Software Engineering and Applications 软件工程与应用, 2023, 12(3), 444-454
Published Online June 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.123044

Hans汉斯

基于GAN和Transformer的人脸图像超分辨率重建

蒯新晨

上海理工大学，光电信息与计算机工程学院，上海

收稿日期：2023年4月7日；录用日期：2023年6月1日；发布日期：2023年6月14日

摘　要

为满足工业领域对低成本、高质量和大批量人脸图像的获取需求。本文提出了一种基于生成对抗网络和Transformer的超分辨率模型。在生成器方面，新设计了一种密集连接的Transformer结构替换了传统的卷积层，以建立全局特征依赖关系，从而提高特征提取能力和图像重建质量。同时，在鉴别器上采用判别能力更强的U-Net结构，以匹配生成器的性能。为了解决以往图像退化泛用性不足的问题，提出了一种图像退化模型，以实时生成训练图像对，大大丰富退化场景和数据集。为了更加细致地呈现人脸特征，所提出的模型还在自制数据集上进一步训练。实验结果表明，本文提出的模型相比其他模型在线条、纹理和清晰度等方面表现更好。

关键词

图像超分辨率，生成对抗网络，Transformer，人脸图像，图像退化

Super-Resolution Reconstruction of Face Images Based on GAN and Transformer

Xinchen Kuai

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Apr. 7th, 2023; accepted: Jun. 1st, 2023; published: Jun. 14th, 2023

Abstract

To meet the demand for low-cost, high-quality and high-volume face image acquisition in industry. In this paper, a super-resolution model based on generative adversarial network and Transformer

文章引用: 蒯新晨. 基于 GAN 和 Transformer 的人脸图像超分辨率重建[J]. 软件工程与应用, 2023, 12(3): 444-454.
DOI: 10.12677/sea.2023.123044

is proposed. For the generator, a new densely connected Transformer structure is designed to replace the traditional convolutional layer to establish global feature dependencies, thus improving the feature extraction capability and image reconstruction quality. Meanwhile, a U-Net structure with stronger discriminative ability is used in the discriminator to match the performance of the generator. To solve the problem of insufficient generalizability of previous image degradation, an image degradation model is proposed to generate training image pairs in real time, which greatly enriches the degradation scenes and datasets. The proposed model is further trained on a home-made dataset in order to present face features in more detail. Experimental results show that the proposed model in this paper performs better in terms of lines, textures and sharpness compared with other models.

Keywords

Image Super-Resolution, GAN, Transformer, Face Image, Image Degradation

1. 引言

超分辨率(Super Resolution, SR)重建旨在从低分辨率(Low Resolution, LR)图像重建出高分辨率(High Resolution, HR)图像。本文的研究动机在于为了满足对高质量人脸图像的获取需求，专注于通过算法模型提高人脸超分辨率重建的效果，以此降低更换高像素相机的成本。

目前，深度学习方法已被广泛应用于图像处理任务。自从文献[1]提出了一种三层结构的超分辨率模型后，又诞生了许多优秀的方法。文献[2]在一个很深的残差模型中，通过调整通道的权重来让模型突出关键部分。文献[3]在通道注意力的基础上使用中间状态矩阵模拟像素间的空间相关性。而文献[4]则通过残差和密集的连接方式加强信息传递，提高模型的泛化能力。这些方法利用残差连接、密集连接和注意力机制等结构，在峰值信噪比(Peak Signal to Noise Ratio, PSNR)和结构相似性(Structural Similarity, SSIM)等评估指标上均有显著提高。但是，它们通常采用 Bicubic 下采样来模拟退化后的训练数据集，这种退化方式是固定且理想的，并且简化了实际的退化场景，因此很难推广到实际应用中。

在实际情况中，受相机传感器质量、拍摄条件、存储方式以及分辨率变化等多种因素影响，图像的退化方式是未知的。为了更好地适应实际应用需求，一些研究人员正试图使用更为复杂的退化模型。文献[5]提出了一个通用的框架与退化策略，采取模糊内核和噪声水平两个退化因素。文献[6]认为真实的模糊核可以通过对图像中高度重复的块进行估计获得，并使用 GAN 来学习这种映射关系。文献[7]在双三次退化的 LR 图像中额外加入自然图像特征，并将这些数据用于 GAN 模型的训练，从而大大提高重建质量。合适的退化模型可能会带来更好的重建效果，但是过度退化会导致性能下降[8]。

