南京農業大學

本科生毕业论文（设计）

题　　目：面向真实场景的低质图像超分辨率重建方法研究与实现

姓　　名：李云帆

学　　号：9203010809

学　　院：人工智能学院

专　　业：电子信息科学与技术

指导教师：王洁　　职称　副教授

2024 年 5 月 13 日

南京农业大学本科生毕业论文（设计）原创性声明

本人郑重声明：所呈交的毕业论文（设计），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

论文作者签名：李云帆　　日期：　2024 年　5 月　13 日

南京农业大学本科生毕业论文（设计）使用授权声明

本学位论文作者完全了解学校有关保留、使用毕业论文（设计）的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权南京农业大学教务处可以将本毕业论文（设计）的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编毕业论文（设计）。

论文作者签名：李云帆　　导师签名：

日期：　2024 年　5 月　13 日　　日期：2024 年　5 月　13 日

面向真实场景的低质图像超分辨率重建方法研究与实现

摘　要

图像超分辨率是机器视觉的热门方向，在生活和生产中的多个领域均具有重要意义。传统超分辨率技术基于插值算法实现，难以兼顾推理速度和推理效果，且无法很好地重构出原图像的细节。机器学习的引入提供了新的研究方向，但在真实场景的应用中还存在着数据集难以获取、没有结合图像先验知识以及伪影等问题。针对上述问题，本文采用以人工合成数据为主，真实退化数据为辅的方法产生数据集，构建面向真实场景的低质图像超分辨率生成对抗网络。通过引入高阶退化模型，只使用原生图像作为数据集训练模型。实验结果表明该方法在结构相似度上优于传统方法，但其他性能指标表现一般，即只使用纯合成数据的模型其性能有限。同时提出了一种利用真实退化生成对照数据的方法，并利用对照数据集加强训练上述模型。结果表明加强训练后的模型在各评价指标上都得到了明显的提升，但难以正确处理文本内容。最后对其超分辨率结果进行伪影检测。结果表明在极端情况下该模型会产生大量伪影，与具备抗伪影特性强化的同类模型相比仍有差距。

关键词：机器视觉；真实场景；超分辨率重建；生成对抗网络

RESEARCH AND IMPLEMENTATION OF LOW-QUALITY IMAGE SUPER-RESOLUTION RECONSTRUCTION METHODS FOR REAL-WORLD SCENARIOS

ABSTRACT

Image super-resolution is a hot direction in machine vision, which has significant importance in multiple areas of life and production. Traditional super-resolution techniques are based on interpolation algorithms, which struggle to balance inference speed and effect, and cannot well reconstruct the details of the original image. The introduction of machine learning provides a new research direction, but there are still problems in the application in real scenarios, such as difficulties in obtaining datasets, not combining image prior knowledge, and artifacts. To address these issues, this paper uses a method of generating datasets with artificially synthesized data as the main and real degraded data as the auxiliary, and constructs a low-quality image super-resolution generative adversarial network for real scenarios. By introducing a high-order degradation model, only native images are used as datasets to train the model. Experimental results show that this method is superior to traditional methods in terms of structural similarity, but other performance indicators are average, that is, the performance of models using only pure synthetic data is limited. At the same time, a method of generating control data using real degradation is proposed, and the control dataset is used to enhance the training of the above model. The results show that the model after enhanced training has significantly improved in all evaluation indicators, but it is difficult to correctly handle text content. Finally, the super-resolution results are subjected to artifact detection. The results show that this model will produce a large number of artifacts under extreme conditions, and there is still a gap compared with similar models with enhanced anti-artifact characteristics.

KEY WORDS： Computer vision; Real scene; Super resolution reconstruction; Generative adversarial networks

目　录

第一章 绪论

1 研究背景及意义

图像超分辨率（Super-Resolution, SR）重建技术是目前机器视觉研究的热点之一。SR是一种将一幅或多幅低分辨率（Low Resolution, LR）图像通过某些手段估计得到对应的高分辨率（High Resolution, HR）图像的技术。其在军用侦察、遥感成像、医学核磁共振成像、视频监控系统及标清视频信号在高清电视上的显示等方面有着广泛的应用需求。譬如以下具体应用场景：当照片、录像等分辨率较低、质量不高或传输带宽有限时，可通过 SR 技术改善图像质量；当进行计算机图像（Computer Graphics, CG）渲染和游戏渲染，而硬件计算能力有限时，可先输出低质量（Low Quality, LQ）原生画面，再进行 SR，得到接近原生高质量画面的输出，从而降低了硬件要求。在消费电子市场上，以 NVIDIA 的 DLSS（Deep Learning Super Sampling）和 AMD 的 FSR（FidelityFX Super Resolution）为代表的 SR 技术也备受推崇。

在 SR 任务中，一个 LR 图像可能对应多个 HR 结果；而在真实场景中，只有特定的结果是符合要求的，因此 SR 是一个典型的不适定问题。在部分场景中已有的 SR 方法已经可以实现较为优秀的效果，例如由 BiliBili AI 团队开发的 Real-CUGAN（Real Cascade U-Nets Generative Adversarial Network）模型[1]针对漫画作品展现出了优秀的性能。但真实场景下从 HR 图像退化到 LR 图像的过程通常是未知且复杂的。例如，从人眼看到的场景到相机拍摄的画面；手持相机的抖动、光线通过相机镜头时的折射与反射、相机 CMOS（Complementary Metal-Oxide-Semiconductor）捕捉光线时的噪点以及 DSP（Digital Signal Processing）对图像的处理等都会造成人眼观察到的图像和照片之间差距；而一张照片在经过社交媒体的多次压缩后，文件体积会大幅下降，像素点和颜色信息又发生了丢失。显然，一种通用且高效的 SR 方法需要充分考虑所有这些因素，寻找最符合真实场景的解。例如，在遥感图像中，依靠硬件升级改善图像分辨率需要投入较大的成本；但使用深度学习对遥感图像进行 SR 又需要大量对应的 HR 与 LR 训练数据集，而这对于遥感图像而言是不现实的[2]。由于缺少训练数据集而难以构建深度学习网络是真实场景 SR 任务中普遍存在的问题。因此，提出一种不完全依赖数据集，而是结合实际退化流程等先验知识的 SR 方法具有显著的研究意义和实际应用价值。

