

基于半监督学习的人脸识别反欺骗方法研究

李莉, 曾伟良, 黄永慧, 孙为军

(广东工业大学自动化学院, 广东 广州 510006)

摘要: 鉴别图像中的真伪人脸是一个长期具有挑战性的问题。当合成的伪造人脸十分逼真时, 机器识别难分真假, 甚至肉眼也难以区分。基于监督学习的真伪人脸识别建模往往需要大量的标签样本, 模型的性能严重依赖样本的规模。提出一种基于半监督学习的人脸识别反欺骗方法, 以减少对大量标签样本的依赖。该方法利用图像修复模型来学习人脸图像潜在的数据分布。在训练过程中, 少量标签样本周期性地提供有监督信号来训练分类器, 以区分真伪人脸。该方法可用于不同场景的伪造人脸, 如基于摄像头拍摄的人脸或生成对抗网络生成的人脸。在NUAA数据集和口罩遮挡人脸数据集RMFD上进行验证, 实验结果表明, 所提方法能够在不降低修复图像质量的情况下达到理想的分类精度。仅依靠少量带标签图像, 所提方法比Improved-GAN方法和常用的半监督机器学习方法优势更明显, 并优于支持向量机和卷积神经网络的监督学习方法。

关键词: 人脸反欺骗; 半监督学习; 生成对抗网络; 图像修复

中图分类号: TP183

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202138

Research on anti-spoofing method of face recognition based on semi-supervised learning

LI Li, ZENG Weiliang, HUANG Yonghui, SUN Weijun

School of Automation, Guangdong University of Technology, Guangzhou 510006, China

Abstract: It is a long-term challenge to identify the real and fake faces in the images. When the synthetic fake faces are very realistic, it is difficult for machines and even naked eyes to distinguish the real and fake ones. The supervised anti-spoofing method often requires a large number of labeled samples for a good performance. An anti-spoofing method of face recognition based on semi-supervised learning was proposed to reduce the dependence on massive labeled samples. The method adopted an image inpainting model to learn the data distribution of face images. During the training process, a few labeled samples periodically provided supervised signals to train the classifier to distinguish real faces from fake ones. The proposed method could be used for face anti-spoofing in different scenario, such as faces captured by cameras or generated by generative adversarial net. Accordingly, it was evaluated on the NUAA and RMFD datasets. Experiment results show that the proposed method can keep the quality of restored images, and achieve desirable classification accuracy. With a few labeled samples, the proposed method outperforms Improved-GAN and common semi-supervised methods, and surpasses supervised learning method based on support vector machine and convolutional neural network.

Key words: face anti-spoofing, semi-supervised learning, generative adversarial network, image inpainting

收稿日期: 2021-06-15; 修回日期: 2021-08-07

通信作者: 曾伟良, weiliangzeng@gdut.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61803100); 广东省重点领域研发计划资助项目 (No.2019B010118001)

Foundation Items: The National Natural Science Foundation of China (No.61803100), Key-Area Research and Development Program of Guangdong Province (No.2019B010118001)

1 引言

人脸识别技术在日常生活中已经被应用到各种身份识别系统中,如门禁系统、公共场合的安检系统、屏幕解锁等,保障了信息安全。尽管如此,人脸识别技术仍然存在安全隐患,例如照片、视频或3D打印技术可以高度模拟人脸图像,欺骗识别系统。为了规避这一问题,近年来人脸反欺骗技术受到了广泛的关注。尤其在新型冠状病毒肺炎疫情期间,口罩遮挡的情况给真伪人脸的识别带来了新的挑战。本文主要研究区分真实和伪造人脸图像的反欺骗方法。该类方法根据特征表示的不同通常分为3种类型,分别为基于传统手工特征、深度学习特征和融合特征的方法。

传统手工特征由特定的算子得到,着重区分活体和非活体之间的区别。Li J W等人^[1]通过估计傅里叶频谱中的高频成分判定真伪人脸图像的区别。Boulkenafet Z等人^[2]提出一种基于颜色纹理分析的方法,用于发现真实和伪造人脸图像在RGB色彩空间的纹理差异。Choi J Y等人^[3]利用颜色局部二值模式算子从图像中提取不同颜色空间的联合颜色纹理信息。Bharadwaj S等人^[4]提出基于方向光流直方图的算子,用于同时捕捉动作和纹理特征。

相较于手工设计的浅层特征,深度学习方法提取的深层语义特征通常具有更强的代表性和泛化能力^[5]。Li L等人^[6]利用神经网络将多个卷积层的输出结合作为图像特征。Liu Y J等人^[7]提出的模型使用空间和时间辅助信息,即深度特征和远程光电容积脉搏波描记法(remote photoplethysmographic, rPPG)信号(即心脏脉冲信号),并采用一种新颖的CNN-RNN架构来学习特征表示。

传统手工设计的浅层特征不能表示高维深层的语义信息,而只采用深度学习的人脸反欺骗方法一般会有过拟合、难收敛等问题。因此,越来越多的学者将两种方法进行融合,两种方法可以互相补充和提升。Patel K等人^[8]提出一种基于深度学习特征和面部运动线索(如眨眼)的鲁棒表示方法,用于防止图像和视频欺骗。Raghavendra R等人^[9]提出表示攻击检测算法,同时捕获人脸图像中的局部特征和全局特征。Feng L T等人^[10]提出一种多信息融合的分层神经网络结构,用于融合动态特征和图像纹理特征。

上述监督学习方法均依赖充足的数据样本和

标签,以保证收敛性和最终性能。但是在实际情况中,获得标签的过程可能会耗费大量的人力物力。对于人脸反欺骗任务,一些人脸图像仅靠肉眼难以辨别真伪,获取的标签并不可靠。因此本文提出采用半监督学习的方法应对人脸反欺骗问题。半监督学习方法有以下两点优势:一是在几乎相同的训练条件(相同的有标签数据)下,可以利用大量的无标签数据进一步提高算法性能;二是利用大量的无标签数据,大大减少有标签数据的需求,降低训练耗时。在半监督学习中,利用生成对抗网络(generative adversarial network, GAN)生成模型^[11-15]、提升网络分类性能,成为近几年新的研究热点。

