Journal of Image and Signal Processing 图像与信号处理, 2025, 14(2), 299-309
Published Online April 2025 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2025.142027

Hans汉斯

基于CSWin-Transformer和WGAN技术的人脸遮挡修复研究

黄施豪，金　钊

云南大学信息学院，云南　昆明

收稿日期：2025年3月28日；录用日期：2025年4月19日；发布日期：2025年4月29日

摘　要

针对于当前人脸遮挡修复方法中出现修复图像信息不完整、纹理模糊、产生伪影、细节欠佳以及模型训练不稳定等问题，提出一种基于CSWin-Transformer和WGAN的人脸遮挡修复方法。该方法以Encoder-Decoder结构作为生成器，在生成器中引入CSWin-Transformer Block来精细识别和处理被遮挡的面部区域，以提高处理的针对性和效率，解码器通过跳跃连接与编码器多尺度特征融合，更好学习图像的细节特征，优化最终效果。在判别器中引入Wasserstein距离，来提高模型训练稳定性以及生成图像的真实性，同时在判别器中引入CSWinSelf-Attention，增强判别器对图像全局结构和细节信息的理解。实验结果显示，文章方法在所使用的CelebA的数据集上有良好的修复效果，在峰值信噪比(PSNR)和结构相似性指数(SSIM)指标上与目前一些图像修复方法相比表现更优。

关键词

图像识别，人脸修复，生成对抗网络，WGAN，CSWin-Transformer

Study on Face Occlusion Repair Based on CSWin-Transformer and WGAN Techniques

Shihao Huang, Zhao Jin

School of Information, Yunnan University, Kunming Yunnan

Received: Mar. 28th, 2025; accepted: Apr. 19th, 2025; published: Apr. 29th, 2025

Abstract

In view of the problems of incomplete repair image information, blurred texture, artifacts, poor details and unstable model training, a face occlusion repair method based on CSWin-Transformer and WGAN is proposed. This method takes Encoder-Decoder, structure as the generator, and introduces

文章引用: 黄施豪, 金钊. 基于CSWin-Transformer和WGAN技术的人脸遮挡修复研究[J]. 图像与信号处理, 2025, 14(2): 299-309. DOI: 10.12677/jisp.2025.142027

CSWin-Transformer Block in the generator to finely identify and process the occluded face areas, so as to improve the pertinacity and efficiency of processing. The decoder integrates with the encoder multi-scale features through jump connection to better learn the detailed features of the image and optimize the final effect. The Wasserstein distance is introduced into the discriminator to improve the stability of the model training and the authenticity of the generated image. Meanwhile, CSWin Self-Attention is introduced in the discriminator to enhance the understanding of the global structure and details of the image. The experimental results show that the method has good repair effect on the data set of CelebA used, and better than some current image repair methods in peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) indicators.

Keywords

Image Recognition, Face Repair, Generative Adversarial Network, WGAN, CSWin-Transformer

1. 引言

随着人脸识别应用的广泛推广，人脸去遮挡恢复技术的发展变得日益重要。该技术在数字身份验证、监控安全、刷脸支付等领域发挥着不可或缺的作用。但是，在生活场景中，人脸常常受因眼镜、口罩、围巾、帽子等物体遮挡而无法被系统识别，甚至犯罪分子也可通过遮挡面部来躲避天眼系统[1]，所以人脸去遮挡修复有着重要的实际运用价值。

传统的图像修复方法有基于纹理合成的方法，如 He 等[2]通过从图像的其他区域复制和粘贴纹理来填补缺失区域，但简单的纹理复制不能很好地恢复图像的全局结构，处理大面积缺失效果差，还有基于扩展的方法如 Criminisi [3]通过扩展图像的边界，逐步填充缺失区域，但当图像缺失区域复杂或边界不明显时，扩展方法可能会产生不自然的填充结果。

近些年，卷积神经网络(CNN) [4]还有生成对抗网络(GANs) [5]及其变种如 DCGAN(深度卷积生成对抗网络) [6]、cGAN (条件生成对抗网络) [7]以及等在图像生成领域展现了巨大的潜力。传统 GAN 通过训练生成器和判别器进行对抗学习，使生成的图像逐渐接近真实图像。DCGAN 它将 GAN 模型和 CNN 进行了融合，把 CNN 的理念应用到生成网络和判别网络中，以提高模型在训练时的性能，从而增强模型的修复能力。cGAN [7]则通过给予条件约束，使得生成的图像能够满足特定的要求或属性。Pathak 等[8]提出了基于 CNN 的上下文编码器方法，通过重构缺失区域来学习图像的高层次特征但生成的图像细节质量较差，容易产生模糊或不连续性，尤其在复杂纹理区域。Iizuka 等[9]提出结合全局和局部鉴别器，确保修复结果在全局结构和局部细节上都具有一致性的图像修复方法，但不足之处在于无法很好处理大面积遮挡修复。刘波宁等[10]提出基于双鉴别网络的生成对抗网络图像修复，但其网络结构复杂且依赖于预处理算法一定程度上限制了网络能力。Yu [11]等提出了上下文注意力机制，使网络能从未遮挡区域中有效提取相关特征并填充缺失部分，但在处理非常细节化的部分(如人脸特征)时，生成的图像可能会有伪影。