针对现实中人脸图像的常见退化场景，本文设计了一种人脸图像退化模型，该模型采用随机退化算法对输入图像进行实时退化，并将其作为训练用的图像对输入超分辨率模型，以此扩展模型的映射能力。人脸图像包含丰富的纹理细节信息，而生成对抗网络(Generative Adversial Network, GAN)被广泛应用于恢复和重建任务。鉴于这点，本文提出了一种新的超分辨率模型 DSTGAN (Dense Swin Transformer Generative Adversial Network)。DSGAN 的生成器不再是传统的卷积，而是采用一种 Transformer 结构，将 GAN 与 Transformer 进行结合。更进一步的，本文通过密集连接多个 Transformer 层，提出了一种新的生成器单元

DST (Dense Swin Transformer)，利用密集连接高效地传递各层信息。同时，鉴别器方面引入性能更强的U-Net 结构，用于提高鉴别能力。

2. 相关工作

2.1. 生成对抗网络(GAN)

GAN 是一种强大的生成模型，已经在图像、语音和文本等领域有许多成功的应用。其中，SRGAN (Super-Resolution Generative Adversarial Network) [9]是首个将生成对抗网络用于超分辨率重建的方法，其生成器输入为 LR 图像，通过感知、对抗和内容损失函数，在与鉴别器对抗的过程中，获得高感知质量的输出。文献[10]提出的 ESRGAN 在 SRGAN 的基础上用残差密集块(Residual Dense Block, RDB) [4]作为生成器的基本单元，堆叠多个 RDB 提取深层特征，并使用激活前的 VGG 网络[11]计算感知损失。ESRGAN 不仅能够生成逼真纹理，而且避免了局部伪影，其感知质量达到了一个新的高度。

2.2. Transformer

Transformer 最初是为自然语言处理[12]构建的，通过多头自注意力模块(MSA)以及前馈多层感知机(MLP)层来捕获单词之间的长距离相关性。Transformer 采用序列表示特征，并建立了特征间的全局相关性。近期，Transformer 的新成果，例如 Vit [13]、IPT [14]和 ST [15]，已经显示出了在计算视觉领域的巨大潜力。在 ST 中，MSA 是按窗口计算的，相对于点到点的注意力计算，复杂度是线性的。通过 MSA 建立的全局依赖关系可以有效地学习图像的自相似信息。由于 ST 具有性能强、复杂度低和兼容性好的优点，文献[16]将 ST 用于图像超分辨率中，并在性能上超越了基于卷积的方法。

2.3. 多头自注意力机制(MSA)

MSA 的结构如图 1 所示。通过对输入进行多次不同的线性变换，然后将变换后的结果作为不同注意力头的输入。输入被划分为 h 个子空间，并对每个子空间计算查询向量 Q、键向量 K 和值向量 V。对于每个子空间 i，注意力计算表示为：

$$\text{head}_i = \text{Attention}\left(Q_i, K_i, V\right)_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}_i}\right)V \tag{1}$$

每个注意力头都会进行权重计算，其中涉及到 Q、K 和 V 三者的权重。通过计算 Q 和 K 之间的注意力权重矩阵以及与 V 的乘积，得到当前头的最终注意力输出。最后把所有注意力头的结果拼接，再将其乘上权重 W，以获得多头注意力的输出：

$$\text{MultiHead}\left(Q,K,V\right) = \text{Concat}\left(\text{head}_1,\cdots,\text{head}_h\right)W \tag{2}$$

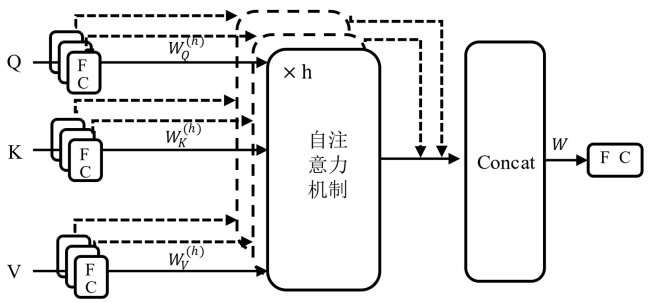

Figure 1. The structure of MSA
图 1. MSA 的结构

3. 本文方法

3.1. 模型结构

DSTGAN 的结构如图 2 所示，该模型由两部分组成：生成器和鉴别器。生成器接收 LR 图像作为输入，输出相应的 HR 图像。然后，将生成的 HR 图像和真实 HR 图像分别送入鉴别器进行真假判别。生成器与鉴别器不断进行对抗，直到达到纳什均衡。