2 超分辨率重建技术发展现状

传统 SR 技术基于插值或重建实现。插值算法使用真正意义上的“算法”来估算低分辨率图像中缺失的像素点。计算效率高、硬件负担低的插值算法在面对复杂场景时其效果往往一般。基于重建的 SR 技术将观测到的 LR 图像看作是 HR 图像经过变形、模糊、降采

样和噪声污染等退化形成的，在退化模型的约束下由多幅 LR 观测图像来逆推 HR 图像；但常规的退化模型不够拟真，不足以描述真实场景中的复杂退化过程。基于机器学习（Machine Learning, ML）的 SR 技术能对复杂场景下的图片展现出良好的超分效果，但需要大量的训练数据集。

2.1 基于插值的方法

插值方法的基本原理是根据目标位置周围像素点的 RGB（Red, Green, Blue）值，估算目标位置像素点的 RGB 值，使得图像的颜色过渡较为平滑。按照估算方法的不同，可将较为常见的传统插值算法分为最近邻插值（Nearest-Neighbor Interpolation, NN）、区域缩放（Area Resize）、双线性插值（Bilinear Interpolation）和双三次插值（Bicubic Interpolation）。即便使用神经网络进行图像 SR，依然离不开传统插值算法作为基础。

最近邻插值是最简单的插值算法。对于新图像中的每个像素点，首先根据缩放倍数找到其在原始图像中对应的像素点。对于坐标为(i,j)的某一输出图像像素点，给定缩放倍数m为原图像的宽与新图像的宽之比，则坐标(i,j)的颜色应与它在原图中的对应像素点(x,y)一致，(x,y)的计算方式如公式 1-1：

$$\begin{cases} x = round(i \cdot m) \\ y = round(j \cdot m) \end{cases} \tag{1-1}$$

其中$round$为四舍五入取整函数。最近邻插值的计算量最小，推理速度最快，对图像原有结构的破坏较少，但图片观感不佳，颜色突变处容易产生锯齿状的缺陷。

区域缩放综合考虑了每个像素点周围的其他像素点提供的信息。使用区域缩放进行降采样时，若给定缩放倍数m，则将每个通道的图像矩阵分割成若干个大小为$m \times m$的区块，求得每个区块内的平均值；将每个区块视作一个像素输出，用区块的平均值作为对应输出像素的颜色强度，再将三个通道叠加即可得到最终结果。由于区域缩放在进行升采样时本质上是一种线性插值，因此通常只用于降采样中。神经网络中常用的平均池化层，实质上就是基于区域缩放的降采样。

双线性插值是对每个像素点在两个方向上依次插值得到最终结果。首先找到与原图中的理论位置相邻的四个像素点；再假定颜色强度随距离线性变化，利用这四个像素点分别进行两次线性插值得到该处对应的颜色强度。对于坐标为(i,j)的某一输出图像像素点，给定缩放倍数m，则相邻的四个像素点的坐标(x_1,y_1)，(x_1,y_2)，(x_2,y_1)，(x_2,y_2)满足式 1-2：

$$\begin{cases} x_1 = floor(i \cdot m) \\ y_1 = floor(j \cdot m) \\ x_2 = ceil(i \cdot m) \\ y_2 = ceil(j \cdot m) \end{cases} \tag{1-2}$$

其中$floor$为“去尾”取整函数，$ceil$为“进一”取整函数。设$f(x,y)$为点(x,y)处的颜色强度，先在x方向上进行线性插值得式 1-3：

$$\begin{cases} f(x,y_1) \approx \frac{x_2-x}{x_2-x_1} f(x_1,y_1) + \frac{x-x_1}{x_2-x_1} f(x_2,y_1) \\ f(x,y_2) \approx \frac{x_2-x}{x_2-x_1} f(x_1,y_2) + \frac{x-x_1}{x_2-x_1} f(x_2,y_2) \end{cases} \tag{1-3}$$

再在y方向上进行线性插值得式 1-4：

$$f(x,y) \approx \frac{y_2-y}{y_2-y_1} f(x,y_1) + \frac{y-y_1}{y_2-y_1} f(x,y_2) \tag{1-4}$$

即可得到目标像素点处的颜色值。与最近邻插值相比，双线性插值的图片观感更好，减少了锯齿的产生，同时计算量不高；但由于假定了颜色是在图像中线性变化的，得到的 SR 结果往往会破坏原图的结构，与真值（Ground Truth, GT）相比产生较大的偏差，不能满足真实场景 SR 任务的需求。

双三次插值在双线性插值的基础上扩大了统计的范围：双线性插值法考虑与理论目标像素点相邻的四个像素点，而双三次插值考虑的是周围的 16 个像素点，并创建了单独的双三次基函数计算每个像素点对目标位置的权重，最后通过加权计算得到目标位置的颜色强度。基函数如式 1-5 所示：

$$W(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^2 + 1 & \text{for } |x| \le 1 \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a & \text{for } 1 < |x| < 2 \\ 0 & \text{otherwise} \end{cases} \tag{1-5}$$