本文提出一种结合 CSWin-Transformer [12]与 Wasserstein GAN 的新策略来克服现有方法的局限。方法首先在判别器中引入 Wasserstein 距离从而使网络的训练更加稳定，然后通过引入 CSWin 自注意力机制有效识别和处理被遮挡的面部区域，提高关键信息的观察精度，提高细节部分的修复效果的自然性。

通过将自注意力机制与 WGAN 相结合，本文所提出的方法能够在复杂场景下，在保证图像真实性的同时，让修复图像更自然、更连续的图像修复结果。

2. 相关技术原理分析

2.1. Wasserstein GAN

Wasserstein GAN (WGAN)由 Arjovsky 提出[13] [14]，旨在解决传统生成对抗网络(GAN)在训练过程中遇到的几个关键问题，包括训练不稳定、模式塌陷以及难以收敛，其工作原理如图 1 所示。其工作流程：初始时，生成器和判别器的参数随机设定，随后生成器开始通过接收随机噪声向量并转换为假数据的方式学习仿造真实数据分布[4]。同时，判别器努力区分接收到的真实数据样本和生成器创造的假数据样本。两者的性能都通过损失函数衡量，该函数刻画了真实数据与生成数据间的差异。随着损失函数指导下的参数更新，生成器越发精于创建逼真数据，而判别器则越来越能准确区分真假数据，整个过程迭代进行，直到生成器与判别器之间的性能达到一种动态平衡为止。

WGAN 通过引入一种新的损失函数——Wasserstein 距离(或称 Earth-Mover 距离)，来衡量生成数据分布与实际数据分布之间的差异。

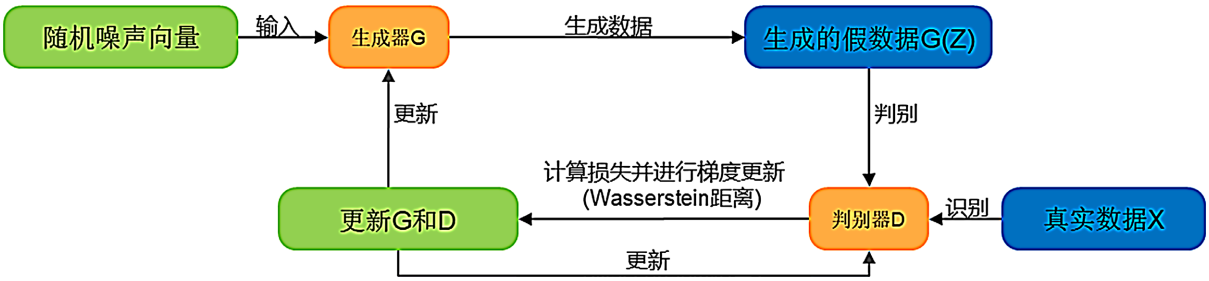


Figure 1.WGAN, working principle
图 1. WGAN 工作原理

相较于传统 GAN 使用的 Jensen-Shannon 散度，Wasserstein 距离提供了一个更平滑、可导的准则，使得梯度下降过程中能够更加稳定地调整生成器的参数。Wasserstein 距离的定义的表达式如下：

$$W(p,q)=\inf_{\gamma\sim\Pi(p,q)}E_{x,y\sim\gamma}\left[\|x-y\|\right]$$

其中，$\gamma\sim\Pi(p,q)$是 p，q 组合起来的所有可能的联合分布的集合，对于每一个 γ 可以从其中采取样本 $x,y\sim\gamma$ 的到一个样本 x 和 y，通过 $\|x-y\|$ 计算得到样本距离，然后就能通过 $E_{x,y\sim\gamma}\left[\|x-y\|\right]$ 计算样本对距离的期望值。

2.2. CSWin 自注意力机制

CSWin(Cross-Shaped Window Transformer)是一种新的视觉 Transformer 架构，旨在通过创新的基于窗口的方法处理图像数据，其网络结构如图 2 所示。其中红色点表示 query，绿色区域表示 key。图 2 中(b)~(e)分别表示 query 点与 global 区域、shiftedlocal 区域、local 区域以及横向区域的相关性计算。

Cross-Shaped Window Self-Attention 将 Multi-Head 平均分成两个部分：一部分用于横向条纹 Self-Attention，另一部分用于纵条纹 Self-Attention，最后将两部分的输出进行拼接。如图 2 中(a)所示，维度为 $H\times W\times H$ 的特征图会线性投影到 K 个头(每个头的特征图通道数为 C/K)，然后平均分成两部分：一部分从位置 h_1 到位置 $h_{K/2}$，另一部分从位置 $h_{k/2+1}$ 到 h_k 位置。对于横向条纹注意力如式(2)，条纹宽度为 sw，