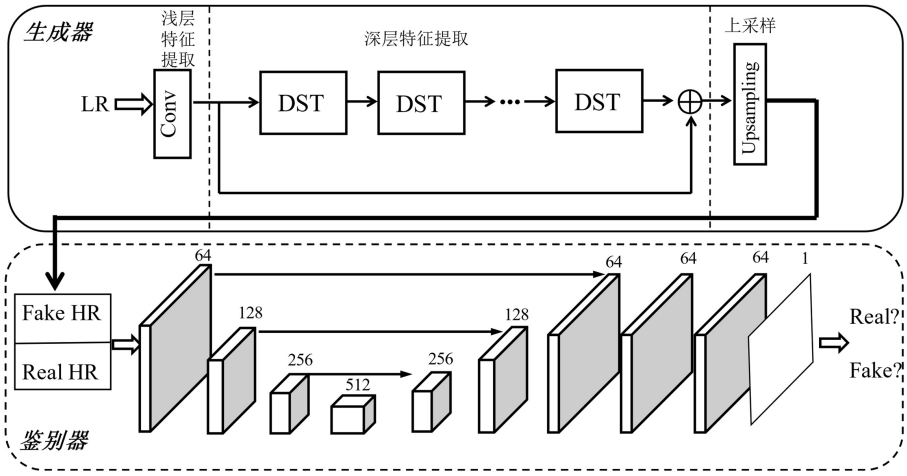

Figure 2. The structure of DSTGAN
图 2. DSTGAN 的结构

3.1.1. 生成器结构

生成器的结构如图 2 实线框所示，主要包含三个阶段：浅层特征提取、深层特征提取和上采样。在浅层特征提取阶段，生成器通过一个 3 × 3 卷积将输入的 LR 图像映射到一个更高维的特征空间。

深层特征提取阶段是生成器的主要阶段。DST 是生成器基本单元，其结构如图 3 所示，由 6 个 ST 和 3 个卷积层组成。ST 引入了文献[15]的一种 Transformer 结构，包含多头自注意力(MSA)和多层感知机(MLP)。在进行 MSA 和 MLP 之前，分别进行一次层归一化(Layer Normalization)，用于将每个输入的均值和方差归一化到相同的范围内，并且以残差的形式连接两部分的输入输出。

在 DST 中，每两个相邻的 ST 作为一个整体密集连接到后续每个卷积层中。值得注意的是，DST 中密集连接的信息来自 Transformer 结构，相比卷积层，Transformer 结构能够更好地关注全局信息。每次在密集连接后，卷积层会将先前层的信息融合，利用卷积的空间不变性增强 Transformer。此外，卷积层的另一个作用是通道进行归一化处理，以便将其输入到下一个 ST。一个 DST 模块总共存在三次密集连接，假设输入的特征图的大小为 $C \times W \times H$，其中 C 为通道，W 和 H 分别是宽度和高度。密集连接后的特征图大小分别为 $2C \times W \times H$、$3C \times W \times H$ 和 $4C \times W \times H$，最后，这些特征图都经过卷积层归一化到 $C \times W \times H$。值得一提的是，在输入和输出 ST 之前，分别进行一次 embedding 和 unembedding 操作，将输入数据在图像特征图和 Transformer 嵌入向量之间进行转换。

人脸包含了丰富的纹理信息，例如毛发、皮肤褶皱、斑点等，这些相似结构可能会在当前图像的全局范围内多次重现，这些重复的信息对重建具有鲁棒性。在深层特征提取阶段，通过堆叠多个 DST 模块，有助于充分提取特征和学习全局相似信息。

在生成器的上采样阶段，为了达到目标的超分辨率倍数，使用子像素卷积[17]的方法对特征图进行上

采样。该方法将通道上的像素排列到同一空间的大矩阵上，通过缩减通道数来扩大空间尺度。相较于其它上采样方法，亚像素卷积具有参数少、效率高和学习性强等优点。最终输出生成的 HR 图像。