通常 a 的取值为-1 或-0.5。若新图像中的坐标(i,j)对应在原图像中的理论坐标为(x_0,y_0)，则公式 1-4 中的 x 为(x_0,y_0)周围的 16 个点到该点的横坐标距离或纵坐标距离。由于新图像中除边缘外的每个像素点都是基于原图像中的 16 个像素点加权得到的，因此双三次插值能够更好地保留原图像中的结构，SR 效果最好，同时计算量也最大。若将 LQ 图像的长和宽均放大四倍，则新图像的像素点数量是原图像的 16 倍，且新图像中的每个像素点都需要遍历原图像中对应的 16 个像素点，因此需要处理的像素点数量实际上是原图像的 256 倍。

综上，传统插值算法虽然在实现原理上较为简单，也容易移植，但往往无法兼顾计算速度和 SR 效果。在实际应用中，使用双三次插值生成的图像也不够平滑，观感欠佳[3]。传统插值算法并没有考虑 LQ 图像是如何由高质图像退化得到的，其 SR 结果只是依赖于纯粹的数学估计，通常只能恢复低频信息而无法恢复高频信息[4]，难以应用于现实场景中。

2.2 基于重建的方法

重建方法的基本思想是构建一种退化模型，用于描述从 HR 图像退化到 LR 图像的过程。传统的重建方法可分为频域法和空域法。频域法致力于通过消除频谱混叠来提高图像分辨率，计算量小且方法直观，但只适用于全局平移运动和线性空间不变的降质模型，对先验知识的应用有限。空域法考虑了多种先验知识，适用范围更广，同时计算量也更大。例如，使用稀疏编码进行 SR[5]，要先对 HR 图像进行退化与插值合成 LR 图像，再进行联合字典训练得到从 LR 图像到 HR 图像的对应关系。退化流程的拟真程度直接影响重建的 SR 效果。同样地，使用合成 LR 数据参与训练的神经网络也需要一种优秀的退化模型。

经典退化模型[6]认为低分辨率图像可以被看作是由高分图像先与模糊核进行卷积，再进行下采样，然后加入噪声，最后进行 JPEG 压缩的结果；如下式 1-6 所示：

$$x = D(y) = [(y \otimes k) \downarrow_r + n]_{\text{JPEG}} \tag{1-6}$$

其中k为模糊核，通常选用尺寸为$2t+1$的高斯模糊，即卷积核内的元素$(i,j) \in [-t,t]$符合高斯分布，如式 1-7 所示：

$$k(i,j) = \frac{1}{N}\exp\left(-\frac{1}{2}C^T\Sigma^{-1}C\right), C = [i,j]^T \tag{1-7}$$

其中C表示空间坐标，Σ为协方差矩阵，N为归一化常数。回到公式 1-7 中，$\downarrow_r$表示降采样，经典退化模型中通常随机选取区域插值、双线性插值和双三次插值作为降采样的方式。n为加性噪声，通常选用可加的高斯噪声和泊松噪声。JPEG 压缩是最常用的图像压缩方式之一，本文选取可微 JPEG 压缩算法[7]：首先将 RGB 色彩转换为 YCbCr 色彩，其中 Y 表示亮度，Cb 和 Cr 分别为蓝色色度与红色色度；再对 Cb 和 Cr 通道进行降采样，随后把每个通道分割成 8×8 的区块，并对每个区块进行离散余弦变换。再对变换后的结果进行量化，此时有大量不重要的数据被丢失，最后进行微分近似。JPEG 压缩后的图片通常能保留人眼可识别的细节，但在某些区域可能产生色块。

高阶退化模型由 Real-ESRGAN（Real- Enhanced Super-Resolution Generative Adversarial Network）[8]引入。考虑到网络过于复杂会严重降低训练速度，因此在实际应用中通常使用二阶退化，流程如下图 1-1 所示。

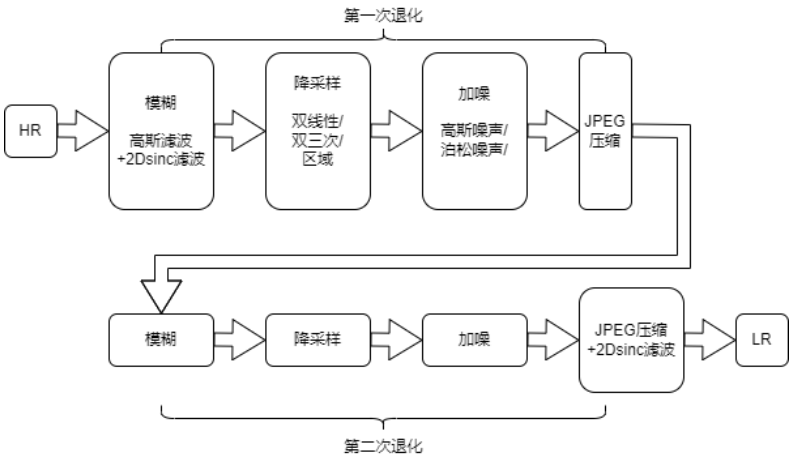


图 1-1　高阶退化流程图

Fig.1-1　High order degradation flow chart

与经典退化模型相比，高阶（二阶）退化模型不止是单纯重复了一次原有的退化流程，还在第一次模糊和第二次 JPEG 压缩中加入了二维 sinc 滤波器，如式 1-8 所示：

$$k(i,j) = \frac{\omega_c}{2\pi\sqrt{i^2+j^2}}J_1\left(\omega_c\sqrt{i^2+j^2}\right) \tag{1-8}$$