每个头的特征图会根据 sw 划分成不重叠的横条纹，记作 $\left[X^1, X^2, \cdots, X^M\right]$，其中，$M = H/sw$。接着，计算所有横条纹和纵条纹的 self-attention，最后将所有输出拼接起来。输入特征 $X \in \mathrm{R}^{(H\times W)\times C}$，$W_k^Q \in \mathrm{R}^{C\times d_k}$，$W_k^K \in \mathrm{R}^{C\times d_k}$，$W_k^V \in \mathrm{R}^{C\times d_k}$ 分别表示 k 个多头注意力的 Q，K，V 投影矩阵[5]。

$$X = \left[X^1, X^2, \cdots, X^M\right], \quad \text{where} \quad X^i \in \mathbf{R}^{(sw\times W)\times C} \quad \text{and} \quad M = H/sw$$

$$Y_k^i = \text{Attention}\left(X^i W_k^Q, X^i W_k^K, X^i W_k^V\right), \quad \text{where} \quad i = 1, \cdots, M \tag{2}$$

$$\text{H-Attention}_k\left(X\right) = \left[Y_k^1, Y_k^2, \cdots, Y_k^M\right]$$

另外，考虑人脸图像不存在方向性偏差，将 K 个 head 平均分成两个平行组，第一组的 head 表现出横向条纹的自注意力，第二组的 head 表现出竖直条纹的自注意力。最后，这两个并行组的输出将被连接在一起，如式 3，其中 $W^0 \in \mathrm{R}^{C\times C}$。通过多头分组，CSWin-Transformer Block 实现了将局部特征扩大至全局范围，为人脸修复过程提供了更充分的特征信息。

$$\text{CSWin-Attention}\left(X\right) = \text{Concat}\left(head_1, \cdots, head_K\right)W^O$$

$$\text{where} \quad head_k = \begin{cases} \text{H-Attention}_k\left(X\right) & k = 1, \cdots, K/2 \\ \text{V-Attention}_k\left(X\right) & k = K/2+1, \cdots, K \end{cases} \tag{3}$$

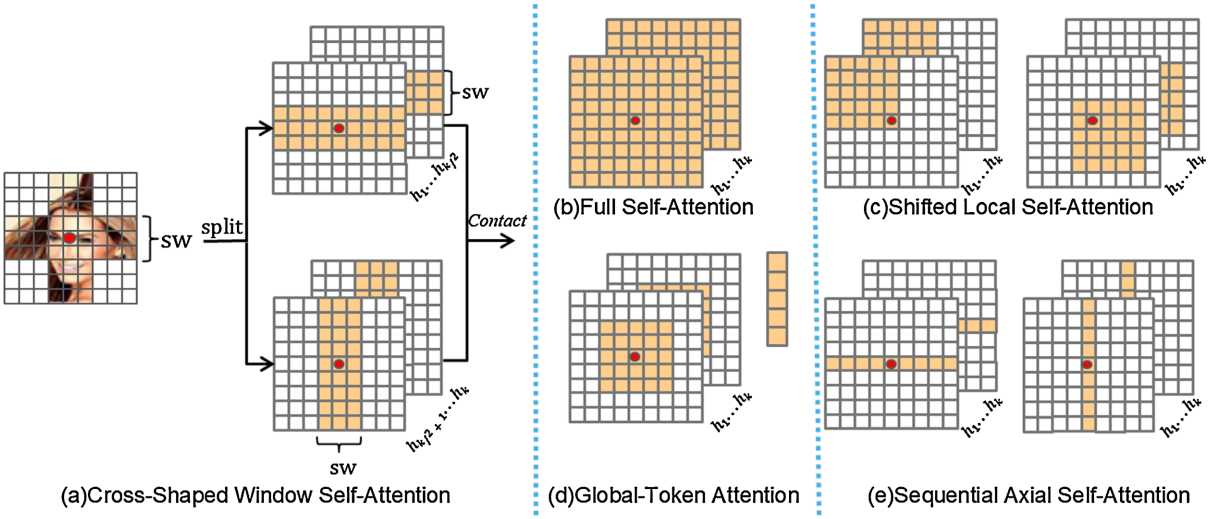


Figure 2. CSWin Self-Attention
图 2. CSWin Self-Attention

3. 人脸遮挡修复方法

3.1. 网络模型

3.1.1. 生成器

传统的卷积判别器通常是简单的卷积层堆叠，缺乏多尺度特征的融合能力。虽然也能有效地进行图像判别，在细节保留和多尺度特征提取方面有所欠缺。本文生成器模型结合了 CSWin-Transfomer 和 encoder-decoderp [15]结构，编码器部分由多个 CSWin-Transformer Block 组成，通过跨阶段窗口注意力机制提取图像的多尺度全局上下文特征，并通过 Merge-Block 实现逐层下采样，逐步减少特征图的空间分辨率并增大通道数。解码器部分则通过反卷积逐层上采样，恢复图像分辨率，并结合跳跃连接机制，将

编码器各层的细节特征与解码器对应层的特征融合，以保留多尺度信息。瓶颈层 Bottleneck 位于编码器与解码器之间，通过最后一层 CSWin-Transformer Block 进一步整合全局和局部特征，确保复杂损伤区域的高效修复。最终，网络通过线性变换将特征映射回图像空间，输出修复后的高质量图像，网络结构如图 3 所示。

