

AI 生成人脸图像鉴别赛题说明及参赛细则

一、赛题背景

目前，人工智能（AI）生成图像的发展现状呈现出蓬勃发展的态势，在艺术创作、医学影像分析、游戏开发等多个应用领域拥有广泛应用。随着深度学习技术的突破，特别是生成对抗网络（GAN）等模型的发展，AI 生成图像的能力得到了显著提升，能够生成高度逼真的图像，仅凭肉眼很难辨别真假。越来越多的研究机构和企业投入到 AI 生成领域，推动技术的不断创新和完善。同时，人工智能图像生成工具的广泛关注和使用，如 Diffusion Studio、Stable Diffusion 和 Midjourney 等，也进一步促进了其在各个领域的应用和普及。

然而，AI 生成图像的广泛应用也带来了许多潜在风险。例如，恶意网络用户可能利用 AI 生成虚假图片进行欺诈、造谣或误导公众，对社会秩序和个人利益造成威胁；在艺术创作、新闻报道等领域，AI 生成图像的滥用可能侵犯原创者的权益，破坏行业的公平竞争。在这种背景下，针对 AI 生成图像开发更有效的检测算法和技术以应对 AI 生成图像带来的风险和挑战，无疑是一个至关重要的研究方向。

二、赛题任务

本次比赛主办方提供由真实图像和 AI 生成图像混合的若干张图片，参赛者需准确判断每张图片是否是 AI 生成图像。AI 生

成的人脸图像鉴别是一个复杂而重要的任务，它涉及到对图像的真实性、自然性和来源的准确判断。

1. 任务描述

AI 生成人脸图像主要依赖于深度学习技术，特别是生成对抗网络（GANs）和变分自编码器（VAEs）等模型。这些模型能够学习真实人脸图像的分布，并生成具有相似特征的新图像。本任务通过深度学习和图像处理等技术来识别真实人脸图像和生成图像之间的差异，可以实现高精度的鉴别，从而通过视觉检查图像的细节和质感可以判断其是否为 AI 合成图像。

2. 任务说明

该任务给定一张人脸图像，需要根据图像的颜色、纹理、光影等特征判断该图像是否为 AI 合成图像。

输入：人脸图像。

输出：是否为 AI 合成：1 表示真实图像，0 表示 AI 合成图像。

三、数据集

举办方不提供训练数据集，报名成功的参赛队伍自行训练并优化算法。

主办方提供测试数据集，包含 5000~6000 张人脸图像。该数据集中包含由 Stable Diffusion 为主的不同 AI 工具网络生成的 AI 生成人脸图像，以及主办方自行收集的或从互联网中爬取得到的原始人脸图像。每张图像可能包含单人人脸或多人人脸，可

能是自然人脸也可能是非自然人脸（如卡通、素描），但有任意一部分（人脸部分或非人脸部分）为 AI 生成即应将该图像视为 AI 生成图像。数据集中 AI 生成人脸图像和原始人脸图像各约占一半，所有图像文件均为 jpg 格式，由主办方以 16 位随机字符（字母+数字）命名，像素大小 $\geq 128 \times 128$ ，但清晰度各不相同。

测试数据集将于 6 月中下旬公布，报名成功的参赛队可从大赛官网自行下载测试数据集。

注：为了控制文件尺寸或美化，图像文件可能采用 Photoshop 等图像处理工具软件进行过人工剪裁、编辑或美颜，此类处理不应视为 AI 生成。

四、解题思路

AI 生成图片识别算法的现有工作和解决思路很多，这里给出几种方案供参考：

1.Tan 等人[1]设计了一个利用梯度来表示 GAN 生成图像中人工痕迹的新型检测框架 LGrad，旨在构建一个可以跨模型和跨数据的通用检测器。具体来说，该检测器利用预训练的 CNN 模型将图像转换成梯度，随后利用这些梯度来呈现图片的人工痕迹，并将其输入分类器以确定图像的真实性。

2.Wu 等人[2]针对在线社交网络（OSN）场景，提出了一种鲁棒的图像伪造检测方案。该方案首先对 OSN 平台引入的噪声进行建模，然后将模拟的噪声集成到一个鲁棒训练框架当中。该方案将 OSN 噪声分解为可预测噪声和不可见噪声。可预测

噪声主要是模拟一些已知操作带来的可以预测的性能损失，而不可见噪声主要针对 OSN 平台的未知操作，Wu 等人利用对抗噪声的思想来对其进行建模，提升检测器对于对抗样本的鲁棒性。

3.Ojha 等人[3]发现针对扩散/自回归模型等新方法生成的图片，在未经真假图像分类训练的信息特征空间中执行近邻法或线性探测，可以大大提高检测模型的泛化能力。

五、评价方式

1.数据指标评价方式：参赛者给出的图片识别结果将与 Ground Truth 比对，赛题成绩由图片鉴别准确率决定（鉴别结果无法识别或输出不全的，视为错误）。准确度相同的，以运行时间少的为优，运行时间也相同的，以提交时间早的为优。源代码无法运行或无法正确输出结果文件的，均视为无效提交。各参赛队伍采用以成绩最好的一次提交。

评审运行环境主要参数：

Intel Xeon 10C CPU * 1

32G 内存 * 1

nVidia 3090 显卡 * 1

Ubuntu 20.04,

2.综合评价方式参考大赛组委会评审总则。

六、成绩提交

1.举办方不提供训练数据集，报名成功的参赛队伍自行训练并优化算法。

2.主办方提供测试数据集，参赛队伍需要给出对于主办方提供的 5000~6000 张人脸图像进行鉴别，并在线提交结果。排行榜开启后，每个队伍每周的提交次数限制为 1 次；每周提交截止时间为周日晚 24 点。

3.具体来说，参赛者需提交一个以“参赛队伍名称+队长姓名+提交日期”命名的文件压缩包，其中包含：

(1) ./src 目录：该目录下为所有算法模型源代码文件，使用 python 3.x 执行其中 main.py 文件，可读取./testset 目录下的测试数据集文件，输出./result_xxxxxxxx.xls 文件；

(2) ./doc 目录：说明文档，以便评审老师理解算法代码；

(3) ./result_xxxxxxxx.xls:图片鉴别结果输出文档（csv 格式，其中 xxxxxxxx 为 MMDDHHMM,月日时分）。该文档中应包含两列，第一列为图片的文件名称（按字母序升序排序），第二列为该图片的鉴别结果，用数字 0 或 1 表示。其中，0 表示该图片是原始图像，1 表示该图片为 AI 生成图像。输出结果从第一行开始，无需表头。

文件提交方式另行通知。

七、参考文献

[1] Tan C, Zhao Y, Wei S, et al. Learning on gradients: Generalized artifacts representation for gan-generated images detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12105-12114.

[2] Wu H, Zhou J, Tian J, et al. Robust image forgery detection against transmission over online social networks[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 443-456.

[3] Ojha U, Li Y, Lee Y J. Towards universal fake image detectors that generalize across generative models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 24480-24489.

八、省赛说明

- 1.获奖比例与大赛组委会公布的获奖比例一致；
- 2.省赛榜单截止时间：10月31日前，具体事宜另行通知。

九、总决赛说明

- 1.获奖比例与大赛组委会公布的获奖比例一致；
- 2.省赛结束后另行通知。

十、联系方式

（一）赛题负责人

联系人：周竞扬

赛题 QQ 群：833020799

（二）国赛组委会

国赛组委会邮箱：lican@digix.org.cn

国赛参赛学生交流 QQ 群：635906376、695491030

大赛官网：www.digix.org.cn