本文提出一种半监督学习方法用于人脸反欺骗,通过基于GAN的图像修复模型,拟合潜在的人脸图像分布,并将少量带标签图像融入图像修复训练过程,以此提升真伪人脸分类性能。实验结果表明,本文提出的半监督学习方法的分类准确率更高。本文的主要贡献概括如下:

(1) 为了减少获取标签的成本,本文提出一种基于GAN的半监督学习方法,用于解决人脸反欺骗问题;

(2) 本文解决了两种情况的人脸图像欺骗,即拍摄人脸和GAN生成的戴口罩人脸,并通过实验验证了所提方法的可行性。

2 相关工作

2.1 半监督学习

半监督学习按照训练方法可被分为自训练、协同训练、无监督预处理3种方法。自训练^[16]是最基本的伪造标签的方法,包括一个单独的监督学习分类器,迭代训练一小部分有标签的数据和之前算法迭代中被贴上伪标记的数据。协同训练^[17-19]是多监督分类器自训练的一种扩展,每一个监督学习分类器都在有标签数据上迭代训练,并将最可信的预测添加到其他分类器的训练集中。基于无监督预处理^[20-22]的方法一般分为利用无标签数据的无监督学习阶段和利用有标签数据的转换阶段。一般来说,无监督学习阶段包括特征提取、聚类^[23]以及初始化学习参数。转换阶段常用的是基于图的标签传播方法,其根据局部网络结构为无标签样本分配标签。在基于无监督预处理的方法的变体^[24-26]中,多种标签传播的方法被提出。利用深度神经网络强大的代表性和判别能力,基于GAN的半监督学习在一些任务

中取得了理想的结果, Salimans T 等人^[27]对 GAN 进行了半监督学习的扩展, 将输出设置为 $Y+1$ 维, 其中前 Y 维对应数据集中的类别, 最后一维表示图像是否来自真实的数据分布。该模型被称为 Improved-GAN, 专门用于半监督学习, 在多个分类任务中取得了理想的结果。Dai Z H 等人^[28]对 GAN 在半监督学习中的应用进行理论分析, 提出为了提高判别器的性能, 生成器所拟合的数据分布应该与实际数据分布互补。

2.2 图像修复

图像修复是对不完整图像的缺失部分进行补全, 从而得到复原图像的一类方法。近年来, 基于 GAN 的图像修复模型得到广泛的研究, 这些模型共享一个类似的框架, 该框架是由 GAN 开发的, 具体来说, 通过生成器修复图像, 并在判别器的指导下提高修复图像的质量。Pathak D 等人^[29]提出了用于图像修复的上下文编码器, 该卷积神经网络 (convolutional neural network, CNN) 可以根据周围环境生成任意图像区域内容, 结合编码器-解码器结构和 GAN, 分别用于图像修复和改善图像质量。Iizuka S 等人^[30]改进了上下文编码器, 将其重构为 3 个部分, 分别是补全网络、全局判别器和局部判别器。与上下文编码器相比, 该模型可以恢复任何形状的掩码图像, 并且能有效地保持一致性。Li Y J 等人^[31]在上下文编码器的基础上引入全局对抗损失, 强调整个图像的真实性, 并保持边界区域的一致性。在上下文编码器的基础上, 其他方法通过使用预训练的分类器、熵损失函数、语义条件等^[32-34]优化图像的一致性和细节。图像修复模型关注的是整个图像的真实性和连贯性以及随机的局部区域, 因此其在捕获数据分布方面比一般的 GAN 更有潜力, 可以提高半监督学习的能力。为了验证这一点, 本文采用 GAN 图像修复模型^[30]进行半监督学习, 有效解决了人脸反欺骗问题。实验结果表明, 所提基于图像修复的半监督学习模型优于 Improved-GAN 模型和其他经典的半监督学习方法。

3 算法

本节介绍基于 GAN 的半监督学习方法如何完成人脸反欺骗任务。其中, 所用 GAN 模型类似于图像修复的 GAN 结构^[30], 分为生成器和判别器两个卷积神经网络。判别器又被分为局部判别器和全局判别器。给定缺失的图像, 生成器根据判别器的

指导进行无监督的训练, 填补缺失的部分, 使补全的图像在视觉上真实, 与缺失前无异。同时, 人脸反欺骗任务被整合到全局判别器的训练过程中, 使全局判别器可使用少量有监督信号来学习识别真实和伪造的人脸图像。

需要明确的是, 本文主要任务是基于图像修复模型进行半监督学习的人脸反欺骗, 即主要目的是对真伪人脸进行分类。本文运用的图像修复模型作为辅助手段服务于训练良好的真伪人脸分类器, 例如在学习图像修复过程中, 模型中的全局判别器捕获图像深层语义信息, 提取潜在的数据分布特征, 用于训练分类器。具体来说, 训练过程中的生成模型不断逼近数据集 (如数据集 NUAA、RMFD) 分布, 而判别器通过不断改进的修复图像来学习数据集中数据的有效特征。此特征能够反映数据集分布, 并提供额外的标签, 从而训练分类器可以对数据集中的真伪人脸进行区分。同时, 修复的图像作为数据集以外的数据, 可以以假乱真, 扰动分类器 (全局判别器), 可以增加人脸分类的难度, 提升准确率。

3.1 生成器

生成器的架构遵循一种编码器和解码器的结构, 与普通 GAN 直接从噪声矢量生成图像不同, 用于图像修复的 GAN 编码器学习到的隐藏表示捕获了未知区域和已知区域之间的更多变化和关系, 可以通过将这些变化和关系输入解码器来填充缺失的图片内容。这种结构先通过降低分辨率来减少内存使用和计算时间, 再进一步生成高分辨率图像。与依赖多个池化层来降低分辨率的架构不同, 本文的网络模型通过步幅卷积将图像的分辨率 (单位面积的像素数目) 降低两次, 每次降低 $1/4$, 该操作对于在缺失区域生成清晰纹理的图像非常重要。此外, 空洞卷积可以通过相同数量的网络参数感知每个像素周围更大的区域, 有利于发掘缺失像素和图像其他部分的关系。解码器主要通过反卷积层将低分辨率图像恢复为原始图像。整体上, 生成网络的输入是存在缺失的 RGB 图像及其二值图像 (1 表示缺失部分像素, 0 表示图像的其他部分), 输出是复原的 RGB 图像。