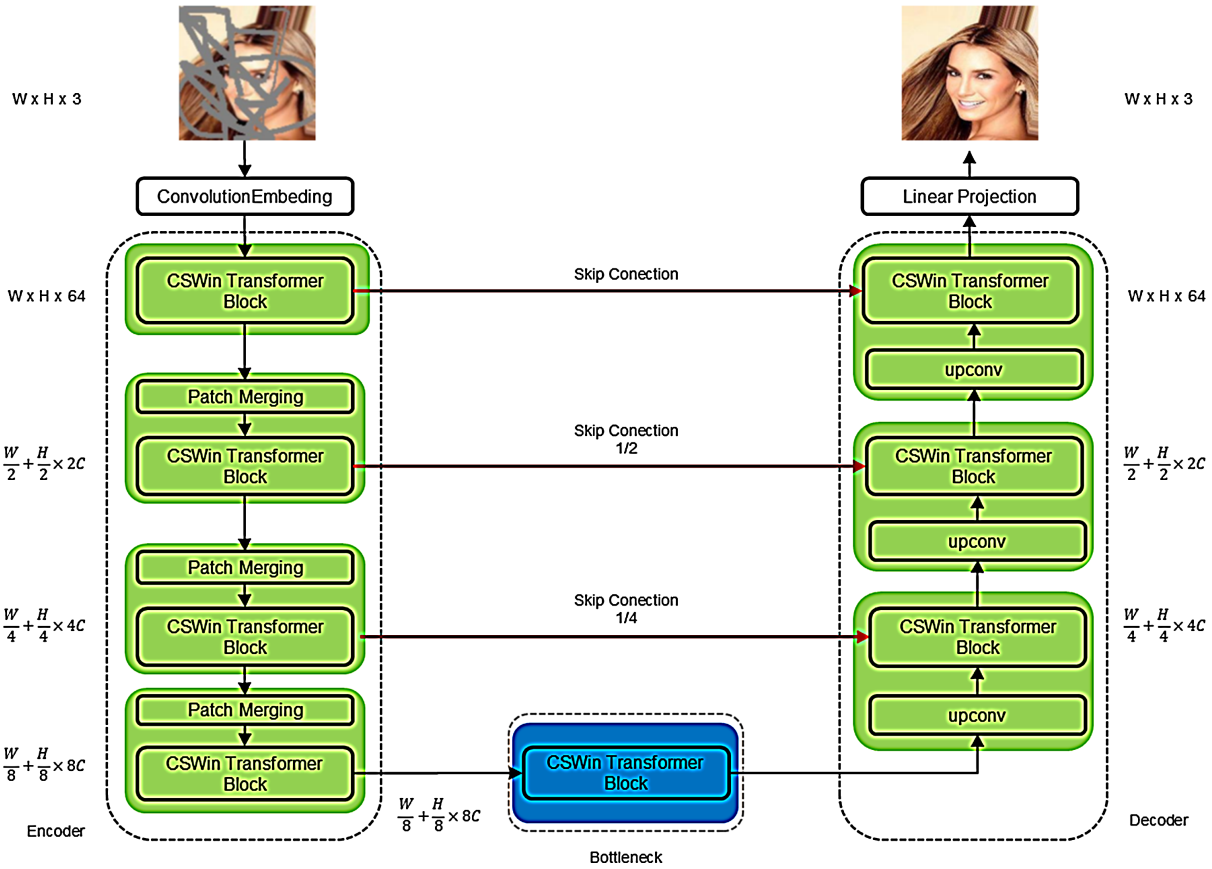


Figure 3. Generator structure
图 3. 生成器结构

3.1.2. 判别器

本判别器模型由多层卷积网络和注意力机制组成，首先通过四个卷积块对输入图像进行逐层下采样，每个卷积块由卷积层、批归一化层(第一层除外)和 LeakyReLU 激活函数组成。卷积层的作用是缩小图像尺寸，同时逐步增加通道数，以提取更高层次的特征，最终将输入图像转化为大小为 8×8 的特征图，通道数为 256。在四层卷积后引入了一个自注意力块 CSWin-Transformer Block，通过跨阶段窗口注意力机制进一步处理特征图。该注意力机制能够有效捕捉远距离的像素关联，增强判别器对图像全局结构和细节信息的理解[7]。输出层是一个卷积层，它将最后一层的特征图通道数减少到 1，输出一个表示图像真实性的得分图，其空间维度是输入图像高度和宽度的 1/16，网络结构如图 4 所示。

3.1.3. CSWin-Transformer Block

CSWin-Transformer Block 结构如图 5 所示，它利用跨形窗口自注意力机制，通过分别在水平方向和垂直方向执行自注意力操作，捕捉图像特征中的局部和全局依赖关系。其核心结构由两部分组成：第一部分通过 LayerNorm (LN)和 Cross-Shaped Window Self-Attention，将输入特征标准化并在划分的窗口内