其中i和j为核内坐标，ω_c为滤波器截止频率，J_1为第一类贝塞尔函数。引入 sinc 函数是为

了消除高频信息被过度放大而导致的过冲效应（Overshoot Artifact），从而抑制伪影的产生。

虽然基于重建的方法已经不再是 SR 研究的重心，但利用退化模型合成 LR 数据的思路在基于 ML 的方法中得到了广泛地应用。

2.3 基于机器学习的方法

较为经典的基于 ML 的 SR 技术有基于卷积神经网络（Convolutional Neural Network, CNN）的方法、基于残差神经网络（Residual Neural Network, ResNet）的方法和基于生成对抗网络（Generative Adversarial Network, GAN）的方法等。

基于 CNN 的 SR 方法有 SRCNN（Super-Resolution Convolutional Neural Network）[9]、ESPCN（Enhanced Super-Resolution Convolutional Network）[10]、RCAN（Residual Channel Attention Networks）[11]和 MSFIN（Multi-Scale Feature Interaction Network）[12]等。SRCNN 是第一个 SR 深度学习网络，结构简单且完全基于数据驱动，没有结合图像中的先验知识，感受野较小。其主要缺陷在于使用插值法将 LR 图像插值到 HR 的尺寸，然后再用 CNN 进行重建，网络计算量大，且只能对相同缩放倍数的图像进行超分。2014 年，Twitter 公司的开发者提出 ESPCN 方法，正式发表于 2016 年的 IEEE CVPR（IEEE Conference on Computer Vision and Pattern Recognition）会议上。ESPCN 基于像素重排列，无需对 LR 上采样，使用亚像素卷积层将 LR 图像恢复到目标分辨率。2018 年，美国东北大学的研究人员提出了 RCAN 方法，旨在引入注意力机制。2021 年，南京邮电大学和华东师范大学的研究者在 IEEE ICME（IEEE International Conference on Multimedia & Expo）会议上展示了移植 SR 网络到移动端设备的轻量级方法 MSFIN。基于 ResNet 的 SR 方法相对少见，2016 年韩国首尔大学学者在当年的 IEEE CVPR 上首次提出利用 ResNet 进行超分，称为 VDSR（Very Deep Super-Resolution）[13]。从网络结构来看，CNN 结构相对简单，容易实现；但卷积层的感受野有限，因此 SR 效果一般。此外，CNN 高度依赖完整的数据集，受限较多。

基于 GAN[14]的方法是目前的主流之一，包括 SRGAN（Super-Resolution Generative Adversarial Network）[15]、ESRGAN（Enhanced Super-Resolution Generative Adversarial Network）[16]、BSRGAN（Blind Super-Resolution Generative Adversarial Network）[17]和 Real-ESRGAN 等。SRGAN 是第一个用于 SR 的 GAN 网络，与其他网络相比更加关注图像的先验知识，而不是只关注像素点间的 RGB 值差异：如 HR 图像中物体的边缘在 LR 图像中会出现细节丢失，而这种丢失在 LR 图像进行 SR 时难以找回。ERSGAN 由中科大、国科大、港中大和南洋理工的多位学者提出，在 SRGAN 的基础上进行了改进，通过修改网络结构减少了 SRGAN 中产生的伪影（artifact）；并使用了更合理的损失函数，使得网络生成的图片不会丢失大量的锐度细节。2021 年，苏黎世理工的张凯等学者提出了著名的 BSRGAN，将针对真实场景的“盲超分”[18]作为主要目标，并取得了优秀的 SR 效果。BSRGAN 和 Real-ESRGAN 这类方法在训练时不直接使用包含 HR 图像及其对应 LR 图像的训练数据集；而

是只收集 HR 图像，通过退化模型合成 LR 数据参与训练。这种方法绕开了真实场景 SR 应用中普遍存在的数据收集困难的问题，大大降低了前期的工作量；但只使用合成数据作为 LR 的做法是否完全可靠，即退化模型是否足够具有代表性，仍缺乏检验。GAN 的对抗结构使其生成的图像往往具有更丰富的高频细节，观感更好，但可能会在训练和推理的过程中产生过冲效应[19]。检测并消除 GAN 模型 SR 结果中的伪影也因此成为了近期重要的研究方向之一。

3 论文主要研究内容及章节安排

本文共包含四章：

第一章是绪论，阐述本课题的研究背景和意义以及 SR 技术的发展现状和对应的技术细节，并给出主要研究内容和章节安排。

第二章提出了基于生成对抗网络的图像超分辨率模型，包括数据集的构成和处理方式、生成模型和鉴别模型的结构以及评估指标和损失函数的选用。研究内容如下：

(1) 研究使用纯合成数据作为 LR 数据的面向真实场景的 LQ 图像 SR 网络。只使用原生 HR 图像作为数据集的 GT，通过高阶退化模型合成 LR 数据，绕开采集对照数据的这一难题；并综合 L1 损失、感知损失和对抗损失作为网络训练时的损失函数。