3.2 判别器

基于 GAN 的对抗学习的性质, 通过训练判别器来区分真实图像和生成图像, 以估计真实数据分布, 并指导生成器生成更真实的图像。与图像生成

任务不同,图像修复不仅要保证填充图像区域的质量,还要关注修复后图像的一致性。因此,判别器又被分为局部判别器和全局判别器。前者用于判断缺失区域中生成的内容是否真实,后者则用于判断生成内容与周围像素环境是否一致。局部判别器的输入是以填充区域为中心的原图像 $1/4$ 大小的区域。但由于局部判别器的局限性,在反向传播过程中,其较难对缺失区域外的图像部分产生直接影响。这容易导致缺失区域边界的像素点明显不一致,造成恢复的图像和原图像不匹配。为了解决这一问题,全局判别器以恢复后的整张图像为输入,同时判断填充内容的真实性和整体图像的一致性。局部判别器和全局判别器的卷积层都使用 5×5 大小的卷积核以及 2×2 的步长降低分辨率,分别得到 1 024 维的向量。最后,将上述两个向量连接为一个 2 048 维的向量,由单个全连接层进行处理,输出一个连续的值。该数值被 Sigmoid 函数映射到 $[0,1]$ 范围内,用于判断图像的来源(真实数据或生成数据)。

本文的目标是将半监督学习与图像修复模型结合用于解决人脸反欺骗问题。全局判别器对整体图像的判断源于其从潜在数据分布学习到的特征表示,因此可以利用全局判别器在真伪人脸数据分布上提取的特征进一步进行学习分类。具体的做法是,在全局判别器的输出上连接额外的子网络。将全局判别器输出的 1 024 维特征向量作为子网络的输入,而子网络的输出表示人脸反欺骗的分类结果,即真实或伪造人脸图像。本文通过一个全连接层将全局判别器的输出转换成一个长度为 2 的向量,并使用 Softmax 层输出真伪人脸图像的概率。本文没有进一步研究局部判别器的输出,因为局部内容对整体图像的分类作用不大。完整的基于图像修复的人脸反欺骗模型如图 1 所示。

3.3 损失函数

本文通过 3 种损失函数对前文中的网络进行训练,分别为作用于生成器的均方误差(mean squared error, MSE)损失函数、同时作用于生成器和判别器的生成对抗损失函数、作用于全局判别器的交叉熵分类损失函数。

3.3.1 均方误差损失

生成器通过将均方误差作为损失函数来更新参数,根据原图像与生成图像, MSE 的损失定义为:

$$L(\mathbf{x}, \mathbf{M}_g) = \|\mathbf{M}_g \odot (G(\mathbf{x}, \mathbf{M}_g) - \mathbf{x})\|^2 \quad (1)$$

其中, $\odot$ 表示像素上的乘法, $\|\cdot\|$ 表示欧几里得范数, $G(\mathbf{x}, \mathbf{M}_g)$ 表示生成器的函数形式, $\mathbf{x}$ 表示输入的缺失部分像素的图像, $\mathbf{M}_g$ 表示与输入图像大小相同的二值图像,缺失部分对应的像素值为 1,其余部分为 0。

3.3.2 生成对抗损失

生成对抗损失即 GAN 损失,常被用于 GAN 生成器和判别器的训练。生成对抗损失函数反映了生成器如何最大限度地欺骗判别器,以及判别器如何最大限度地区分真实和生成的数据分布。该损失函数用于图像修复,可以逐步提高生成器恢复图像的质量以及判别器的预测能力。GAN 损失的优化可以被描述为如下最小最大化问题:

$$\min_G \max_D E_{x \in \mathcal{X}} [\log D(\mathbf{x}, \mathbf{M}_d) + \log(1 - D(G(\mathbf{x}, \mathbf{M}_g), \mathbf{M}_g))] \quad (2)$$

其中, $\mathbf{M}_d$ 为类似于 $\mathbf{M}_g$ 的二值图像,表示输入图像中缺失的部分,被输入判别器中; $D(\mathbf{x}, \mathbf{M}_d)$ 表示判别器的函数形式,包括局部判别器和全局判别器。式(2)中的期望值是根据实际数据分布 x 计算得到

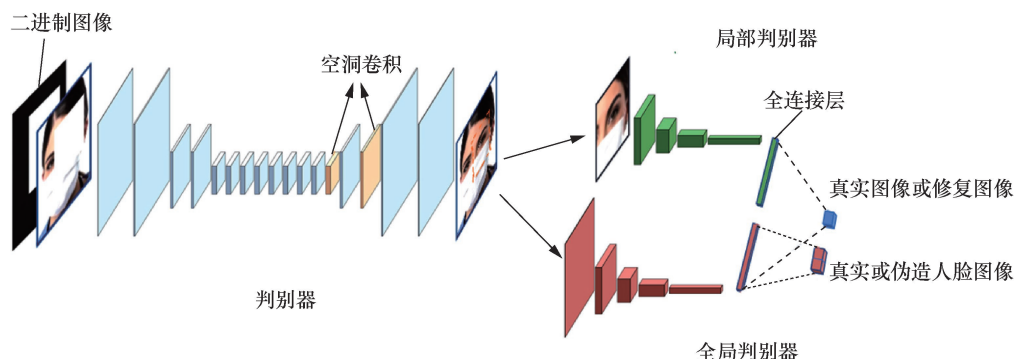


图1 基于图像修复的人脸反欺骗模型

的。由于 MSE 损失在每次迭代中会更新生成器参数，因此其与 GAN 损失可以整体表示为：

$$\min_G \max_D E_{x \in \mathcal{X}} [L(\mathbf{x}, \mathbf{M}_g) + \alpha \log D(\mathbf{x}, \mathbf{M}_d) + \alpha \log(1 - D(G(\mathbf{x}, \mathbf{M}_g), \mathbf{M}_g))] \quad (3)$$

