AI 生成人脸图像鉴别算法说明文档

1 概述

本文提出了一种基于双路径特征提取的 AI 生成人脸图像识别算法。该算法采用并行的 ResNet50 网络结构，分别处理 MTCNN 人脸检测后的图像和原始图像。在特征提取阶段，算法不仅利用 ResNet50 提取特征，还融合了边缘检测、色彩分布等额外特征。通过将两个 ResNet50 网络预测结果的置信度以及多维特征进行融合，最终使用决策树分类器实现对 AI 生成人脸的精确识别。整体算法可视化展示在图1中。

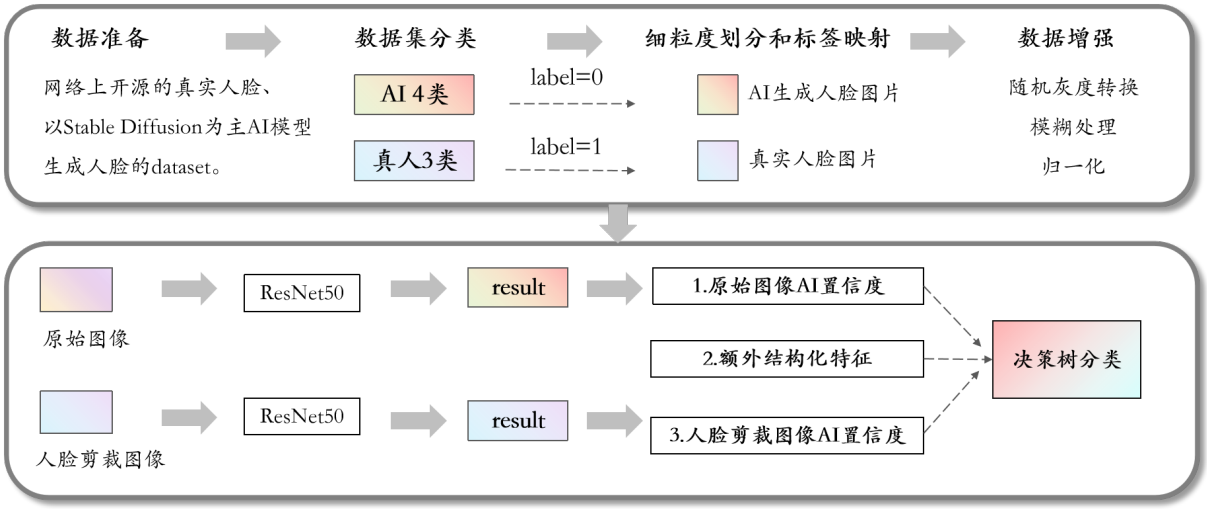

图 1　算法流程

2 数据准备

2.1 训练集准备

为了模型能够更准确、高效地区分以 Stable Diffusion 为主的 AI 生成工具产生的图像和真实图像，同时正确预测省赛国赛可能处于不同数据分布的测试集的 SD 生成图片，我们搜集了网络上开源的真实人脸和不同 AI 模型生成人脸的 dataset。

考虑到从互联网上收集而来的数据集存在标签噪声以及长尾分布等问题，有效的 AI 生成人脸图片的 dataset 有限，我们使用了 Stable Diffusion webui 工具，导入各种 LoRa 模型来生成各种人脸图片。

为了提高模型在不同类型 AI 生成人脸和真实人脸的预测准确性，我们准备了 7 类训练集数据（ Stable Diffusion 生成的人脸 / StyleGAN 生成的人脸 / AI 素描人脸 / AI 卡通人脸 / 真实人脸 / 卡通人脸 / 素描人脸 ），图2展示了每个训练集类别的典型。初步观察可以看出图像大小，分辨率、五官比例以及光感等存在差异，为了让模型更好的学习到不同人脸之间的这些差异，接下来我们进行细粒度划分和标签映射。

