩戴的物品以人工方式附加到真实图像上. 遮挡物体首先被附加到一张与真实图像等大的透明图像上, 计算掩码矩阵, 如 图 5 (b) 所示; 然后, 包含遮挡物的透明图像被附加到真实图像上, 得到输入图像, 如图 5 (a) 所示. 实验结果如图 5 (c) 所示, 表明生成的人脸补全图像能够较好地保持与真实图像一致的上下文信息.



(a) 输入图像 (a) Input images　(b) 掩码矩阵 (b) Masks

(c) 补全图像 (c) Output images　(d) 真实图像 (d) Real images

图 5　不规则局部遮挡图像的补全结果

Fig. 5　Image completion results on images with irregular occlusion

4) 对 COFW[47] 人脸数据库的测试集进行图像补全实验. COFW 测试集由 503 幅在自然条件下采集的人脸图像组成, 包含较大程度的局部遮挡, 遮挡物包括人手, 眼镜和口罩等. 图 6 展示了相应的补全结果, 表明补全网络能够较好地补全像素分布不同于 CelebA 的人脸图像, 这意味着补全网络具有一定的泛化能力. 与图 5 相比, 观察到补全图像的细节损失较多, 尤其是当遮挡物体在人脸产生阴影时.

5) 在 CelebA 测试集上将所提方法与其他方法进行比较, 结果如图 7 所示. 从整体上来看, 由于图

像中的眼睛和鼻子几乎被完全遮挡, 三种方法生成的补全图像与真实图像之间均存在一定差异. 观察局部遮挡区域的补全结果, 本文补全图像中的局部色差比 Yeh 等[30] 的补全图像明显, 因为后者对补全图像使用了泊松编辑. Pathak 等[29] 没有对补全图像进行后处理, 本文补全图像更加平滑和自然.

图 6 在 COFW 测试集上的补全结果

Fig. 6 Image completion results on images taken from the COFW test set

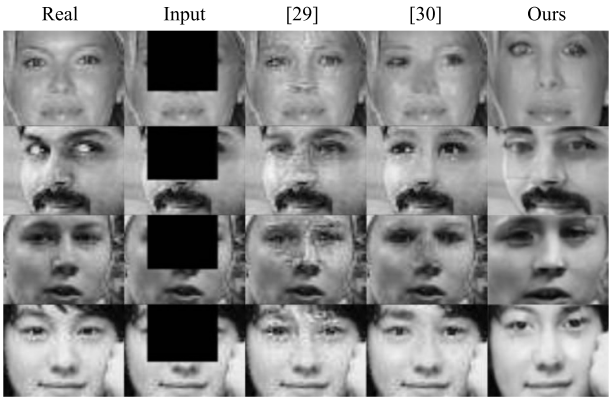
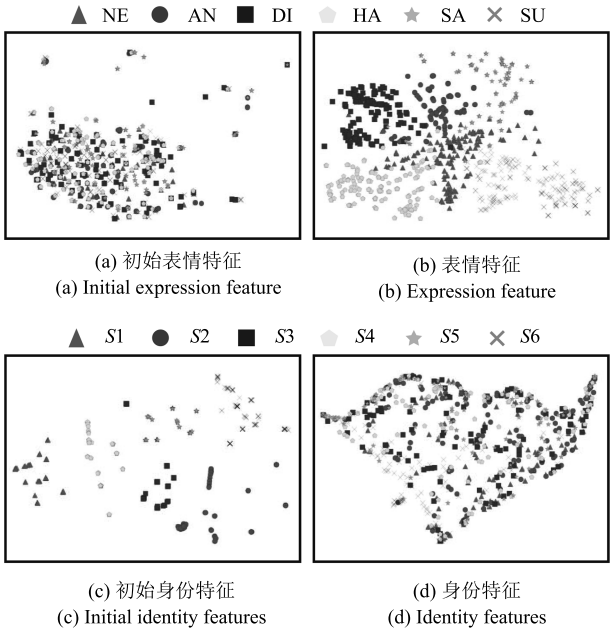


图 7 不同方法补全 CelebA 测试集图像的结果

Fig. 7 Comparisons of different image completion methods on images taken from the CelebA test set

4.4 类内差异抑制的作用

通过抑制类内差异的方法, 表情识别网络提取到一种有效的表情特征. 从表情特征和身份特征的角度, 分别对识别过程中的特征变换进行 t-SNE[48] 可视化分析. 图 8 展示了随机选取的 6 名用户 ($S1 \sim S6$) 的表情图像在识别模型中的特征分布. 其中, 图 8 (a) 和图 8 (b) 显示了表情识别网络中 FC2 层的 t-SNE 可视化结果, 表明表情特征在模型训练过程中逐渐占据支配地位, 直到最后变得几乎完全可分. 另一方面, 图 8 (c) 和图 8 (d) 显示了身份识别网络中 FC2 层的 t-SNE 可视化结果, 反映出用户身份在训练过程中变得越来越不可分, 表明模型在抑制身份信息的同时, 突出了表情特征. 当模型训练完成时, 表情特征已经变得完全可分, 其 t-SNE 可视化结果如图 9 所示.