进行自注意力操作，从而有效获取不同窗口区域的上下文信息；第二部分通过再次标准化的 LayerNorm 和多层感知机(MLP)，对特征进行线性和非线性变换，增强特征表达能力。每个部分都使用了残差连接，直接将输入信息跳过自注意力模块与 MLP 结合，保证了信息流的连续性，避免梯度消失问题。

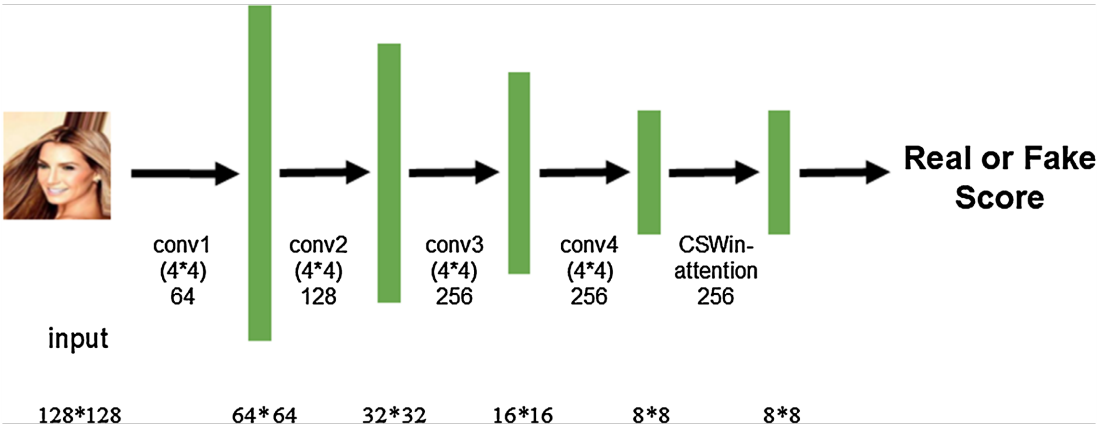


Figure 4. Discriminator structure
图 4. 判别器结构

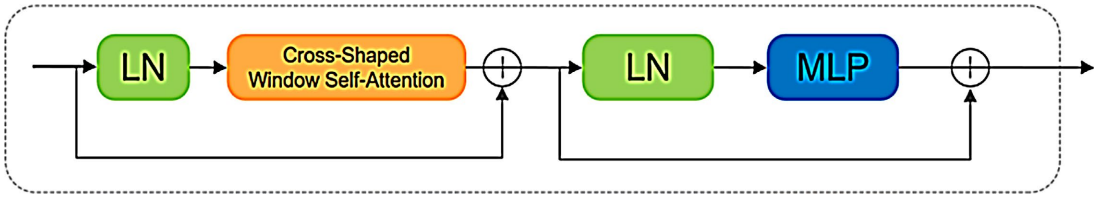


Figure 5. CSWin-Transformer Block structure
图 5. CSWin-Transformer Block 结构

CSWin-Transformer Block 形式上定义为式(4)，其中 X^l 是第 l 个 CSWin-Transformer Block 的输出。本文在生成器的编码阶段，CSWin-Transformer Block 应用于每个分辨率层级，处理逐步下采样的图像特征。在生成器的瓶颈层，CSWin Transformer Block 处理了高度下采样的特征图，通过在最深层次捕捉图像的全局结构信息，为后续的解码过程提供了丰富的语义信息。在解码阶段，每个解码层中的 CSWin Transformer Block 处理特征拼接后的输出[8]。

$$\hat{X}^l = \text{CSWin-Attention}\left(\text{LN}\left(X^{l-1}\right)\right) + X^{l-1}$$

$$X^l = \text{MLP}\left(\text{LN}\left(\hat{X}^l\right)\right) + \hat{X}^l \tag{4}$$

3.2. 损失函数

WGAN 所定义的生成器损失函数为：

$$\mathcal{L}_G = -\mathbb{E}_{s\to P_g}\left[D(\hat{x})\right] \tag{5}$$

式(5)中，D 代表判别器，P_g 是生成器产生的数据分布，$\hat{x}$ 表示是生成器从 P_g 中采样得到的生成样本。加入感知损失可以使生成的人脸图像在视觉上更加接近真实图像。感知损失通过比较深层特征来评估图像质量，这有助于生成器学习到更加精细和高级的图像细节，其损失函数如下：

$$\mathcal{L}_p(x,y)=\frac{1}{C_jH_jW_j}\sum_{c=1}^{C_j}\sum_{h=1}^{H_j}\sum_{w=1}^{W_j}\left(F_j(x)_{chw}-F_j(y)_{chw}\right)^2 \tag{6}$$

式(6)中，($F_j(x)$和($F_j(y)$分别是在第 j 层的提取特征，C_j，H_j 分别是第 j 层的通道数、高度和宽度。注意力图损失如下面公式(7)所示，其有助于保持生成图像的结构一致性。在人脸图像中，注意力机制可以专注于关键区域，如眼睛、鼻子和嘴巴，确保这些区域在去遮挡过程中保持正确的位置和形状[9]。

$$\mathcal{L}_A(x,y)=\frac{1}{N}\sum_{n=1}^{N}\left\|A(x)_n-A(y)_n\right\|_2^2 \tag{7}$$