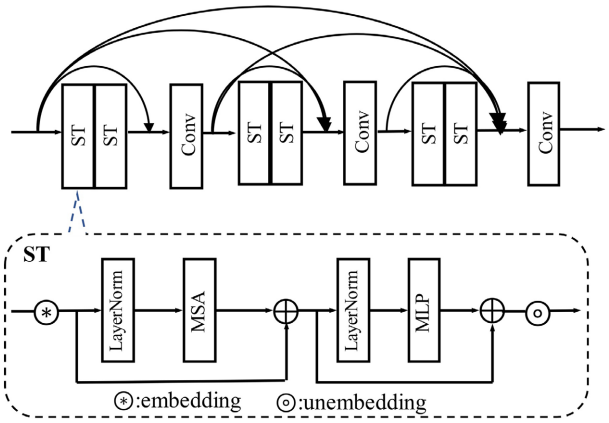

Figure 3. The structure of DST
图 3. DST 的结构

3.1.2. 鉴别器结构

关于鉴别器的选择，文献[9]和文献[10]都采用了基于 VGG 结构的鉴别器。虽然 VGG 结构在图像分类和识别方面有出色的表现，但其鉴别图像真假能力有限，无法满足本章对鉴别器性能的需求。因此，本文使用了最近工作中常用的 U-Net [18]，其结构如图 2 虚线框所示。U-Net 结构具有对边缘的敏感性，可以通过真假图像的细微边缘差异来更加准确地识别出真假图像。

在起始阶段，输入特征图经过 4 个卷积层的处理，通道数逐渐从 3 增加到 512。这些卷积层使用两倍步长的卷积核，每次特征图通道增加时，分辨率都会缩减到原来的四分之一。当通道数达到 512 后，每次进行卷积操作之前都会将特征图上采样(Bicubic 插值)两倍，然后通过卷积操作逐渐将通道数压缩到 1，这个过程正好与起始阶段相反。在 U-Net 的第 5、6、7 层的特征图分别与 3、2、1 层的对应尺度的特征图进行跨层连接，这种残差连接可以有效地融合浅层信息，有助于得到更精确的结果。为了增强训练的稳定性，避免出现过度的尖锐和伪影，每次卷积操作之后都会进行一次 LeakyRelu 激活和谱归一化。

3.1.3. 损失函数

DSTGAN 的损失函数包括内容损失、对抗损失和感知损失三部分。下面以 x 表示输入 LR 图像，$G(x)$ 表示生成器生成的 HR 图像，y 表示真实 HR 图像。

内容损失：L_1 损失广泛地使用在超分辨率任务中，它是一种一阶运算，对异常值的敏感度相对较低，另外计算简单、收敛速度快。L_1 损失通过绝对差值平均和的方式，计算了两个图像间像素级别的差异，内容损失 $L_{content}$ 表示如下：

$$L_{content} = \frac{1}{N}\sum_{(i,j)}\left\| G\left(x_{(i,j)}\right) - y_{(i,j)} \right\|_1 \tag{3}$$

其中 N 是特征图的像素数量，(i,j)为 $G(x)$和 y 中对应的坐标。

对抗损失：对抗损失的主要目的是让鉴别器无法区分 $G(x)$与 y，文献[10]中提到这种损失有助于学习更清晰的边缘和更精细的纹理。鉴别器和生成器各自的对抗损失分别表示为式 4 和式 5，这里计算的是一种相对概率。对于鉴别器来说，它的目标是最大化 $G(x)$与 y 的概率差异。而对于生成器，它的目标是最小化这个差异。

$$L_D^{Ra} = -E_y\left[\log\left(D_{Ra}\left(y, G(x)\right)\right)\right] - E_{G(x)}\left[\log\left(1 - D_{Ra}\left(G(x), y\right)\right)\right] \tag{4}$$

$$L_G^{Ra} = -E_y\left[\log\left(1 - D_{Ra}\left(y, G(x)\right)\right)\right] - E_{G(x)}\left[\log\left(D_{Ra}\left(G(x), y\right)\right)\right] \tag{5}$$

其中 $D_{Ra}(a,b) = D(a) - E_b\left[D(b)\right]$，$E_{G(x)}$和 E_y 分别计算了 $G(x)$和 y 所在 mini-batch 的平均值。