(a) 初始表情特征
(a) Initial expression feature

(b) 表情特征
(b) Expression feature

(c) 初始身份特征
(c) Initial identity features

(d) 身份特征
(d) Identity features

图 8 人脸表情特征和人脸身份特征的 t-SNE 可视化

Fig. 8 t-SNE visualization of facial expression features and facial identity features

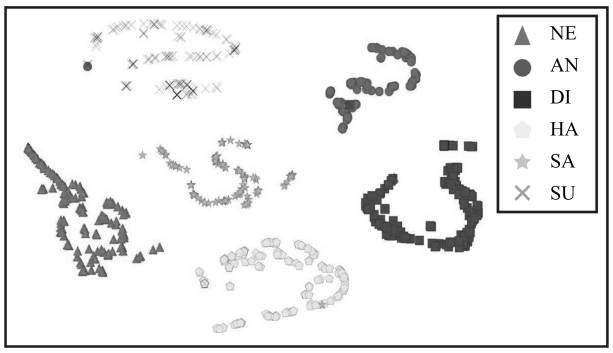


图 9 人脸表情特征的 t-SNE 可视化

Fig. 9 t-SNE visualization of facial expression features

进一步地, 在 CK + 测试集上使用不同方法对学习到的表情特征进行识别, 比较识别准确率, 如表 6 所示. 实验结果表明, 在相同的测试条件下, 提取的表情特征在各类表情上全都取得了比 LPQ[49] 和 SIFT[50] 特征更高识别准确率, 意味着通过对抗学习弱化表情特征中用户身份信息的方法有利于获得鲁棒性较强的表情特征.

5 结论

本文针对具有局部遮挡的人脸表情识别问题, 提出了一种基于 WGAN 的人脸图像补全算法, 能够

表 6 不同特征和识别方法在 CK + 上的识别准确率 (%)
Table 6 Comparisons of different descriptors and methods on CK + (%)

识别方法	AN	DI	HA	SA	SU	平均
LBP + SRC[51]	75.56	87.93	98.55	64.29	100.00	85.26
Gabor + SRC[52]	84.44	98.28	97.10	64.29	98.78	88.58
LPQ + SVM	70.00	65.30	59.31	76.40	88.32	71.87
SIFT + SVM	72.86	56.70	88.75	72.22	96.30	77.37
EFM + SVM	**88.37**	**90.20**	**87.16**	**78.38**	**94.00**	**87.62**

为遮挡区域生成上下文一致的补全图像. WGAN 模型满足 Lipschitz 连续性条件, 因此能够比较稳定地生成更加接近真实图像的结果. 在此基础上, 为提出的补全网络设置了关于图像真实性, 平滑性和上下文相似性的优化目标, 进一步加强补全图像的质量, 使其能够直接用于人脸表情识别. 进一步地, 受到对抗学习的启发, 本文提出了一种对用户身份鲁棒的表情识别方法, 能够在提取表情特征的过程中抑制用户身份信息的影响, 增强表情识别的准确性. 通过在表情识别任务和身份识别任务之间进行对抗训练, 使得识别用户身份的难度不断增大直至不可分, 从而提取到这种鲁棒的表情特征. 通过在由 CK +, Multi-PIE 和 JAFFE 构成的混合数据库上进行的实验, 发现提取的表情特征对用户身份信息不敏感, 能够为 CK + 上用户无关的识别准确率带来大约 4.5 % 的性能提升. 在 Multi-PIE 上识别非正面人脸表情的实验结果表明, 在 45° 头部旋转范围内, 本文方法有助于提升表情识别的鲁棒性和准确性. 此外, 尽管本文主要针对静态图像中的人脸表情识别进行了实验评估, 但本文方法也能够对视频帧进行逐帧表情识别. 由于生成式图像补全及其预处理过程具有较高的计算负荷, 针对视频的表情识别过程可以离线方式进行.

下一步, 我们将沿着识别连续表情的方向进行探索, 考虑如何使用更有效的语义约束提升人脸图像补全的真实性和相似性, 以及如何面向视频中的时序信息进行类内差异抑制.