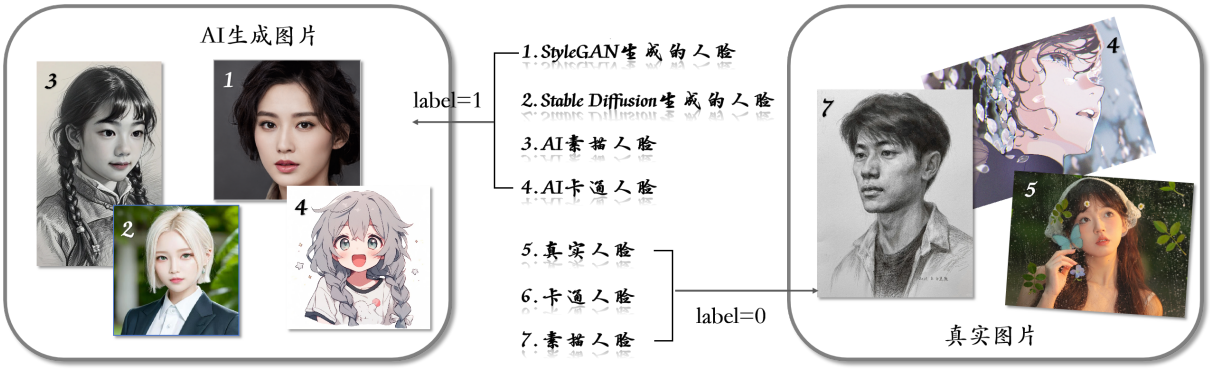

图 2　7 类训练集数据

2.2 细粒度划分和标签映射

细粒度分类任务可以很好地识别出较大的类内差异和细微的内间差异，因此我们采用了细粒度划分策略，将上述的 7 类 dataset 作为 7 个 label 来预测。

进行 7 分类之后，再对 label 映射到 0-1 的 label，代表是否是 AI 生成人脸，具体映射规则如下：

- AI 生成图像（label=1）：Stable Diffusion 生成人脸、StyleGAN 生成人脸、AI 素描人脸、AI 卡通人脸
- 真实图像（label=0）：真实人脸、卡通人脸、素描人脸

这种细粒度分类策略不仅提升了模型对细微特征的识别能力，保持了训练过程中的特征区分度，同时也满足了最终二分类的评估需求，有效提高了模型在真实应用场景中的泛化性能。

2.3 数据增强

考虑到 C1 数据集中存在的照片特征，我们引入了随机灰度转换（概率 0.05）和模糊处理（概率 0.4，范围 0.25-0.45）来增强模型的鲁棒性。最后使用 ImageNet 的标准化参数（均值 [0.485, 0.456, 0.406] 和标准差 [0.229, 0.224, 0.225]）进行归一化。数据准备整体过程展示在图3中。

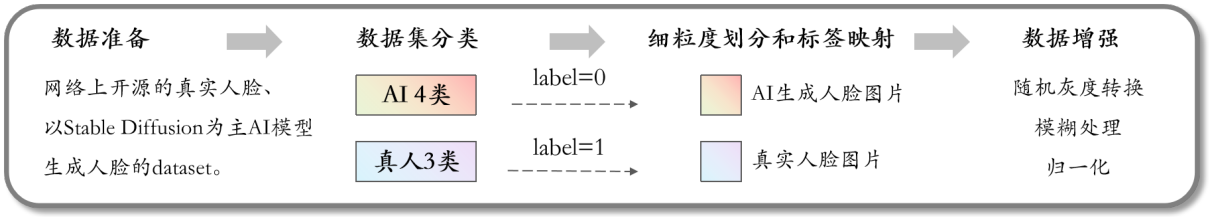

图 3　数据准备流程

3 改进的神经网络架构和消融实验

3.1 特征融合

在本实验中，我们选取了 ResNet50 和 Swin Transformer 两个强大的神经网络作为基座网络进行对比实验。由于 ResNet50 和 Swin Tranformer 的特征融合方式一致，此处只介绍 ResNet50 的特征融合方式。

首先，针对 ResNet50 的输入层进行了改进，将标准的 3 通道 RGB 输入扩展为 5 通道结构。除了保留原始 RGB 图像 (3 通道) 外，我们额外引入了 Canny 边缘检测图 (1 通道) 和 Laplacian 锐度图 (1 通道)。显著增强了模型对图像边缘、纹理和局部结构特征的感知能力。在特征提取环节，我们设计了多维度的特征融合机制，将神经网络提取特征后的特征向量与以下三类标量特征进行 concat 操作。

- Laplacian 平滑度系数。
- RGB 颜色直方图。
- 灰度直方图

3.2 消融实验

本算法的输入图像大小设置为 512x512，但是考虑到不同的图像分辨率统一化会导致模型预测精度有略微的差异，本文使用暴力 resize 和 padding 后 resize 两种方法进行对比实验。结果如下表所示。

使用的图像	预处理	网络类型	准确率
原始图像	暴力 resize	Swin-T	0.80
		resnet50	0.79
	padding	Swin-T	0.87
		resnet50	0.90
人脸裁剪后图像	暴力 resize	Swin-T	0.87
		resnet50	0.88
	padding	Swin-T	0.88
		resnet50	0.86

通过实验发现，原始图像 +padding+ResNet50 组合取得最佳性能 (0.90)，其次是人脸裁剪后图像 +resize+ResNet50 组合 (0.88)。我们发现，通过神经网络架构去分类图片的准确率已经到了瓶颈，所以我们后续采用二阶段预测方法，加上机器学习模型进一步分类。

4 决策树分类

4.1 决策树的输入

在经过改进的 ResNet50 完成初步分类后，我们设计了基于决策树的集成判别机制。决策树的 Input 来自三个维度:

- 两个核心输入特征: 经过 MTCNN 人脸检测裁剪后的人脸图像通过 ResNet50 预测得到的 AI 人脸置信度、原始完整图像经 ResNet50 处理预测得到的 AI 人脸置信度。