(2) 研究合成真实退化对照数据的方法，并使用该方法产生少量真实退化数据集，用于对(1)中模型进行加强训练。

第三章进行了实验并对结果进行总结与分析，以证实方法的有效性和优越性，同时探讨其局限性。其内容包括：

(1) 完成模型预训练和正式训练，并对其 SR 效果进行性能评估。与传统插值方法相比，该方法的 SR 结果观感提升明显，结构相似性保持的较好，但在 L1 损失、L2 损失和峰值信噪比上均弱于其他方法。上述结果表明使用纯合成数据作为 LR 训练神经网络具有一定的局限性，但加入感知损失有助于提升 SR 结果与原 GT 值的结构相似程度。因此，该 GAN 网络存在进一步研究的价值，但在某些方面仍需优化。

(2) 加入真实退化数据完成强化训练，并对其 SR 性能进行检测。实验结果表明，强化训练后的模型在各项评价指标上都得到了明显的提升，不仅在性能指标上与双三次插值的表现持平，而且拥有更平滑的图像观感以及短得多的推理时间。但该 GAN 模型在对文本内容进行 SR 时可能失效。

(3) 研究上述加强训练后的 GAN 模型的伪影问题，对其 SR 结果进行伪影检测。实验结果表明，在某些极端情况下该模型会产生大量的伪影，并丢失部分 GT 中真实存在的纹理，与具备抗伪影特性强化的 GAN 模型相比仍有差距。

第四章是全文总结与展望，对全文内容进行总结，并对下一步的研究方向进行分析与展望。

第二章　基于生成对抗网络的图像超分辨率模型

1 数据集介绍

本文使用的数据集分为两部分：一是由 DIV2K 和 Flickr2K 组成的 HR 图像数据集用于训练 GAN 网络，共计 3450 张，均为现实生活中的照片；部分 HR 图像如下图 2-1 所示。

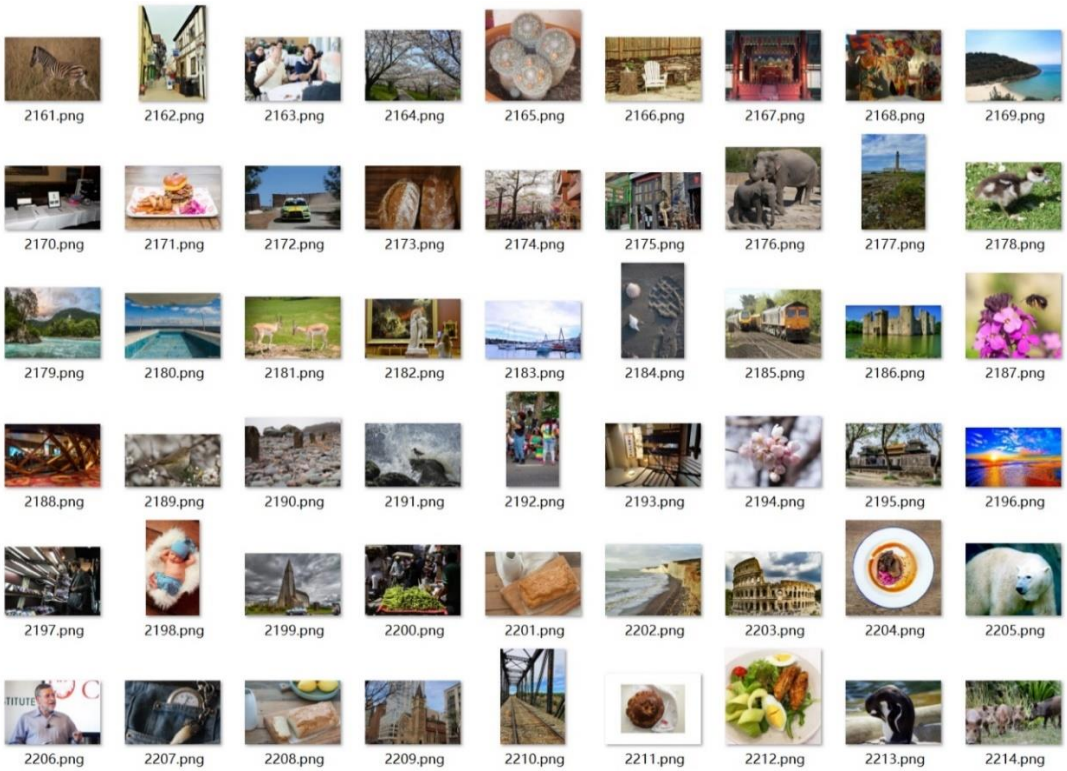


图 2-1　HR 图像训练集示例

Fig.2-1　Examples of HR training datasets

二是经过手动退化得到的已配对数据集，包括原生高清图像和退化图像各 110 张，其中 100 张用于调整 GAN 网络，10 张用于测试各类方法和模型的超分效果。其既包含现实中实拍的照片，也加入了其他内容，如 CG 图片、文档和图纸的扫描件等；因为在实际应用中也大量存在对这些内容的 SR 需求。部分 HR 和 LR 图像如下图 2-2 至图 2-4 所示。