其中 $A(x)$和 $A(y)$分别是生成图像和真实图像的注意力图，N 是注意力图的元素数量。所以生成器最终的损失函数就为：

$$\mathcal{L}_{\text{tatal}}=\mathcal{L}_G+\alpha\mathcal{L}_P+\beta\mathcal{L}_A \tag{8}$$

对于判别器，目标是最大化其对真实图像的评分和对生成图像评分的差距。因此，损失函数可以表示为真实图像的评分的平均值减去生成图像的评分的平均值，WGAN 使用的 Wasserstein 距离损失相较于传统 GAN 中的 JS 散度更稳定，解决了模式崩溃问题，其公式如下：

$$\mathcal{L}_D=\mathbb{E}_{x\sim P_r}\left[D(x)\right]-\mathbb{E}_{x\sim P_g}\left[D(\hat{x})\right] \tag{9}$$

式(9)中，P_r 代表真实数据分布，x 是在 P_r 中采样得到的真实样本。

4. 实验分析

4.1. 实验环境状况

实验在 ubuntu20.04 操作系统下采用 PyTorch2.0.0 框架 cuda11.8 进行模型训练，CPU 使用 Intel(R) Core i7-13700K CPU @3.40 GHz，16.0 GB RAM，GPU 为 NVIDIA Geforce RTX4080，显存 16 G。

4.2. 数据集和实验设置

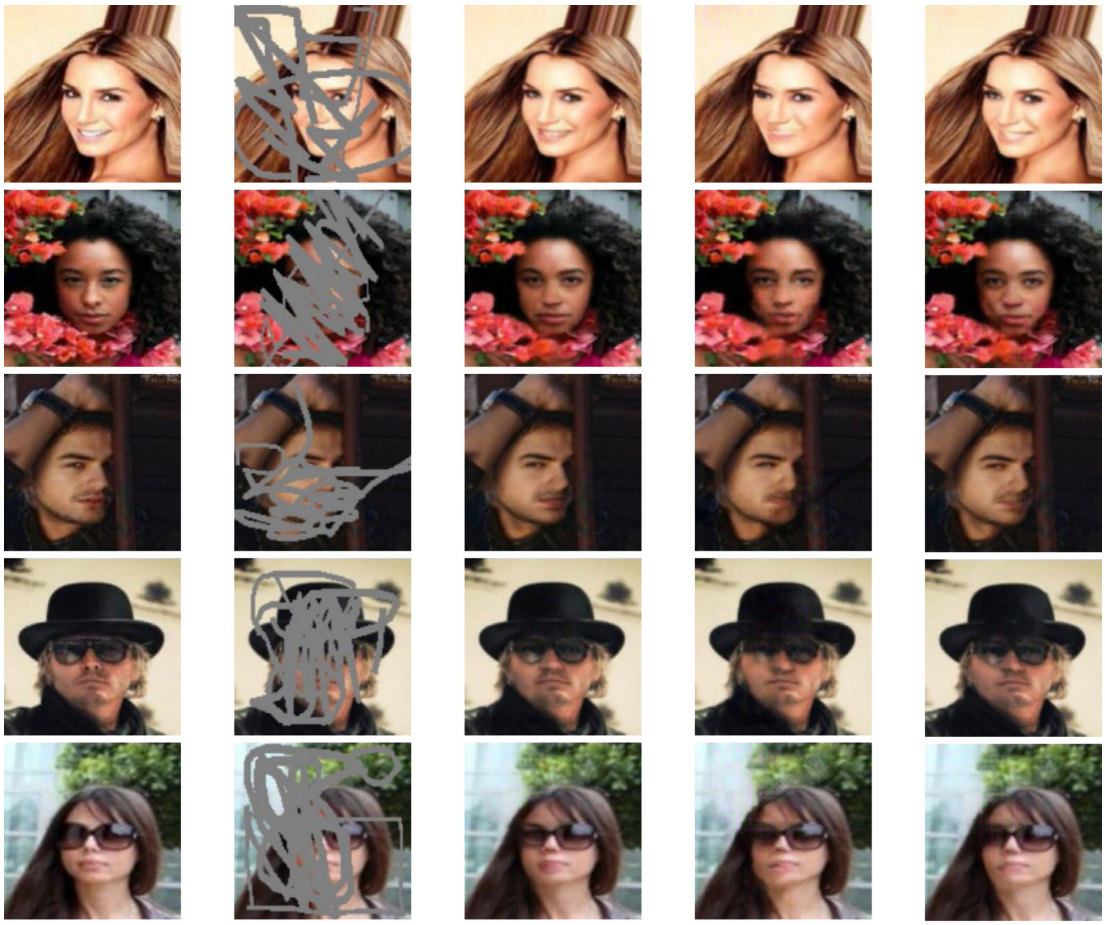
本实验采用 CelebA [16]中的部分数据 15,964 张图像作为人脸数据集，并使用了 NVIDIA Irregular Mask 数据集[17]包含 55,116 个掩码。实验将人脸图像和掩码调整为 128 × 128 像素。同时将它们转换为 PyTorch 张量。把人脸图像的像素值从[0, 255]范围归一化到[−1, 1]范围。将掩码转换为二值图像，其中 1 表示缺失部分，0 表示可见部分。batch-size 为 4，lr 为 0.0002，判别器使用权重剪裁(范围设定为[−0.01, 0.01])，生成器和判别器分别使用 AdamW 和 RMSprop 优化器。