其中， α 表示一个权重超参数，用来衡量不同损失函数的权重，这里 $\alpha = 0.004$ 。由上述损失函数可以得到生成器参数的标准随机梯度为：

$$E[\nabla_{\theta_g} L(\mathbf{x}, \mathbf{M}_g) + \alpha \nabla_{\theta_g} \log(1 - D(G(\mathbf{x}, \mathbf{M}_g), \mathbf{M}_g))] \quad (4)$$

其中， θ_g 表示生成器 G 的参数。在反向传播过程中，将损失函数对参数 θ_g 求导得到的梯度用于更新生成器参数，判别器参数保持不变。同理，还可以使用类似方式更新判别器 D ，梯度可表示为：

$$E[\nabla_{\theta_d} L(\mathbf{x}, \mathbf{M}_g) + \alpha \nabla_{\theta_d} \log(1 - D(G(\mathbf{x}, \mathbf{M}_g), \mathbf{M}_g))] \quad (5)$$

其中， θ_d 表示判别器 D 的参数。此时的梯度被用于更新判别器参数，不会影响生成器。这里的判别器 D 由局部判别器和全局判别器组成，两者同时更新反向传播的梯度。总体来看，式 (4) 和式 (5) 中的梯度用于更新网络时，分别满足了式 (2) 中最小化和最大化的要求。

3.3.3 交叉熵分类损失

本文的目的是使用少量带标签图像完成人脸反欺骗的半监督学习。在图像修复的训练过程中，判别器可以捕捉数据潜在分布，因此具备辅助图像分类的能力。在此基础上，将全局判别器的输出通过一个子网络进行转换（本文采用全连接层和 Softmax 层），转换结果表示真伪人脸的判别结果。全局判别器及其子网络通过少量带标签图像进行监督训练，相应的交叉熵分类损失函数 C 定义如下：

$$C(\mathbf{x}_l, y) = -\log(D'_g(\mathbf{x}_l)[y]), y=0,1 \quad (6)$$

其中， $(\mathbf{x}_l, y)$ 表示有标签的样本， $D'_g(\mathbf{x}_l)$ 表示全局判别器延伸的子网络的输出。 $D'_g(\mathbf{x}_l)[y]$ 表示真实人脸图像和伪造人脸图像的预测概率， $y=0$ 、 $y=1$ 分别表示伪造人脸图像和真实人脸图像。

因此，网络更新所用的最终损失函数如式 (7) 所示。值得注意的是， D'_g 是在 D_g 的基础上添加子网络构建得到的，意味着人脸反欺骗分类器的部分

网络层和图像修复的全局判别器共享结构与参数。通过这种方法，少量带标签图像也可以利用无监督学习得到的特征表达（即全局判别器）训练出良好的分类器。

$$\min_G \max_{D, D'_g} E_{x_l \in \mathcal{X}, x \in \mathcal{X}} [L(\mathbf{x}, \mathbf{M}_g) + \alpha \log D(\mathbf{x}, \mathbf{M}_d) + \alpha \log(1 - D(G(\mathbf{x}, \mathbf{M}_g), \mathbf{M}_g)) + C(\mathbf{x}_l, y)] \quad (7)$$

3.4 神经网络训练及测试

整个神经网络的训练过程被分为如下 3 个阶段。

(1) 首先运用 MSE 损失函数训练生成器模型，此过程中判别器参数固定不变。

(2) 将生成器参数固定，采用无标签图像计算 GAN 损失函数，回传梯度并更新判别器参数，进行无监督学习训练，并在每次训练完判别器后，利用带标签图像计算交叉熵分类损失函数，从而训练全局判别器。

(3) 采用无标签数据，通过 GAN 损失函数对生成器和判别器进行联合训练，先训练判别器再训练生成器，并在每次迭代完成后基于带标签图像用交叉熵分类损失函数训练全局判别器。GAN 损失函数旨在训练判别器，使其可以判别被修复图像是否真实；交叉熵分类损失函数用于训练全局判别器，使得全局判别器具有区分真伪人脸图像的能力，即图像分类的能力。

测试过程中，可以通过训练得到的生成器对输入的缺失图像进行补全修复，同时可以利用全局判别器对输入的完整图像进行真伪判别。

在优化过程中，使用 ADADELTA 算法^[35]自适应地为网络中的每个权值设置学习率。

4 实验结果

4.1 数据集

本文方法首先在 NUAA 人脸反欺骗数据集上进行验证。该数据集使用通用的网络摄像头对实验对象及其照片进行拍照，得到真实和伪造人脸图像。图像反映了人脸识别系统中常见的各种外观变化（如性别、光照、戴/不戴眼镜），并通过处理得到较高的分辨率，逼近真实人脸图像。该数据集模拟了基于摄像头等电子设备产生的伪造人脸的欺骗手段。

此外，考虑到在新型冠状病毒肺炎疫情期期间，口罩遮挡对人脸反欺骗任务造成了新的困难，本文进一步验证了所提方法在这种情况下效果。RMFD (real-world masked face dataset) 是武汉大学

免费开放的全球首个口罩遮挡的人脸数据集。该数据集包含了 525 个受试者的近 5 000 张真实口罩遮挡人脸图像。实验中通过网络爬虫进一步补充口罩遮挡人脸图像，共收集到 5 600 张图像。经过预处理，这些图像中包含的复杂背景被去除，只保留口罩遮挡的人脸区域，如图 2 所示。



图2 口罩遮挡人脸图像

此数据集中的伪造人脸图像由上述口罩遮挡人脸图像通过深度卷积生成对抗网络 (deep convolutional generative adversarial network, DCGAN) 模型^[36]生成，从而模拟基于深度学习模型产生的伪造人脸图像的欺骗手段。

最后，本文在更新的 CelebA-Spoof 大型人脸反欺骗数据集上进行了实验。该数据集拥有 10 177 个受试者的 625 537 张图像，其中包括面部、照明、环境和欺骗类型等 43 个丰富的属性。在 43 个属性中，40 个属性属于真实图像，包括所有面部器官（如鼻子、眼睛、嘴唇）、头发、帽子和眼镜等；3 个属性属于欺骗图像，包括欺骗类型、环境和照明条件等。