References

1 Zeng Z H, Pantic M, Roisman G I, Huang T S. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(1): 39−58

2 Valstar M, Pantic M, Patras I. Motion history for facial action detection in video. In: Proceedings of the 2004 IEEE International Conference on Systems, Man, and Cybernetics. The Hague, The Netherlands: IEEE, 2004. 635−640

3 Sandbach G, Zafeiriou S, Pantic M, Yin L J. Static and dynamic 3D facial expression recognition: a comprehensive survey. *Image and Vision Computing*, 2012, **30**(10): 683−697

4 Gunes H, Schuller B. Categorical and dimensional affect analysis in continuous input: current trends and future directions. *Image and Vision Computing*, 2013, **31**(2): 120−136

5 Yang M, Zhang L, Yang J, Zhang D. Robust sparse coding for face recognition. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA: IEEE, 2011. 625−632

6 Li Y J, Liu S F, Yang J M, Yang M H. Generative face completion. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 5892−5900

7 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: ACM, 2012. 1097−1105

8 Yu Z D, Zhang C. Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. Seattle, USA: ACM, 2015. 435−442

9 Kim B K, Dong S Y, Roh J, Kim G, Lee S Y. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 1499−1508

10 Mollahosseini A, Chan D, Mahoor M H. Going deeper in facial expression recognition using deep neural networks. In: Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision. Lake Placid, USA: IEEE, 2016. 1−10

11 Sun Xiao, Pan Ting, Ren Fu-Ji. Facial expression recognition using ROI-KNN deep convolutional neural networks. *Acta Automatica Sinica*, 2016, **42**(6): 883−891
(孙晓, 潘汀, 任福继. 基于 ROI-KNN 卷积神经网络的面部表情识别. 自动化学报, 2016, **42**(6): 883−891)

12 Kan M N, Shan S G, Chang H, Chen X L. Stacked progressive auto-encoders (SPAE) for face recognition across poses. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 1883−1890

13 Zhu Z Y, Luo P, Wang X G, Tang X O. Deep learning identity-preserving face space. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 113−120

14 Yim J, Jung H, Yoo B, Choi C, Park D S, Kim J. Rotating your face using multi-task deep neural network. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 676−684

15 Zhang J, Kan M N, Shan S G, Chen X L. Occlusion-free face alignment: deep regression networks coupled with decorrupt AutoEncoders. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 3428−3437

16 Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: 2014. 2672−2680

17 Wang Kun-Feng, Gou Chao, Duan Yan-Jie, Lin Yi-Lun, Zheng Xin-Hu, Wang Fei-Yue. Generative adversarial networks: the state of the art and beyond. *Acta Automatica Sinica*, 2017, **43**(3): 321−332
(王坤峰, 苟超, 段艳杰, 林懿伦, 郑心湖, 王飞跃. 生成式对抗网络 GAN 的研究进展与展望. 自动化学报, 2017, **43**(3): 321−332)

18 Wright J, Yang A Y, Ganesh A, Sastry S S, Ma Y. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(2): 210−227

19 Zafeiriou S, Petrou M. Sparse representations for facial expressions recognition via $l1$ optimization. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 32−39

20 Chen H, Li J D, Zhang F J, Li Y, Wang H G. 3D model-based continuous emotion recognition. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 1836−1845

21 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [Online], available: https://arxiv.org/abs/1409.1556, February 11, 2018

22 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions [Online], available: https://arxiv.org/abs/1409.4842, February 11, 2018

23 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770−778

24 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [Online], available: https://arxiv.org/abs/1511.06434, February 11, 2018

25 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN [Online], available: https://arxiv.org/abs/1701.07875, February 11, 2018

26 Ding L, Ding X Q, Fang C. Continuous pose normalization for pose-robust face recognition. *Signal Processing Letters*, 2012, **19**(11): 721−724

27 Li S X, Liu X, Chai X J, Zhang H H, Lao S H, Shan S G. Morphable displacement field based image matching for face recognition across pose. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012. 102−115

28 Zhu X Y, Lei Z, Yan J J, Yi D, Li S Z. High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 787−796

29 Pathak D, Krähenbühl P, Donahue J, Darrell T, Efros A A. Context encoders: feature learning by inpainting. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 2536−2544

30 Yeh R A, Chen C, Lim T Y, Schwing A G, Hasegawa-Johnson M, Do M N. Semantic image inpainting with deep generative models [Online], available: https://arxiv.org/abs/1607.07539, February 11, 2018

31 Lee S H, Plataniotis K N, Ro Y M. Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. *IEEE Transactions on Affective Computing*, 2014, **5**(3): 340−351

32 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: IEEE, 2015. 448−456

33 Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA: JMLR, 2011. 315−323

34 Hinton G, Srivastava N, Swersky K. Neural networks for machine learning: overview of mini-batch gradient descent [Online], available: http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, February 11, 2018

35 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, **115**(3): 211−252

36 Ganin Y, Lempitsky V S. Unsupervised domain adaptation by backpropagation. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: IEEE, 2015. 1180−1189

37 Liu Z W, Luo P, Wang X G, Tang X O. Deep learning face attributes in the wild. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 3730−3738