感知损失：感知损失可以帮助生成器学到更真实、语义性和结构性的特征。为此，引入了训练好的未激活的 VGG-19 模型，用于计算生成图像和真实图像之间的感知损失，其定义为两者的欧式距离：

$$L_{percep} = \frac{1}{N}\left\|\phi G(x) - \phi(y)\right\|_2^2 \tag{6}$$

其中 N 是特征图像素总数，ϕ 是特征提取函数。通过比较生成图像和真实图像之间的差异，促使它们向全局内容和结构相似的方向逐渐收敛。

DSTGAN 的生成器使用上面三种损失的加权和作为整体损失，表示为：

$$L_G = \alpha L_{content} + \beta L_G^{Ra} + \lambda L_{percep} \tag{7}$$

其中 α，β，λ 是权重系数。鉴别器的损失是对抗损失中的 L_D^{Ra}，在训练中，为了避免污染生成器的参数，鉴别器的反向传播与生成器的相互独立。

3.2. 人脸图像退化模型

在图 4 中，HR 和 Bicubic 两张图像视觉上差异较小，如果只使用 Bicubic 下采样建立 HR/LR 图像对，这样训练出的模型仅能学习有限的退化信息，无法充分发挥深度学习模型强大的预测能力。为此，可以进一步退化图像，以建立一个更广阔的从低质量到高质量的映射空间，从而让模型找寻更优的重建结果。

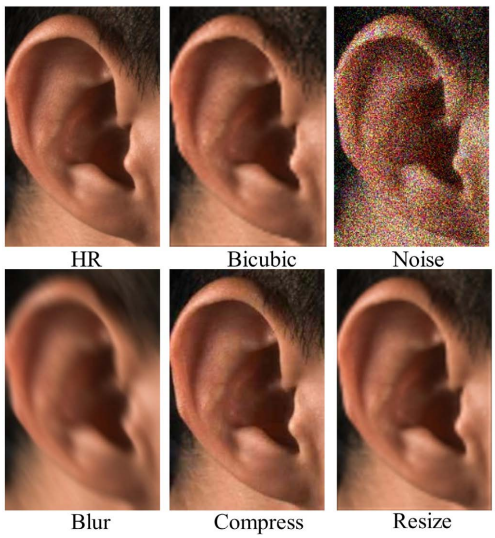

Figure 4. Effect of various degradations
图 4. 各种退化的效果

本文的退化模型有 4 种退化方式，分别是噪声、模糊、压缩和放缩。不同退化方式的效果也列在图 4 中，噪声为泊松噪声和高斯噪声，模糊是高斯滤波器，压缩是 JPEG 压缩，放缩通过多次尺度改变来实现。退化模型的退化策略是由本文提出的随机退化算法(详见算法 1)决定的，该算法主要考虑以下几个因素：1) 一张图像很难同时包含所有的退化情况，比较常见的退化有压缩和放缩，因此压缩和放缩的阈值

k_2 和 k_3 需要适当放大。而噪声的情况相对较少，因此噪声的阈值 k_0 被设置得比较小；2) 同一张图片在不同的退化顺序下得到的结果是不同的，因此随机退化算法一开始就设置一个随机序列来决定退化的顺序；3) 由于实际中退化的强度通常是正态分布的，因此对每种退化方式都设置了一个强度系数，且该系数是正态分布随机数。通过该随机退化算法，本文的退化模型可以生成理论上无限多的LR和HR的图像配对，这大大扩展了数据集。

算法 1: 图像随机退化算法

输入： HR 图像 x

1: A = [0, 1, 2, 3]

2: SET k_0, k_1, k_2, k_3 //定义概率阈值，范围 0~1

3: Shuffle(A) //打乱序列 A

4: FOR i IN A DO

5: k = rand() //定义随机数，范围 0~1

6: $\alpha, \beta, \gamma, \lambda$ = rand() //随机退化强度

7: IF A[i] == 0 AND $k < k_0$ THEN

8: Noise(x, α) END IF //加入噪声

9: IF A[i] == 1AND $k < k_1$ THEN

10: Blur(x, β) END IF //加入模糊

11: IF A[i] == 2AND $k < k_2$ THEN

12: Compress (x, γ) END IF //加入压缩

13: IF A[i] == 3AND $k < k_3$ THEN

14: Resize(x, λ) END IF //放缩

15: END FOR

16: Down sample to LR //下采样

输出： 退化后的 LR 图像

4. 实验部分

4.1. 数据集和评估指标

目前开源的高质量人脸图像数据集 FFHQ [19] (图 5 上层)。尽管图 5 是筛选出的高质量图像，但在细节方面仍有不足。经过退化处理后，图像细节进一步减少，这会影响模型对纹理细节的学习。因此，在本文工作中拍摄了大量超高分辨人脸图像用于训练超分辨率模型。在拍摄过程中，搭设了光场、幕布等设备，可以调节多种色温及亮度。使用海康威视 MV-CE200-10GM 工业摄像头拍摄了 100 张 2000 w 像素的超高分辨率的人脸图像，并使用康成 800 RYS-800WAF 拍摄了分辨率相对较低的 5 张 800 w 像素的图像用于图像测试。图 5 下层是截取的自摄图像，相比 FFHQ 数据集，人脸的纹理明显更加清晰了。此外，还使用了公开数据集：DIV2K、Set5、Set14、B100 以及 Urban100，表 1 给出了具体的数据集组成。