图 2-2　已配对数据集示例 1

Fig.2-2　The first example of pair-datasets

图 2-3　已配对数据集示例 2

Fig.2-3　The second example of pair-datasets

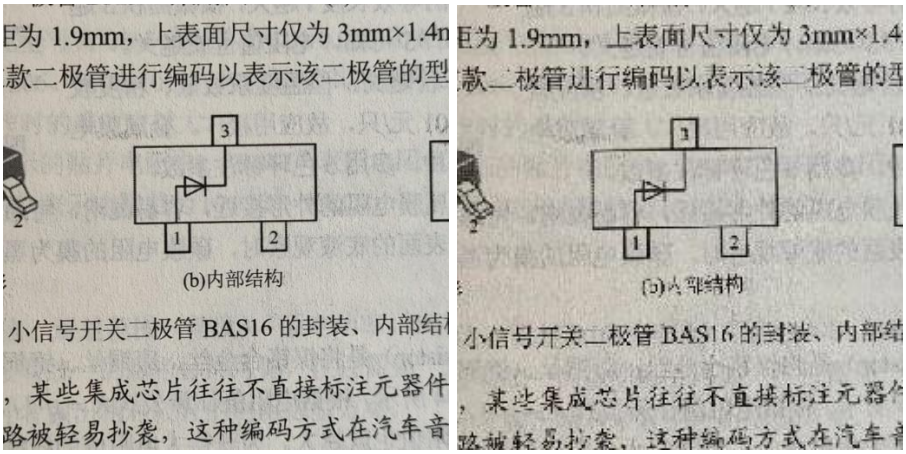


图 2-4　已配对数据集示例 3

Fig.2-4　The third example of pair-datasets

手动退化模拟了原生高清图像在各网站和软件内反复压缩得到低分图像的过程，退化流程如图 2-5 所示。首先选取分辨率大于 3000×2000 像素的原始图片下载。本文中所有方法的目标放大倍数均为 4 倍，因此 LQ 图像的长和宽应为高分图像的 1/4。故首先需要裁切（但不缩放）下载的图像，使其长和宽可以被 4 整除，将裁切后的图像作为 HR 的 GT。GT 图像需要首先经过两次压缩，再经过一次缩放（不一定是压缩）得到 LQ 数据集。本数据集采取了 B 站、网易云音乐、QQ 和微信朋友圈作为手动压缩的途径，每次随机选取一种网站或应用程序对图片进行压缩。由于经过两次压缩后的图片长和宽不一定是 GT 图像的 1/4，因此还要再进行一次缩放，使其长和宽满足要求。为了使从 LQ 图像到 GT 图像的像素点间对应关系尽可能少地被破坏，同时兼顾计算速度，最后一次缩放采取最近邻插值法。通过这种方法得到的 LQ 图像和 GT 图像即可构成已配对的数据集。

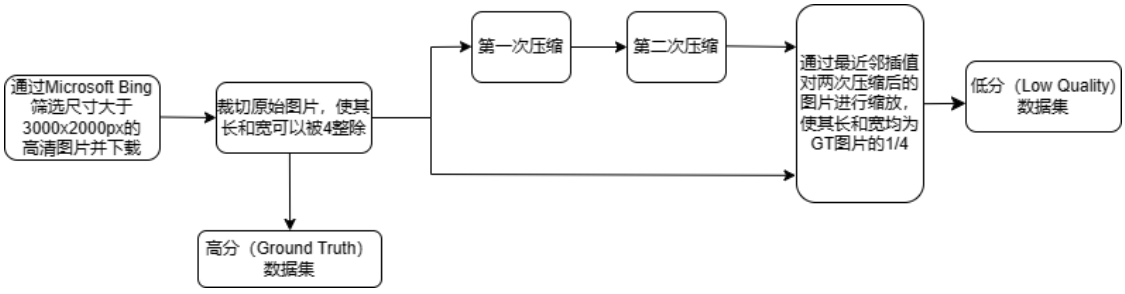


图 2-5　手动退化流程图

Fig.2-5　Manually degradation flow chart

2 生成对抗网络理论基础

GAN 是一种无监督学习的复杂分布上的生成模型，主要包含生成模型（Generator, G）和鉴别模型（Discriminator, D）。生成模型（又称生成器）的目标是生成尽可能接近真值的

结果，而鉴别模型（又称鉴别器）则用于判断生成的结果是否足够接近真值。随着训练迭代次数的增加，两个网络的性能会越来越强大，迭代完成后的 G 网络是本文最终需要的模型。GAN 的结构图如下图 2-6 所示。

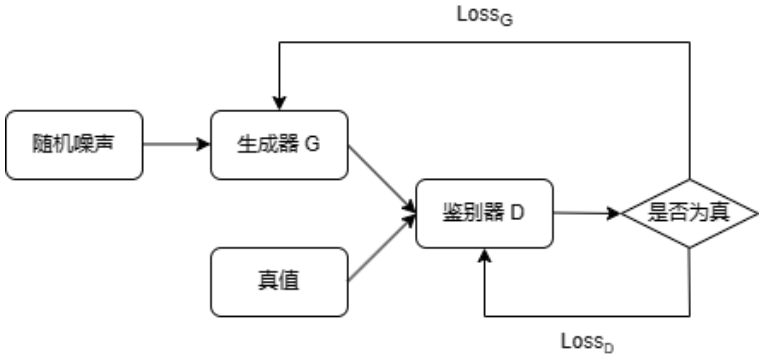


图 2-6　生成对抗网络结构图

Fig.2-6　Framework of Generative Adversial Network

在 GAN 中，随机噪声输入生成器，其输出结果越接近对应的 GT，表明其性能越强大。鉴别器将生成器的输出与其对应的 GT 作为输入，计算鉴别器认为 GT 为真的概率和认为生成器的输出为真的概率，并计算对应的损失值，返回给鉴别器自身以及生成器，从而同步强化二者的性能。在实际应用中，先选取一个随机分布作为噪声源$p_z(z)$，并准备好真值数据集，初始化生成器的参数θ_G和鉴别器的参数θ_D。在每次迭代中，$p_z(z)$产生m个噪声向量：$\{z^1, z^2, z^3, \ldots, z^m\}$，作为$G$的输入，得到输出$\{G(z^1), G(z^2), G(z^3), \ldots, G(z^m)\}$；从真值数据集$P_{data}(x)$中随机选取$m$个样本：$\{x^1, x^2, x^3, \ldots, x^m\}$，与$G(z^i)$ $(i \in [1, m])$一起作为D的输入，得到$D(x^i)$和$D(G(Z^i))$，分别表示鉴别器判定x为真的概率，和判定$D(G(Z^i))$为真的概率。G和D网络都是多层感知器，其中G为无监督网络，D为有监督网络。

