北京大学

硕士研究生学位论文

题目：基于扩散模型的人脸眼部图像修复

姓　　名：冯浩然

学　　号：2001213100

院　　系：计算机学院

专　　业：计算机科学与技术（智能科学与技术）

研究方向：人工智能与产业创新

导　　师：王亦洲 教授

二〇二三 年 五 月

摘要

人脸眼部图像修复是指对人脸照片中因闭眼、皱纹等存在的眼部缺陷进行美化。相比人工手动调整图像等传统方式，基于深度学习的修复方法处理效率高，操作门槛低，能解放普通用户用照片记录生活中的美好，具有较大的应用价值。该任务中，一个好的修复结果应满足三个标准：1）真实性：细节清晰、光照自然、形状合理、边缘平滑，看起来是真实的人脸照片；2）身份特征：修复结果忠实保留眼睛的瞳色、形状等特征；3）表情神态：修复结果的眼睛部位与脸部整体神态连贯，情绪一致。

眼睛区域形状多样，纹理和细节复杂，且对照片修复的整体观感影响巨大，给任务带来了巨大困难。目前该领域的方法主要基于生成对抗网络和扩散模型，现有方法主要面临三个问题：1）修复结果中眼部身份特征与原待修复图片不匹配；2）以 ExGAN 为代表，基于生成对抗网络的方法修复结果的精确度和真实性较好，但训练稳定性和模型泛化性不足；以 RePaint 为代表，基于扩散模型的方法虽然训练稳定性更好，但是在约束较多的条件生成任务中却难以生成足够精确和真实的修复图像；3）待修饰图像面部姿态倾斜或旋转时，容易产生模糊和杂乱结果。针对上述问题，本文创新性地提出一个基于去噪扩散模型的人脸眼部图像修复框架，主要的创新点如下：

1. 针对修复结果无法保留身份特征的问题，本文提出在框架中引入额外信息引导生成。在输入中给出同一个人的睁眼照片和相应眼部蒙板，利用眼部特征引导模型修复时恢复真实的眼部特征，并在训练时利用身份特征保留损失约束模型有效利用额外的引导信息。为减少引导图片中与眼部无关的特征影响，本文利用泊松融合方法对引导图片进行预处理。实验结果证明，引入额外信息引导的框架能在修复结果中较好保留瞳色、形状等身份特征。

2. 针对去噪扩散模型修复结果细节精确度和真实性不足的问题，本文提出了一种结合空间多尺度注意力模块的 U-Net 网络结构。空间多尺度注意力模块在传统多尺度注意力的基础上，考虑人脸修复问题中某些已知区域的信息更重要的特性，对得到的注意力图进行逐像素权重处理，帮助模型排除掩码区域的干扰，并更充分的利用重要区域的信息，确保生成细节真实、语义合理的结果。消融实验证明，该方法可有效提高生成图像的细节精确性和结构合理性。

3. 针对待修饰图像面部姿态倾斜或旋转时，容易产生模糊和杂乱结果的问题，本文利用自监督方法解决人脸图像旋转时纹理缺失问题，对输入修复模型的引导图片进行姿态调整，使其与待修补照片的姿态进行对齐，增强引导效果。实验结果证明，该方法能减少修补结果中较大区域的模糊和杂乱结构。

综上所述，本文对人脸眼部图像修复任务的精确度和真实性问题、身份特征保留问题和姿态旋转倾斜问题三个难点问题进行分析，并提出了具有创新性的解决框架，能生成良好的修复结果。在 celebA 数据集上的实验结果证明，本文提出框架在主要评测指标上超越现有方法，并可以生成真实、细节精确、结构合理且保留身份特征的修复结果。

关键词：图像修补，扩散模型，人脸修复

Facial eye region image restoration based on diffusion model

Feng Haoran (Computer Software and Theory)
Directed by Prof. Wang Yizhou

ABSTRACT

Facial eye region image restoration refers to the beautification of eye defects in facial photos such as closed eyes and wrinkles. Compared with traditional methods like manual image adjustment, deep learning-based restoration methods offer higher processing efficiency, lower operational barriers, and enable ordinary users to capture the beauty of life in photos, thus providing significant practical value. In this task, a good restoration result should meet three criteria: 1) authenticity: clear details, natural lighting, reasonable shape, and smooth edges, resulting in a realistic facial photo; 2) identity features: restoration results should faithfully retain eye features such as pupil color and shape; 3) expression consistency: the restored eye region should be coherent with the overall facial expression and convey the same emotion.

The diverse shapes, complex textures, and intricate details of the eye region, as well as its significant impact on the overall visual appeal of the restored photo, present considerable challenges to this task. Current methods in this field are primarily based on generative adversarial networks (GANs) and diffusion models, facing three main issues: 1) restored eye features in the results do not match the original images; 2) GAN-based restoration results exhibit good accuracy and authenticity, but lack training stability and model generalization; diffusion models offer good training stability but insufficient accuracy and authenticity; 3) when the face in the image to be restored is tilted or rotated, the results are likely to be blurry and disorganized. To address these issues, this paper innovatively proposes a facial eye region image restoration framework based on denoising diffusion models, with the following key innovations:

1. To address the inability of restoration results to retain identity features, this paper introduces additional information to guide the generation process within the framework. Open-eye photos of the same person and corresponding eye masks are added to the input. During training, the eye features are extracted using an eye feature recognition model and fed into the denoising process of the U-Net network. Identity feature preservation loss constrains the model to effectively utilize the additional guiding information. To reduce the impact of unrelated features in the guiding image, this paper employs the Poisson blending method for preprocessing the

guiding image.

2. To address the insufficient accuracy and authenticity of restoration results using denoising diffusion models, this paper proposes a U-Net network structure with the spatial multi-scale attention modules. The spatial multi-scale attention module, building upon traditional multi-scale attention, considers the importance of certain known regions in facial restoration problems, performing pixel-wise weighted processing on the obtained attention maps. This helps the model eliminate interference from masked regions and fully utilize important region information, ensuring the generation of realistic and semantically reasonable results.

3. To tackle the issue of blurry and disorganized results when the face in the image to be restored is tilted or rotated, this paper uses a self-supervised method to address texture loss issues during facial image rotation. The guiding image input to the restoration model is adjusted for pose, aligning it with the pose of the image to be restored, thus reducing large blurry and disorganized structures in the restoration results.

During training, the model utilizes identity-matched paired data obtained from the CelebA dataset, while an additional user photo is employed for testing. Considering practical applications, paired data will not cause inconvenience for users and can improve restoration effects.

In summary, this paper analyzes the three main challenges in facial eye region image restoration: accuracy and authenticity, identity feature preservation, and pose rotation and tilt. It proposes an innovative solution framework addressing these challenges and generating high-quality restoration results. The experimental results on the CelebA dataset demonstrate that the framework proposed in this paper outperforms existing methods in terms of major evaluation metrics and is capable of generating restoration results that are realistic, accurate in details, structurally reasonable, and preserve identity features.

KEY WORDS: Image inpainting,Diffusion model,Face restoration