4.3. 实验方法

4.3.1. 定性分析

本文方法修复效果如图 6 所示，从左往右依次为，原人脸图像、遮挡人脸、本文方法、卷积神经网络方法、生成对抗网络方法。本文方法修复后的图像中，整体视觉效果具有色彩的一致性和图像的真实感。例如对第一行人脸进行修复时，相较于卷积神经网络和生成对抗网络，本文方法修复后五官清楚真实，形状和位置与原图较为一致，没有明显的错位或变形。本文方法修复人脸后，皮肤颜色和纹理在修复后没有明显的色差或不自然的过渡，但卷积神经网络方法人物左脸部分出现了皮肤色彩偏差并产生一些伪影。通过对比第三行修复效果可以看出，在修复较大区域遮挡以及细节过度问题时，本文方法修复区域与周围图像边缘平滑，修复效果较为清晰，几乎不存在伪影等现象，但是对于某些非刚性遮挡(如第

四行图片中的眼镜)，修复后的图像出现轻微结构扭曲现象。这可能是由于模型对非刚性物体的几何结构理解不足，导致修复结果与原图有细微差别。



(a) 原图 (b) 遮挡人脸 (c) 本文方法 (d) 文献 [8] 方法 (e) 文献 [9] 方法

Figure 6. Comparison diagram of the repair effect
图 6. 修复效果对比图

4.3.2. 定量分析

针对修复后图像质量的定量分析，本文采用 PSNR (峰值信噪比) [18]和 SSIM (结构相似性指数) [19]来进行评估。PSNR 是一种基于图像敏感度的图像质量评价指标，用来评估图像的恢复质量，衡量图像中的噪声或失真程度。PSNR 越高，表示恢复图像和原始图像之间的差异越小，噪声越少。本文方法得到平均 PSNR 为 35.62 dB。SSIM 是一种基于结构信息的图像质量评价指标，考虑了图像的亮度、对比度和结构信息。与 PSNR 不同，SSIM 不仅关注像素级的差异，还衡量图像在视觉感知上的相似性。SSIM 值的范围在 0 到 1 之间。1 表示图像在亮度、对比度和结构上的完全一致，0 表示完全不一致。SSIM 值越接近 1，说明修复图像与原始图像在结构和视觉信息上越相似，修复质量越高，本文方法 SSIM 平均为 0.9515。本文方法与其他方法的对比如表 1 所示。

可以看出在 PSNR 和 SSIM 两个指标上本文方法均有提升，两个指标的提高说明本文方法在减小修复图像与原始图像之间的误差方面更有效，保持图像结构细节方面具有优势，更好地恢复了图像的原始

结构信息，提高了图像的整体质量。

4.3.3. 消融实验

为了验证本文方法各模块对模型的影响效果，对每一模块进行消融实验，Baseline 是基于基础的 Encoder-Decoder 结构的 GAN 网络，消融实验主观结果如图 7 所示，在只有 WGAN 模块情况下，恢复效果不佳，人物五官、皮肤模糊，且仍有遮挡痕迹。当加入 WGAN 模块后，PSNR 和 SSIM 分别提升至 23.83 dB 和 0.8482，修复图像的全局一致性有所改善，但细节恢复仍然不足。这一现象可能与 WGAN 的对抗训练机制有关，其虽然能够生成全局一致的图像，但在细节恢复上表现有限。单独使用 CSWinTF (CSWin-Transformer)模块时，PSNR 和 SSIM 进一步提升至 30.61 dB 和 0.9226，修复图像的纹理更加清晰，但在遮挡区域边缘仍存在少量伪影。这可能是由于 Transformer 模块在缺乏对抗训练的情况下，难以生成高质量的细节信息。

Table 1. PSNR and SSIM, compare results
表 1. 各模型方法 PSNR 和 SSIM 比较结果

方法	PSNR	SSIM
卷积神经网络	27.63	0.8674
生成对抗网络	30.1	0.8985
本文方法	35.62	0.9515

(a) 原图　(b) 遮挡人脸　(c)WGAN　(d)CSWinTF　(e) 本文方法

Figure 7. Subjective results of the ablation experiment
图 7. 消融实验主观结果

相比之下，本文方法在同时引入 WGAN 和 CSWinTF 模块后，修复效果显著提升，PSNR 和 SSIM 分别达到 35.62 dB 和 0.9515。从主观结果来看，修复图像在视觉上更加真实自然，细节特征也更加明显。例如，修复后的五官轮廓清晰，皮肤纹理细腻，遮挡区域与周围背景的过渡更加平滑。这一结果表明，WGAN 和 CSWinTF 模块在功能上具有互补性：WGAN 通过对抗训练提升了图像的全局一致性，而 CSWinTF 则通过自注意力机制捕捉了局部细节信息。