4.2 实验装置

4.2.1 对比实验

对比实验中使用了部分监督学习方法和本文提出的半监督学习方法，展示了在有限带标签样本情况下本文方法的优势。具体的半监督学习方法包括如下几种。

(1) 标记传播 (label propagation, LP)^[37]，一种基于样本间图关系的半监督学习方法^[38]，根据标签样本点预测未标记样本点。

(2) 标记扩散 (label spreading, LS)^[37]，一种类似 LP 的图网络方法，因为其最小化了具有正则化属性的损失函数，所以更适用于噪声数据。

(3) K -means+支持向量机 (support vector machine, SVM)，利用 K -means^[39]对训练集进行聚类，得到带标签样本，然后训练 SVM 作为分类器；

(4) 基于 Improved-GAN^[29]模型的半监督学习方法，在多个分类任务上取得了较好的效果。

(5) SHOT-VAE^[40]模型为最近提出的基于变分自动编码器的半监督学习方法，通过改进证据下界

(evidence lower bound, ELBO) 损失函数，模型更加注重学习图像分类，并在常用数据集上取得了目前最优的分类结果。

对比的监督学习方法包括如下几种。

(1) SVM^[41]，一种被广泛应用且通常有效的机器学习模型。

(2) CNN，即本文图像修复模型中的全局判别器。在训练过程中，全局判别器采用交叉熵分类损失函数进行迭代训练，具有较强的分类能力，其网络结构见表 1。这里，CNN 的结构与全局判别器保持一致。在此条件下，将本文方法与通过监督学习训练得到的分类器进行对比，更能突出本文的半监督学习方法在带标签样本较少的情况下的优势。

表1 全局判别器的网络结构

网络层类型	卷积核	步长	通道数量/个
卷积层	5×5	2×2	64
卷积层	5×5	2×2	128
卷积层	5×5	2×2	256
卷积层	5×5	2×2	512
卷积层	5×5	2×2	512
卷积层	5×5	2×2	512
全连接层	—	—	1 024

4.2.2 实现细节

NUAA 数据集包含 5 105 张真实人脸图像和 7 509 张伪造人脸图像。本文实验分别将真实和伪造人脸图像按照 8:2 进行分组，作为监督学习的训练集和测试集。对于半监督学习，在训练集中随机选取 100 张、200 张和 300 张带标签图像，其中真实和伪造人脸图像各一半，剩余图像均假设没有标签，用于图像修复的半监督学习。对于 RMFD，本文实验分别采用 5 000 张真实和伪造人脸图像。与 NUAA 数据集相似，监督学习中的训练集与测试集比例也设置为 8:2。为了更好地展示本文方法的优势，在两个数据集中，只选用带标签图像训练监督学习模型，减少无标签图像的辅助功能。

在训练过程中，在两个数据集上每种方法的参数设置相同。LP 方法和 LS 方法分别选用径向基函数 (radial basis function, RBF) 和 K 近邻 (K -nearest neighbor, KNN) 核函数。LP 方法的最大迭代次数设置为 1 000。LS 方法的最大迭代次数设置为 100，夹持因子 (clamping factor) 设置为 0.2。SVM 方法的正则化系数 C 设置为 1，核函数默认设置为 RBF。

Improved-GAN 方法的批大小设置为 64, 迭代次数设置为 100, 其余参数均与参考文献[30]保持一致。本文方法的训练过程分为 3 个阶段, 对应的迭代次数分别为 10 000、20 000 和 100 000, 针对 MSE 损失函数、GAN 损失函数和交叉熵分类损失函数, 图片批大小分别设置为 16, ADADELTA 优化算法保持 PyTorch 默认设置。对于卷积神经网络的监督学习方法, 图片批大小设置为 20, 迭代次数设置为 100, 同样使用 ADADELTA 优化算法。对于本文方法训练的第三阶段, 仅对 GAN 损失函数使用权重控制, 此权重设置参考了图像修复模型^[31]中的参数, 即 0.004。此外, 本文方法也对不同的权重值进行了测试, 见表 2。

表 2 NUAA、RMFD 两个数据集上不同权重值的测试结果

权重值	NUAA	RMFD
1	89.49%	86.95%
0.4	89.14%	87.05%
0.04	88.63%	86.65%
0.004	88.43%	87.80%
0.000 4	89.21%	86.35%
0.000 04	88.82%	86.30%

由表 2 可知, 测试的 6 个权重值 (呈递减顺序排列) 在 NUAA、RMFD 两种数据集上没有明显规律, 且没有任何一个数值同时在两个数据集上表现最好, 因此后续实验依旧采用图像修复模型^[26]中的原有参数。

4.2.3 结果与分析

本文方法基于图像修复模型, 且训练分为 3 个阶段, 本文方法在 NUAA 和 RMFD 两个数据集上训练后验证得到的修复图像分别如图 3~图 5 和图 6~图 8 所示。由图 3~图 8 可以看到, 3 个阶段对应的修复图像在视觉上越来越逼近真实人脸图像。第一阶段仅通过 MSE 损失对生成器进行训练, 填充区域遗漏较多细节, 修复后的图像在视觉上模糊, 如图 3、图 6 所示; 第二阶段仅通过 GAN 损失训练判别器, 修复的图像在一定程度上可以描述比较细节的面部特征, 提高了图像整体质量, 如图 4、图 7 所示; 第三阶段同时使用 MSE 和 GAN 损失更新生成器和判别器, 此时修复的图像比前两个阶段修复的图像更真实, 如图 5、图 8 所示。

在本文实验中, 人脸遮挡部分的大小是随机的, 即从图像上随机挑选一点作为顶点, 挖取一个



图 3 NUAA 数据集第一阶段训练后得到的修复图像

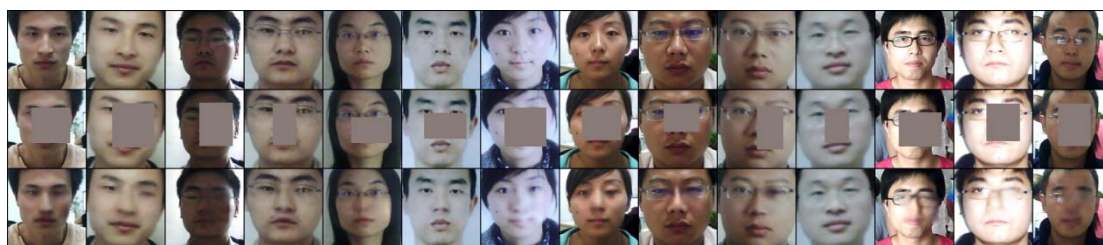


图 4 NUAA 数据集第二阶段训练后得到的修复图像



图 5 NUAA 数据集第三阶段训练后得到的修复图像



图6 RMFD 第一阶段训练后得到的修复图像



图7 RMFD 第二阶段训练后得到的修复图像



图8 RMFD 第三阶段训练后得到的修复图像

长方形，其长和宽均为(96, 128)区间中的一个随机数。这样做的目的是证明图像修复模型对图像缺失部分的大小和位置不敏感，具有较强的图像修复能力，避免仅对某一位置修复效果较好。此遮挡方法被广泛应用于图像修复领域^[26]。