Figure 5. Comparison of FFHQ dataset and this paper's dataset
图 5. FFHQ 数据集和本文数据集对比

Table 1. Data set composition
表 1. 数据集组成

预训练数据集	DIV2K (800 张)
训练集	2000 w 像素人脸图像(100 张) 自采高分辨率图像(200 张)
测试图像	800 w 像素人脸图像(5 张)
测试集	Set5、Set14、B100 和 Urban100

4.2. 实施细节

在训练过程中，Batch Size 设置为 32，超分辨率的倍数固定为 4。输入的 HR 图像是随机裁剪的 64 × 64 大小的块，经过所提出的退化模型生成 LR 图像。DST 的窗口大小为 8。生成器总共包含 6 个 DST 模块，其损失函数的权重分别为 $\alpha=1,\beta=0.1,\gamma=1$。随机退化算法的参数为 $k_0=0.2$， $k_1=0.8$， $k_2=0.6$， $k_3=0.6$。使用 ADAM 优化器，其中 $\beta\_1=0.9$，$\beta\_2=0.999$，生成器和鉴别器的学习率均为 5e−3，并且每迭代 5000 轮衰减一半，总共训练 30,000 轮。

本文方法均使用 Pytorch 框架实现，并在操作系统为 Ubuntu v18.04 的服务器上训练，其中 CPU 为 Intel Xeon Gold 6140，内存为 64 GB，显卡为 6 块显存容量为 11264MB 的 NVIDIA GTX1080Ti。

4.3. 实验结果分析

图 6 给出了 DSTGAN 与其他模型的重建结果对比，包括 RDN [4]、ESRGAN [10]和 SwinIR [16]。从图中可以看出，RDN、ESRGAN 和 SwinIR 的表现差异不大。然而，这些模型在重建头发、睫毛等高频纹理方面存在困难，并且会出现明显的锯齿感，从而导致整体效果不够自然和真实。这些问题主要原因是 LR 图像中存在较多的缺失像素，导致线条的过渡不够光滑。仔细观察 RDN 和 ESRGAN 的结果，会发现图像上存在一些不自然的亮点和噪声，而 SwinIR 相对较少。这是因为 SwinIR 中采用基于全局的自注意力机制，考虑了自身的相似信息，避免了这部分的不自然，这也是本文选用 ST 结构的原因。

图 6 最后一列对应 DSTGAN 的重建结果。DSTGAN 对 LR 图像中无法辨认的模糊区域进行了光滑处理，并消除了模糊感，从而使毛发的边界更加清晰，视觉上更加流畅和连贯。在处理皮肤的纹理时，通过突出皮肤的褶皱来增强其视觉效果。综合来看，DSTGAN 在纹理和细节方面的恢复效果更精细，其重建质量和视觉观感优于其他模型。

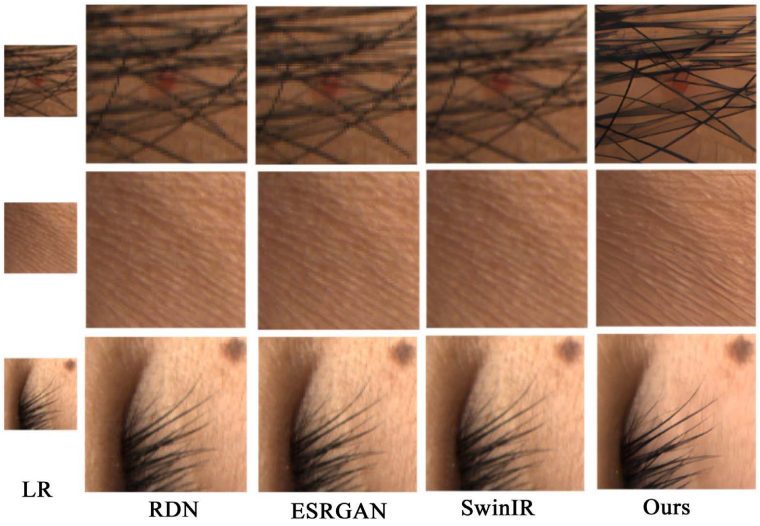

Figure 6. Partial comparison between the results of this model and other models
图 6. 本文模型和其他模型的结果局部对比