38 Lucey P, Cohn J F, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 94−101

39 Gross R, Matthews I, Cohn J, Kanade T, Baker S. Multi-PIE. *Image and Vision Computing*, 2010, **28**(5): 807−813

40 Lyons M, Akamatsu S, Kamachi M, Gyoba J. Coding facial expressions with gabor wavelets. In: Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition. Nara, Japan: IEEE, 1998. 200−205

41 Zhao G Y, Pietikäinen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**(6): 915−928

42 Ptucha R, Tsagkatakis G, Savakis A. Manifold based sparse representation for robust expression recognition without neutral subtraction. In: Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops. Barcelona, Spain: IEEE, 2011. 2136−2143

43 Liu M Y, Li S X, Shan S G, Wang R P, Chen X L. Deeply learning deformable facial action parts model for dynamic expression analysis. In: Proceedings of the 12th Asian Conference on Computer Vision. Singapore, Singapore: Springer, 2014. 143−157

44 Jung H, Lee S, Yim J, Park S, Kim J. Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 2983−2991

45 Zhang K H, Huang Y Z, Du Y, Wang L. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 2017, **26**(9): 4193−4203

46 Kim Y, Yoo B, Kwak Y, Choi C, Kim J. Deep generative-contrastive networks for facial expression recognition [Online], available: https://arxiv.org/abs/1703.07140, February 11, 2018

47 Burgos-Artizzu X P, Perona P, Dollár P. Robust face landmark estimation under occlusion. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 1513−1520

48 Zhang S Q, Zhao X M, Lei B C. Robust facial expression recognition via compressive sensing. *Sensors*, 2012, **12**(3): 3747−3761

49 Ojansivu V, Heikkilä J. Blur insensitive texture classification using local phase quantization. In: Proceedings of the 3rd International Conference on Image and Signal Processing. Cherbourg-Octeville, France: Springer, 2008. 236−243

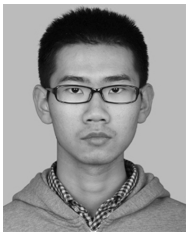
50 Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, **60**(2): 91−110

51 van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, **9**(11): 2579−2605

52 Huang M W, Wang Z W, Ying Z L. A new method for facial expression recognition based on sparse representation plus LBP. In: Proceedings of the 3rd International Congress on Image and Signal Processing. Yantai, China: IEEE, 2010. 1750−1754

姚乃明 中国科学院软件研究所博士研究生. 2011 年获得首都师范大学硕士学位. 主要研究方向为情感交互, 机器学习和计算机视觉.
E-mail: naiming2014@iscas.ac.cn
(**YAO Nai-Ming** Ph. D. candidate at the Institute of Software, Chinese Academy of Sciences. He received his master degree from Capital Normal University in 2011. His research interest covers areas of affective interaction, machine learning, and computer vision.)

郭清沛 中国科学院软件研究所硕士研究生. 2014 年获得华中科技大学学士学位. 主要研究方向为机器学习, 计算机视觉和情感计算.
E-mail: qingpei2014@iscas.ac.cn
(**GUO Qing-Pei** Master student at the Institute of Software, Chinese Academy of Sciences. He received his bachelor degree from Huazhong University of Science and Technology in 2014. His research interest covers machine learning, computer vision, and affective computing.)



乔逢春 中国科学院软件研究所硕士研究生. 2016 年获得北京林业大学学士学位. 主要研究方向为深度学习和图像生成.
E-mail: qiaofengchun16@mails.ucas.ac.cn
(**QIAO Feng-Chun** Master student at the Institute of Software, Chinese Academy of Sciences. He received his bachelor degree from Beijing Forestry University in 2016. His research interest covers areas of deep learning and image synthesis.)

陈 辉 中国科学院软件研究所研究员. 2006 年获得香港中文大学计算机学院博士学位. 主要研究方向为人机交互, 情感交互, 力触觉交互和虚拟现实. 本文通信作者. E-mail: chenhui@iscas.ac.cn
(**CHEN Hui** Professor at the Institute of Software, Chinese Academy of Sciences. She received her Ph. D. degree from the School of Computer Science, The Chinese University of Hong Kong in 2006. Her research interest covers areas of human-computer interaction, affective interaction, haptics, and virtual reality. Corresponding author of this paper.)

王宏安 中国科学院软件研究所研究员. 1998 年获得中国科学院软件研究所博士学位. 主要研究方向为人机交互和实时智能. E-mail: hongan@iscas.ac.cn
(**WANG Hong-An** Professor at the Institute of Software, Chinese Academy of Sciences. He received his Ph. D. degree from the Institute of Software, Chinese Academy of Sciences in 1998. His research interest covers areas of human-computer interaction and real-time intelligence.)