以上渐进的图像修复质量表明生成器不断生成逼近真实图像的图像，且判别器能够逐渐学习到数据集分布的特征表达，有利于半监督学习的展开。因此实验在第二和第三阶段的训练中加入有监督学习过程，利用少量带标签图像通过交叉熵分类

损失更新全局判别器，从而学习区分真伪人脸图像。下面对所提方法和一些现有方法进行对比。

半监督学习分类结果分别见表3和表4。由表3和表4可以看出，同样的方法在RMFD上的分类准确率明显低于NUAA数据集。一方面，与NUAA数据集相比，RMFD中的人脸图像均存在口罩遮挡，大量的关键面部特征不能利用，如鼻子、嘴巴和下巴等；另一方面，RMFD中的伪造人脸图像由DCGAN生成，在噪声有限的情况下可以最大限度地模拟真实人脸图像特征，因此较难区分。

表3 NUA A数据集上半监督学习对不同数量的带标签图像的分类准确率

方法	100 张	200 张	300 张
LS(RBF)	59.53%	59.53%	59.53%
LS(KNN)	71.90%	80.74%	86.68%
LP(RBF)	50.00%	50.00%	50.00%
LP(KNN)	59.65%	59.89%	60.76%
K-means+SVM	92.91%	94.54%	97.82%
Improved-GAN	94.49%	96.07%	96.27%
SHOT-VAE	95.88%	96.12%	98.80%
本文方法	98.01%	98.89%	99.21%

表4 RMFD上半监督学习对不同数量的带标签图像的分类准确率

方法	100 张	200 张	300 张
LS(RBF)	49.34%	49.34%	49.34%
LS(KNN)	50.27%	51.80%	53.30%
LP(RBF)	40.47%	40.47%	40.47%
LP(KNN)	50.00%	51.40%	51.60%
K-means+SVM	60.52%	61.08%	62.78%
Improved-GAN	67.14%	69.75%	73.19%
SHOT-VAE	65.05%	67.10%	69.96%
本文方法	62.27%	74.55%	83.05%

由表3和表4还可以看出,在NUAA和RMFD两个数据集上,本文方法整体上可以得到比其他半监督学习方法更高的准确率。这表明本文方法可应对不同伪造人脸的反欺骗任务,且仅依靠少量的带标签图像。使用KNN核函数时,LP和LS的性能总是优于其使用RBF核函数的性能。RBF核函数下的分类准确率不会随着带标签图像数量的增加而变化,实验中参数的调整也并未带来结果上的变化。反之,使用KNN核函数的LP和LS的分类准确率会随着带标签样本数量的增加不断提升。然而,LP和LS相较于其他半监督学习方法仍较差,尤其在RMFD上,使用KNN核函数时的准确率仅略高于50%。相比之下,K-means方法通过对训练集进行聚类,收集质心作为代表样本。由这些样本训练的SVM分类器得到了比LP和LS更好的结果。但是上述方法的结果比深度学习模型差,没有提取出有代表性的特征进行分类。对于深度学习方法,本文方法在两个数据集上整体优于半监督学习方法Improved-GAN和SHOT-VAE。具体来说,在NUAA

数据集上,本文方法利用不同数量的带标签图像均得到较高的准确率;在RMFD上,本文方法利用200和300张带标签图像时,分类准确率明显高于另外两种方法。这可能是由于图像修复模型中的全局判别器能够更好地学习数据的深度表征,拟合数据的潜在分布,从而提高少量带标签图像情况下的分类性能。

本文实验也对比了少量带标签训练图像情况下所提半监督学习方法和一般的监督学习方法(即SVM和CNN)。实验中分别选用100、200、300张带标签图像作为训练集,并在训练完成后,使用相同的测试集评估各模型。测试的结果见表5和表6。由表5和表6可以看出,在提供少量带标签图像的情况下,本文的半监督学习方法优于其他两种方法,且在RMFD上的优势更加明显。这再次验证了作为一种半监督学习方法,本文方法可以充分利用少量带标签图像来学习图像特征和估计数据集的潜在分布,从而提高分类性能。同时可以发现,当带标签图像越来越多时,本文方法的分类准确率相较于监督学习方法虽有提升,但提升不多,这说明本文方法更加适用于少量带标签图像的情况。

表5 NUA A数据集监督学习的分类准确率

方法	100 张带标签图像	200 张带标签图像	300 张带标签图像	所有训练集
SVM	87.36%	89.38%	95.01%	99.80%
CNN	96.55%	98.65%	98.53%	99.96%
本文方法	98.01%	98.89%	99.21%	—

表6 RMFD监督学习的分类准确率

方法	100 张带标签图像	200 张带标签图像	300 张带标签图像	所有训练集
SVM	53.95%	58.20%	62.50%	67.45%
CNN	51.35%	54.80%	59.36%	87.40%
本文方法	69.27%	74.55%	83.05%	—

在上述两个数据集中,本文方法的有效性得到了证实。此外,在人脸反欺骗数据集CelebA-Spoof上进行实验。实验选择10 000张图像样本,其中真实人脸图像样本5 000张,欺骗人脸图像样本5 000张。半监督学习中带标签图像分别为100、200和300张。在CelebA-Spoof数据集上,本实验仅选择准确率较高且方法较新的Improved-GAN半监督学习方法和SHOT-VAE半监督学习方法与本文方法进行比较。实验结果见表7。