4.4. 消融实验

4.4.1. 退化模型有效性分析

为了验证所提图像随机退化模型的有效性，进行了多项消融实验，针对不同退化强度进行了比较。实验结果如图 7 所示，其中，通用的 Bicubic 下采样方法作为弱退化模型，强度系数减半的图像随机退化模型作为等半退化模型，以及实际的随机退化模型。从图中可以看出，弱退化模型视觉效果与上一节的 RDN 和 ESRGAN 类似，整体上具有较强的模糊感，边缘细节不够清晰。等半退化模型的效果有所改善，但仍存在一定的朦胧感。相比之下，本文实际的退化模型表现相对更好，进一步提升了头发和皮肤纹理的重建质量，在保留细节的同时带来更具有鉴别性的效果。

Figure 7. Degradation model validity ablation experiment
图 7. 退化模型有效性消融实验

4.4.2. DST 有效性分析

为了验证 DST 结构的有效性，比较了 RDB、ST 以及 DST 三种生成器基本单元的性能，使用 PSNR 和 SSIM 这两个指标来定量评估性能，其中 PSNR 和 SSIM 数值均越大性能越好。为了减少退化模型的随机性带来的影响，采用 Bicubic 方法进行下采样，即作为一个通用的模块来验证效果，结果如表 2 所示。将生成器的基本单元从 RDB 更换为 ST 后，PSNR 提升了 0.36 dB。相比文献[10]中使用的 RDB，ST 建立了全局上下文依赖关系，因此表现更佳。表中最后一行是 DST 的评估结果，该结构在 ST 的基础上进一步提高了性能，证明 DST 中的密集连接是有效的。

Table 2. Comparison experiments of different generator basic units
表 2. 不同生成器基本单元对比实验

基本单元	PSNR	SSIM
RDB	27.59	0.7873
ST	27.95	0.8096
DST	28.01	0.8118

4.4.3. DSTGAN 整体有效性分析

为了进一步评估 DSTGAN 模型的整体有效性，在基准数据集 Set5、Set14、B100 和 Urban100 上将 DSTGAN 与目前流行的其他几个生成对抗网络模型进行了× 4 超分辨率的定量比较，包括 SRGAN [9]、ESRGAN [10]和 NatSRGAN [20]。由于退化模型对 PSNR 及 SSIM 指标的影响较大，为了公平起见，DSTGAN 使用与其他几种模型相同的退化方式，即 Bicubic 下采样，实验结果如表 3 所示。DSTGAN 在所有测试集上都取得了最好成绩。与次好性能的 NatSRGAN 相比，DSTGAN 在所有数据集上的 PSNR 分别提升了 0.16 dB、0.36 dB、0.23 dB 及 0.27 dB，在 SSIM 指标上分别提升了 0.0049、0.0142、0.0127 及 0.0104。此外，相比第一个生成对抗网络方法 SRGAN，DSTGAN 性能的提升较大，特别是在 Urban100 数据集上相比 SRGAN 在 PSNR 和 SSIM 指标上分别提升了 2.19 dB 和 0.0794。实验结果表明，DSTGAN 作为一个普通的生成对抗网络超分辨率模型依旧有着出色的性能。

Table 3. Quantitative comparison of DSTGAN and other models (PSNR/SSIM)
表 3. DSTGAN 和其他模型的定量比较(PSNR/SSIM)

模型	Set5	Set14	B100	Urban100
SRGAN [9]	29.41/0.8345	26.02/0.6934	24.93/0.6401	23.54/0.6912
ESRGAN [10]	30.32/0.8474	26.41/0.7159	24.48/0.6184	24.36/0.7208
NatSRGAN [20]	30.98/0.8606	27.42/0.7329	26.44/0.6827	25.46/0.7602
DSTGAN (Ours)	**31.14/0.8655**	**27.78/0.7471**	**26.67/0.6954**	**25.73/0.7706**

5. 结束语

在本文中，结合生成对抗网络和 Transformer 模型，构建了一种用于人脸图像超分辨率重建模型 DSTGAN。使用 DST 模块作为生成器的基本单元，以提升模型对深层特征的提取能力。此外，提出了一种图像退化模型，用于实时生成训练的图像对，从而让模型学习更广泛的映射关系。通过主观视觉评价，所提模型在视觉感知方面表现最佳，并通过多项消融实验验证了模型的有效性。

综上所述，所提模型可用于生成具有良好视觉感知的人脸图像，并可用于实际场景的图像优化。但是，目前该模型对算力和存储资源的要求较高，限制了其应用场景。因此，在未来的工作中，将致力于构建一个可移植性强的模型，并且提高对其他图像类别的适用性。

参考文献

[1] Dong, C., Loy, C.C., He, K. and Tang, X. (2015) Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 295-307. https://doi.org/10.1109/TPAMI.2015.2439281

[2] Zhang, Y., Li, K., Li, K., *et al.* (2018) Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018. *ECCV* 2018.