GAN 网络训练的核心在于价值函数（Value Function），如式 2-1：

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log_2 D(x)] + E_{z\sim p_z(z)}[\log_2 (1 - D(G(z)))] \quad (2\text{-}1)$$

先固定生成器，优化鉴别器，考虑G为确定映射，则式 2-1 可转化为式 2-2：

$$\max_D V(D,G) = E_{x\sim p_{data}(x)}[\log_2 D(x)] + E_{z\sim p_z(z)}[\log_2 (1 - D(G(z)))] \quad (2\text{-}2)$$

要求D使$V(D,G)$最大，则$D(x)$和$1 - D(G(z))$应尽可能大。再固定鉴别器，优化生成器，考虑D为确定映射，则式 2-1 可转化为式 2-3：

$$\min_G V(D,G) = \log_2 (1 - D(G(z))) \quad (2\text{-}3)$$

要求G使$V(D,G)$最小，则$1 - D(G(z))$应尽可能小。

由于实际的网络中，生成器只接受噪声输入，没有接受真值，因此影响式 2-2 的只有$1 - D(G(z))$ [2] 。从价值函数来看，GAN 中的“对抗”表现在生成器和鉴别器对$D(G(z))$的不同目标。

3 生成模型

本文中的生成模型采用 ESRGAN[16]中残差内残差密集块（Residual-in-Residual Dense Block, RRBD）。RRBD 是由 SRGAN[15]的残差块（Residual Block, RB）改进而来。整个生成器的架构如图 2-7 所示。

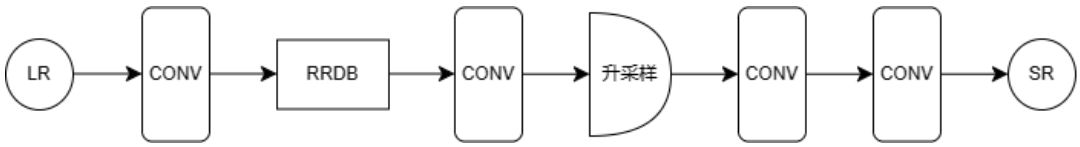


图 2-7　本文使用的生成器架构图

Fig.2-7　Architecture of the Generator

3.1 残差块

SRGAN 的生成器部分和其中使用的 Residual Block 如下图 2-8 所示。

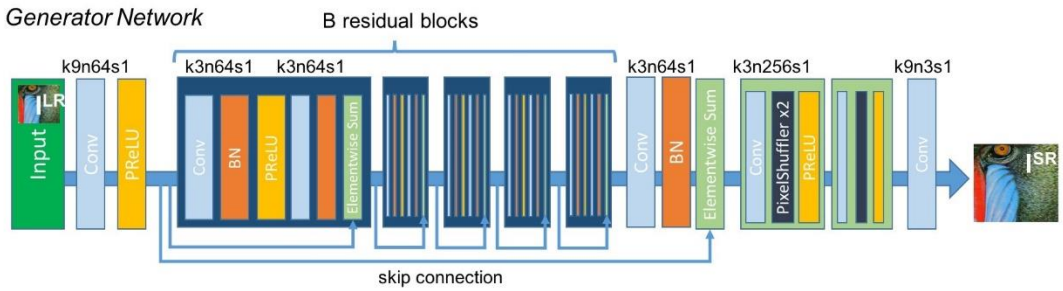


图 2-8　SRGAN 生成器架构图

Fig.2-8　Architecture of Generator Network in SRGAN

图中，Conv 表示二维卷积，对于格式为(N, C_{in}, H, W)的四维输入张量，通过一次二维卷积得到的结果如式 2-4 所示：

$$out(N_i, C_{out_j}) = bias(C_{out_j}) + \sum_{k=0}^{C_{in}-1} weight(C_{out_j}, k) \star input(N_i, k) \qquad (2\text{-}4)$$

N为批尺寸（batch size），C为通道数，H和W分别为单个二维平面的高和宽，$\star$是二维互相关运算符。对于 RGB 图像输入而言，C是颜色通道数，即为 3；H和W为图片的高和宽。

PReLU[20]是对激活函数 ReLU 的改进。当输入大于 0 时，ReLU 的输出与输入一致；当输入小于 0 时，ReLU 的输出为 0。为了改善模型的过拟合问题，PReLU 函数对为负值的输入提供了线性的输出，如式 2-5 所示：

$$f(y_i) = \begin{cases} y_i, & if\ y_i > 0 \\ a_i y_i, & if\ y_i \le 0 \end{cases} \qquad (2\text{-}5)$$

其中a_i为 PReLU 函数的负向斜率，可在学习中进行参数更新。

BN（Batch Normalization）[21]为批归一化算法。通过计算当前批（Batch）中每个特征

的均值和方差，对该批次中的数据进行归一化（Normalization），从而减轻梯度消失问题。

3.2 残差内残差密集块

与 SRGAN 的 Residual Block 相比，ESRGAN 使用的 RRDB 去掉了 RB 中的 BN 从而减少了网络参数量[22]，并用多个 Dense Block（DB）替换了原来的 RB 结构。DB 没有使用 PReLU 作为激活函数，而是使用了 LeakyReLU。其主要区别在于 LeakyReLU 具有固定的负向斜率，不再需要进行参数更新。RRBD 的结构如图 2-9 所示。