表7 CelebA-Spoof数据集上半监督学习
对不同数量带标签图像的分类准确率

	100 张	200 张	300 张
Improved-GAN	61.09%	62.34%	63.96%
SHOT-VAE	58.90%	59.27%	64.15%
本文方法	60.47%	62.96%	64.57%

由表7的实验结果可知,在选用 CelebA-Spoof数据集进行半监督实验时,与 Improved-GAN 和 SHOT-VAE 方法相比,在 100、200、300 张带标签图像的情况下,本文方法的分类准确率依然较好,这再次证明了本文方法的有效性。

5 结束语

本文提出一种半监督学习方法,用于解决人脸反欺骗问题。采用基于 GAN 的图像修复模型,在其训练过程中融合监督学习。在 NUAA 和 RMFD 两种数据集上对本文方法进行了验证,实验结果表明,在不影响修复图像质量的情况下,本文方法可以仅通过少量带标签图像获得较高的真伪人脸图像分类准确率,优于常用的半监督学习和监督学习方法。未来的研究主要集中于以下两个方向:(1)通过数据增强或发掘代表性样本的标签进一步提高半监督学习方法的性能;(2)将现有方法拓展并应用于基于视频或连续帧的人脸反欺骗任务,通过实时的图像修复学习动态特征。

参考文献:

- [1] LI J W, WANG Y H, TAN T N, et al. Live face detection based on the analysis of Fourier spectra[C]//Proceedings of SPIE. [S.l.:s.n.], 2004: 296-303.
- [2] BOULKENAFET Z, KOMULAINEN J, HADIDA A. Face anti-spoofing based on color texture analysis[C]//Proceedings of 2015 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2015: 2636-2640.
- [3] CHOI J Y, PLATANOTIS K N, RO Y M. Using colour local binary pattern features for face recognition[C]//Proceedings of 2010 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2010: 4541-4544.
- [4] BHARADWAJ S, DHAMECHA T I, VATSA M, et al. Face anti-spoofing via motion magnification and multifeature videolet aggregation[R]. 2014.
- [5] 张钹. 人工智能进入后深度学习时代[J]. 智能科学与技术学报, 2019, 1(1): 4-6.
ZHANG B. Artificial intelligence is entering the post deep-learning era[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 4-6.
- [6] LI L, FENG X Y, BOULKENAFET Z, et al. An original face anti-spoofing approach using partial convolutional neural network[C]//Proceedings of 2016 6th International Conference on Image Processing Theory, Tools and Applications. Piscataway: IEEE Press, 2016: 1-6.
- [7] LIU Y J, JOURABLOO A, LIU X M. Learning deep models for face anti-spoofing: binary or auxiliary supervision[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 389-398.
- [8] PATEL K, HAN H, JAIN A K. Cross-database face antispoofing with robust feature representation[C]//Proceedings of 2016 Chinese Conference on Biometric Recognition. Heidelberg: Springer, 2016: 611-619.
- [9] RAGHAVENDRA R, BUSCH C. Novel presentation attack detection algorithm for face recognition system: application to 3D face mask attack[C]//Proceedings of 2014 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2014: 323-327.
- [10] FENG L T, PO L M, LI Y M, et al. Integration of image quality and motion cues for face anti-spoofing: a neural network approach[J]. Journal of Visual Communication and Image Representation, 2016, 38: 451-460.
- [11] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.
- [12] SPRINGENBERG J T. Unsupervised and semi-supervised learning with categorical generative adversarial networks[J]. arXiv preprint, 2015, arXiv:1511.06390.
- [13] SUTSKEVER I, JOZEFOWICZ R, GREGOR K, et al. Towards principled unsupervised learning[J]. arXiv preprint, 2015, arXiv: 1511.06440.
- [14] ODENA A. Semi-supervised learning with generative adversarial networks[J]. arXiv preprint, 2016, arXiv:1606.01583.
- [15] 余翀. 基于半监督生成对抗网络的三维重建云工作室[J]. 智能科学与技术学报, 2019, 1(1): 70-82.
YU C. Three-dimensional reconstruction cloud studio based on semi-supervised generative adversarial networks[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 70-82.
- [16] TRIGUERO I, GARCÍA S, HERRERA F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study[J]. Knowledge and Information Systems, 2015, 42(2): 245-284.
- [17] WANG W, ZHOU Z H. A new analysis of co-training[C]//Proceedings of the 27th International Conference on Machine Learning. [S.l.:s.n.], 2010.
- [18] DU J, LING C X, ZHOU Z H. When does cotraining work in real data?[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(5): 788-799.
- [19] CHEN M, WEINBERGER K. Q, CHEN Y. Automatic feature decomposition for single view co-training[C]//Proceedings of the 27th International Conference on Machine Learning. [S.l.:s.n.], 2011.
- [20] SHEIKHPOUR R, SARRAM M A, GHARAGHANI S, et al. A survey on semi-supervised feature selection methods[J]. Pattern Recognition, 2017, 64: 141-158.
- [21] GOLDBERG A B, ZHU X J, SINGH A, et al. Multi-manifold semi-supervised learning[C]//Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. [S.l.:s.n.], 2009: 169-176.
- [22] ERHAN D, BENGIO Y, COURVILLE A, et al. Why does unsuper-