Lecture Notes in Computer Science, Vol. 11211, Springer, Cham, 286-301. https://doi.org/10.1007/978-3-030-01234-2_18

[3] Woo, S., Park, J., Lee, JY. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision—ECCV* 2018. *ECCV* 2018. *Lecture Notes in Computer Science*, Vol. 11211, Springer, Cham, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[4] Zhang, Y., Tian, Y., Kong, Y., Zhong, B. and Fu, Y. (2018) Residual Dense Network for Image Super-Resolution. *Proceedings of* 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 2472-2481. https://doi.org/10.1109/CVPR.2018.00262

[5] Zhang, K., Zuo, W. and Zhang, L. (2018) Learning a Single Convolutional Super-Resolution Network for Multiple Degradations. *Proceedings of the* 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3262-3271. https://doi.org/10.1109/CVPR.2018.00344

[6] Bell-Kligler, S., Shocher, A. and Irani, M. (2019) Blind Super-Resolution Kernel Estimation Using an Internal-GAN. https://arxiv.org/abs/1909.06581

[7] Fritsche, M., Gu, S. and Timofte, R. (2019) Frequency Separation for Real-World Super-Resolution. 2019 *IEEE*/*CVF International Conference on Computer Vision Workshop* (*ICCVW*), Seoul, 27-28 October 2019, 3599-3608. https://doi.org/10.1109/ICCVW.2019.00445

[8] Efrat, N., Glasner, D., Apartsin, A., Nadler, B. and Levin, A. (2013) Accurate Blur Models vs. Image Priors in Single Image Super-Resolution. *Proceedings of* 2013 *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 2832-2839. https://doi.org/10.1109/ICCV.2013.352

[9] Ledig, C., Theis, L., Huszár, F., *et al*. (2017) Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *Proceedings of* 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 105-114. https://doi.org/10.1109/CVPR.2017.19

[10] Wang, X., Yu, K., Wu, S., *et al*. (2019) ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In: Leal-Taixé, L. and Roth, S., Eds., *Computer Vision—ECCV* 2018 *Workshops*. *ECCV* 2018. *Lecture Notes in Computer Science*, Vol. 11133, Springer, Cham, 63-79. https://doi.org/10.1007/978-3-030-11021-5_5

[11] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/abs/1409.1556

[12] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. https://arxiv.org/abs/1706.03762

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929

[14] Chen, H., Wang, Y., Guo, T., *et al*. (2021) Pre-Trained Image Processing Transformer. *Proceedings of* 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 12294-12305. https://doi.org/10.1109/CVPR46437.2021.01212

[15] Liu, Z., Lin, Y., Cao, Y., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *Proceedings of* 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[16] Liang, J., Cao, J., Sun, G., *et al*. (2021) Swinir: Image Restoration Using Swin Transformer. *Proceedings of* 2021 *IEEE*/*CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October 2021, 1833-1844. https://doi.org/10.1109/ICCVW54120.2021.00210

[17] Shi, W., Caballero, J., Huszár, F., *et al*. (2016) Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *Proceedings of* 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 1874-1883. https://doi.org/10.1109/CVPR.2016.207

[18] Schönfeld, E., Schiele, B. and Khoreva, A. (2020) A U-Net Based Discriminator for Generative Adversarial Networks. *Proceedings of* 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 8204-8213. https://doi.org/10.1109/CVPR42600.2020.00823

[19] Karras, T., Laine, S. and Aila, T. (2019) A Style-Based Generator Architecture for Generative Adversarial Networks. *Proceedings of* 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4396-4405. https://doi.org/10.1109/CVPR.2019.00453

[20] Soh, J.W., Park, G.Y., Jo, J. and Cho, N.I. (2019) Natural and Realistic Single Image Super-Resolution with Explicit Natural Manifold Discrimination. *Proceedings of* 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 8114-8123. https://doi.org/10.1109/CVPR.2019.00831