- 神经网络中 concat 的相同的额外特征:Laplacian 平滑度系数、RGB 颜色直方图、灰度直方图。

- 颜色分布特征:R,G,B 每个通道的均值和标准差、H,S,L 每个通道的均值、标准差、偏度、峰度。

其中，AI 人脸置信度的计算策略如下：

- 预测结果为真实类别任何一类（真实人脸、卡通人脸、素描人脸）：AI 人脸置信度 = 1 - 置信度 score
- 预测结果为 AI 类别任何一类（Stable Diffusion 生成人脸、StyleGAN 生成人脸、AI 素描人脸、AI 卡通人脸）：AI 人脸置信度 = 置信度 score

我们的双路径特征提取与 AI 人脸图像分类的整体架构流程如图4所示。

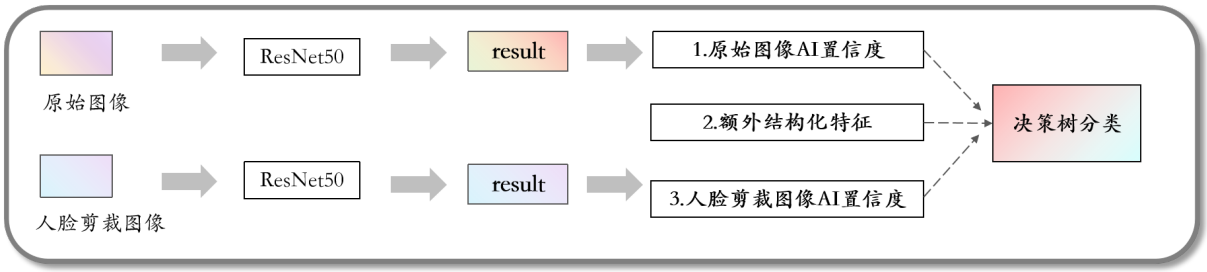

图 4　算法流程图

4.2 决策树分类结果

通过集成这些特征，决策树在测试集上取得了优异的性能表现：

特征重要性分析表明，ResNet50 模型提取的置信度特征的特征重要性大于 0.8，在最终分类中起到了决定性作用，同时结构化统计特征也提供了有效的补充信息，共同构建了一个鲁棒的 AI 图像检测系统。

表 1　决策树分类性能

评估指标	性能值
准确率 (Accuracy)	0.991
精确率 (Precision)	0.988
召回率 (Recall)	0.993
F1 分数 (F1-Score)	0.990

5 总结与创新

本文提出了一种基于 ResNet50 和决策树相结合的 AI 生成人脸图像识别算法。先使用细粒度划分的 7 分类训练 ResNet50，然后通过标签映射实现最终的二分类目标。采用双路特征提取并行的 ResNet50 网络分别处理 MTCNN 人脸检测裁剪后的图片和原图的输入，通过多维特征提取和融合策略，使用决策树分类实现对 AI 生成人脸图像的识别，最终取得了 99.1% 的准确率，展现出优异的分类性能和泛化能力。

创新点:

1. **精细化数据分类策略**：提出 7 类细粒度分类体系，包含 Stable Diffusion、StyleGAN 等 AI 生成图像和真实人脸、卡通、素描等真实图像，通过标签映射机制提升模型对细微特征的识别能力和泛化性能。

2. **双路径并行特征提取**：创新性地设计双路径 ResNet50 架构，一路处理原始完整图像获取全局特征，另一路专注于 MTCNN 检测的人脸区域提取局部特征，实现多尺度特征的协同分析。

3. **多维特征融合机制**：设计了创新的特征融合策略，将深度学习自动提取的特征与手工设计的结构化特征 (包括 Canny 边缘检测图、Laplacian 锐度图、Laplacian 平滑度系数、RGB 颜色直方图、灰度直方图、18 维 RGB 统计特征、12 维 HSL 空间特征) 进行有效整合，增强了模型的特征表达能力。

4. **二阶段级联分类框架**：构建了深度学习与决策树相结合的二阶段分类机制，利用改进的 ResNet50 进行初步特征提取和分类，再通过决策树进行精细化判别，充分发挥了深度学习和传统机器学习的优势。

参考文献

[1] Z. Liu, Y. Lin, Y. Cao, et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012-10022.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[3] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, “Attention augmented convolutional networks,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3286-3295.

[4] J. Li, H. Xiao, and Y. Li, “HPViT: A hybrid visual model with feature pyramid transformer structure,” in 2023 8th International Conference on Control, Robotics and Cybernetics (CRC), 2024, pp. 212-216.

[5] J. Gao, M. Zhang, L. Ma, M. Huang, and Z. Li, “A unified binary classification network for weld image detection,” in 2023 8th International Conference on Control, Robotics and Cybernetics (CRC), 2024, pp. 275-279.