- vised pre-training help deep learning[C]//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. [S.l.:s.n], 2010: 201-208.
- [23] 王晓霞, 蒋伏松, 王宇, 等. 基于 ICD-10 诊断编码的慢性病并发症聚类算法[J]. 大数据, 2018, 4(3): 37-45.
WANG X X, JIANG F S, WANG Y, et al. Chronic disease complications clustering based on ICD-10 diagnoses code[J]. Big Data Research, 2018, 4(3): 37-45.
- [24] LIU W, HE J F, CHANG S F. Large graph construction for scalable semi-supervised learning[C]//Proceedings of the 27th International Conference on Machine Learning. [S.l.:s.n], 2010: 679-686.
- [25] DHILLON P S, KEERTHI S S, BELLARE K, et al. Deterministic annealing for semi-supervised structured output learning[C]//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. [S.l.:s.n], 2012: 299-307.
- [26] BREVE F, ZHAO L, QUILES M, et al. Particle competition and co-operation in networks for semi-supervised learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(9): 1686-1698.
- [27] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training GANs[C]//Proceedings of the 30th Conference on Neural Information Processing Systems. [S.l.:s.n], 2016: 2234-2242.
- [28] DAI Z H, YANG Z L, YANG F, et al. Good semi-supervised learning that requires a bad GAN[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 6510-6520.
- [29] PATHAK D, KRÄHENBÜHL P, DONAHUE J, et al. Context encoders: feature learning by inpainting[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2536-2544.
- [30] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics, 2017, 36(4): 1-14.
- [31] LI Y J, LIU S F, YANG J M, et al. Generative face completion[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5892-5900.
- [32] KERZHNER Y, ELAD A, ROMANO Y. Image inpainting using pre-trained classification CNN[C]//Proceedings of 2018 International Conference on Sustainable Energy Engineering. [S.l.:s.n], 2018: 1-9.
- [33] ALTINEL F, OZAY M, OKATANI T. Deep structured energy-based image inpainting[C]//Proceedings of 2018 24th International Conference on Pattern Recognition. Piscataway: IEEE Press, 2018: 423-428.
- [34] LAHIRI A, JAIN A, BISWAS P K, et al. Improving consistency and correctness of sequence inpainting using semantically guided generative adversarial network[J]. arXiv preprint, 2017, arXiv:1711.06106.
- [35] ZEILER M D. ADADELTA: an adaptive learning rate method[J]. arXiv preprint, 2012, arXiv:1212.5701.
- [36] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint, 2015, arXiv:1511.06434.
- [37] YOSHUA B, OLIVIER D, ROUX N L. Label propagation and quadratic criterion[M]//Semi-supervised learning. Cambridge: The MIT Press, 2006: 192-216.
- [38] WANG F, ZHANG C S. Label propagation through linear neighborhoods[J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(1): 55-67.
- [39] HARTIGAN J A, WONG M A. Algorithm AS 136: a K -means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100-108.
- [40] FENG H Z, KONG K Z, CHEN M H, et al. SHOT-VAE: semi-supervised deep generative models with label-aware ELBO approximations[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 7413-7421.
- [41] SAUNDERS C, STITSON M O, WESTON J, et al. Support vector machine[J]. Computer Science, 2002, 1(4):1-28.

[作者简介]



李莉 (1993—), 女, 广东工业大学自动化学院研究助理, 主要研究方向为人工智能、图像处理。



曾伟良 (1986—), 男, 博士, 广东工业大学自动化学院副教授, 主要研究方向为人工智能和物联网系统。



黄永慧 (1975—) 女, 广东工业大学自动化学院讲师, 主要研究方向为数据科学。



孙为军 (1975—), 男, 博士, 广东工业大学自动化学院副教授, 主要研究方向为人工智能和工业互联网系统。

Chinese Journal of Intelligent Science and Technology

Vol.3 No.3 September 2021

Contents

Review Intelligence

Parallel philosophy and intelligent technology: dual equations and testing systems for parallel industries and smart societies

.....WANG Fei-Yue (245)

Special Issue: Intelligent Object Detection and Recognition

Retinal multi-disease screening and recognition method based on deep convolution ensemble network.....

.....WANG Heyang, YANG Qiming, ZHU Qi (259)

3D convolution-based image sequence feature extraction and self-attention for license plate recognition method

.....ZENG Ganxiong, KE Xiao (268)

A method based on visual saliency for vehicle-mounted monocular camera ego-motion estimation and vehicle scale estimation

.....AI Mingxin, LIU Tie, WANG Jing, DING Jiali, YUAN Zejian, SHANG Yuanyuan (280)

Anchor free multispectral pedestrian detection algorithm based on differential feature attention mechanism

.....SHEN Jifeng, LIU Yue, WEI Hao, ZUO Xin, YANG Wankou (294)

Plant leaf detection technology based on multi-scale CNN feature fusion

.....LI Ying, CHEN Long, HUANG Zhaohong, SUN Yang, CAI Guorong (304)

Insulator detection based on SE-YOLOv5s.....TIAN Qing, HU Rong, LI Zuoyong, CAI Yuanzheng, YU Zhaochai (312)

Light weight rotating object detector based on angle sensitive spatial attention mechanism

.....YIN Kaishi, YANG Meng, GU Xi, WANG Zhicheng (322)

Collaborative representation based classifier with maximum correntropy criterion and locality constraint.....

.....YU Qinru, LU Guifu (334)

Robust non-negative supervised low-rank discriminant embedding algorithm

.....YAO Yu, WAN Minghua (342)

End-to-end speech enhancement based on ultra-lightweight channel attention.....

.....HONG Yi, SUN Chengli, LENG Yan (351)

Sparse representation for image recognition based on semi-genetic algorithm in feature space

.....SHI Linrui, HUANG Yijing, FU Jinwu, GUO Xinyue, FAN Zizhu (359)

Research on anti-spoofing method of face recognition based on semi-supervised learning

.....LI Li, ZENG Weiliang, HUANG Yonghui, SUN Weijun (370)

收录声明

本刊对发表的文章, 拥有出版电子版、网络版版权, 并拥有和其他网站交换信息的权利。本刊支付的稿酬中已经包含上述费用。

Chinese Journal of Intelligent Science and Technology has the copyright to publish electronic edition, online edition of the published articles, and has the right to exchange information with other sites. The expenses have been included in the fee paid by editorial department.

道德声明

本刊发表的论文是作者独立取得的原创性研究成果, 无一稿多投; 论文内容不涉及国家机密; 未曾以任何形式用任何文种在国内外公开发表过; 论文内容不侵犯他人著作权和其他权利。若发生一稿多投、侵权、泄密等问题, 论文作者将承担全部责任。

The authors of *Chinese Journal of Intelligent Science and Technology* guarantee that their submitted articles are original and contain nothing confidential. The said article is only submitted to *Chinese Journal of Intelligent Science and Technology*. The said article has not been published before and has not been submitted elsewhere for print or electronic publication consideration. The said article is no way whatever a violation or an infringement of any existing copyright or license from the third party. Otherwise, the authors of the said article shall take the blame for the violation or infringement of the related copyright and the leakage of secrets.