从表 2 的定量结果可以看出，在同一数据集上，随着不同模块的加入，模型的 PSNR 和 SSIM 指标均有显著提升。具体而言，WGAN 模块使 PSNR 和 SSIM 分别提升了 17.6%和 6.6%，而 CSWinTF 模块使 PSNR 和 SSIM 分别提升了 50.9%和 16.0%。当同时使用 WGAN 和 CSWinTF 模块时，PSNR 和 SSIM 的提升幅度最大，分别达到 75.8%和 19.6%。这一结果进一步验证了 WGAN 和 CSWinTF 模块对人脸修复效果的协同促进作用。

Table 2. Ablation experiment results
表 2. 消融实验结果

Baseline	WGAN	CSWinTF	PSNR	SSIM
√			20.27	0.7955
√	√		23.83	0.8482
√		√	30.61	0.9226
√	√	√	35.62	0.9515

5. 结论

针对于人脸遮挡修复问题，本文提出了基于生成对抗网络的的方法。模型采用 CSWin-Transformer 和 WGAN 相结合的策略，通过 WGAN 中的 Wasserstein 距离来解决模型在训练过程中训练不稳定、不收敛、模式坍塌等问题，同时利用 CSWin-Transformer Block 跨阶段窗口注意力机制提取图像的多尺度全局上下文特征，提升细节处理程度提高图形修复质量。实验结果显示，本文方法对于遮挡人复有良好的修复效果，生成图像真实自然，在修复细节以及完整度上优于对比方法。

基金项目

本文受到云南大学大学生创新创业训练项目(S202310673251)的资助。

参考文献

[1] 石雪梅, 朱克亮, 张祥民, 张树涛, 陈良锋. 基于生成对抗网络的有遮挡人脸修复方法[J]. 数据与计算发展前沿, 2022, 4(4): 123-131.

[2] Chen, D. and Hashimoto, T. (2004) Transmit Diversity Schemes for an Overloaded Space-Time Spreading System over a Flat Rayleigh Fading Channel. 2004 *IEEE* 15*th International Symposium on Personal*, *Indoor and Mobile Radio Communications*, Barcelona, 5-8 September 2004, 2664-2668.

[3] Barakova, E.I. and Lourens, T. (2004) Novelty Gated Episodic Memory Formation for Robot Exploration. 2004 2*nd International IEEE Conference on* '*Intelligent Systems*', Varna, 22-24 June 2004, 116-121. https://doi.org/10.1109/is.2004.1344648

[4] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

[5] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C. and Bengio, Y. (2014) Generative Adversarial Nets. In Neural Information Processing Systems, arXiv:1406.2661.

[6] Radford, A., Metz, L. and Chintala, S. (2015) Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434.

[7] Mirza, M. and Osindero, S. (2014) Conditional Generative Adversarial Nets. arXiv:1411.1784.

[8] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. and Efros, A.A. (2016) Context Encoders: Feature Learning by Inpainting. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 2536-2544. https://doi.org/10.1109/cvpr.2016.278

[9] Li, Y., Liu, S., Yang, J. and Yang, M. (2017) Generative Face Completion. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5892-5900. https://doi.org/10.1109/cvpr.2017.624

[10] 刘波宁，翟东海. 基于双鉴别网络的生成对抗网络图像修复方法[J]. 计算机应用, 2018, 38(12): 3557-3562, 3595.

[11] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X. and Huang, T.S. (2018) Generative Image Inpainting with Contextual Attention. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 5505-5514. https://doi.org/10.1109/cvpr.2018.00577

[12] Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., *et al*. (2022) Cswin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 12114-12124. https://doi.org/10.1109/cvpr52688.2022.01181

[13] Arjovsky, M. and Bottou, L. (2017) Towards Principled Methods for Training Generative Adversarial Networks. ArXiv abs/1701.04862.

[14] Arjovsky, M., Chintala, S. and Bottou, L. (2017) Wasserstein Gan. ArXiv abs/1701.07875.

[15] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. and Wang, M. (2021) Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In ECCV Workshops. arXiv:2105.05537.

[16] Taigman, Y., Yang, M., Ranzato, M. and Wolf, L. (2015) Web-Scale Training for Face Identification. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 2746-2754. https://doi.org/10.1109/cvpr.2015.7298891

[17] Gohring, M., Shulman, H. and Waidner, M. (2018) Path MTU Discovery Considered Harmful. 2018 *IEEE* 38*th International Conference on Distributed Computing Systems* (*ICDCS*), Vienna, 2-6 July 2018, 866-874. https://doi.org/10.1109/icdcs.2018.00088

[18] Sankur, B. (2002) Statistical Evaluation of Image Quality Measures. *Journal of Electronic Imaging*, **11**, 206-213. https://doi.org/10.1117/1.1455011

[19] Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, **13**, 600-612. https://doi.org/10.1109/tip.2003.819861