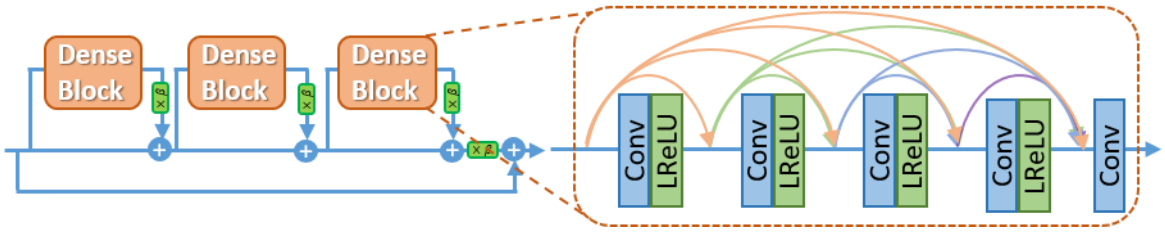


图 2-9　RRDB 结构图

Fig.2-9　Structure of RRDB in ESRGAN

RRDB 的特点在于，将多个残差密集块[23]（Residual Dense Block, RDB）进行残差级联，而每个残差块内又是多个卷积层的密集连接：把先前的 RDB 状态与后面每一个卷积层相连，从而充分利用低分辨率图像的层次特征。

4 鉴别模型

本文中的鉴别模型使用具有谱归一化（Spectral Normalization, SN）的 U-Net 结构，并支持跳跃连接（Skip Connection），其结构如下图 2-10 所示。

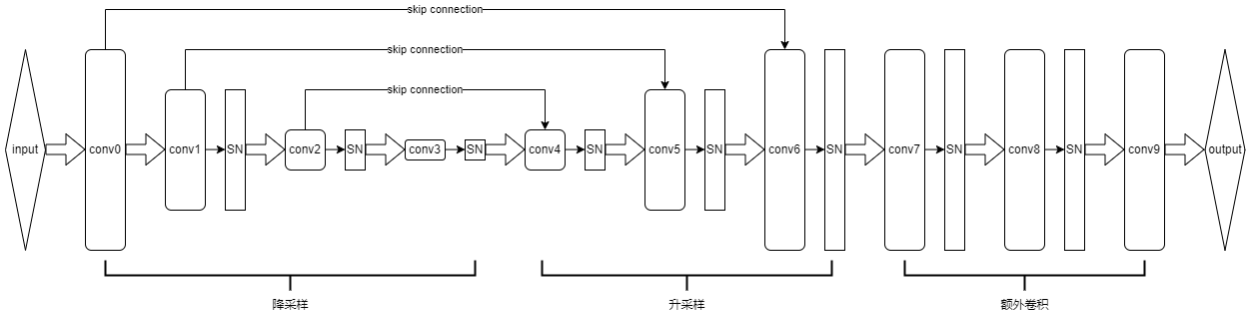


图 2-10　本文使用的鉴别器架构图

Fig.2-10　Architecture of the Discriminator

4.1 U-Net 结构

U-Net 是一种卷积神经网络结构，最初用于生物医学图像分割[24]，现在被广泛使用为 GAN 的鉴别器架构[25]。如图 2-10 所示，U-Net 的特点是先对输入进行多次降采样（又称为编码），再进行升采样（又称为解码），最终得到输入图像中每个像素对应的“真实程

度”。由于在提取特征的过程中，矩阵的尺寸先变小后变大，形状类似 U 型，故因此得名。

4.2 谱归一化

SN[26]是一种用于抑制神经网络中权重矩阵的 Lipschitz 常数的技术，能够降低梯度爆炸的可能性。所谓梯度爆炸即在反向传播更新网络参数的过程中，由于神经网络层数过多，或权重初始化值太大等原因，导致部分梯度的数值溢出，造成训练不稳定，甚至无法收敛。SN 使用幂迭代法计算的权重矩阵的谱范数σ来重新标定权重张量，对于权重矩阵 W，谱范数$\sigma(W)$的计算如式 2-6 所示：

$$\sigma(W) = \underset{h:h\neq 0}{max} \frac{||Wh||_2}{||h||_2} \tag{2-6}$$

得到谱范数后，即可计算谱归一化后的权重矩阵，如式 2-7 所示：

$$W_{SN} = \frac{W}{\sigma(W)} \tag{2-7}$$

目前 PyTorch 提供了内置的 torch.nn.utils.spectral_norm 函数，可以直接调用。

5 评估指标

评估指标在神经网络中非常重要；这不仅是因为可以通过评估指标判断模型生成结果的优劣，更是由于神经网络需要依靠损失函数调整其训练过程。如本章 4.2 节中提到的，网络参数需要通过梯度下降法在反向传播中进行更新，这对于大多数神经网络而言都是适用的，包括本文使用的 GAN。梯度下降法即通过迭代找到损失函数的最小值，如果损失函数的设置不合理，更新后的参数就无法实现既定目标。在本文中，若损失函数不能全面且真实地反映出 SR 结果与 GT 的差距，该模型的性能自然难以满足真实场景下的 SR 需求。

5.1 常用评估指标

本文共使用了六种评估指标，分别为平均绝对误差（Mean Absolute Error, MAE）、均方误差（Mean Squared Error, MSE）、峰值信噪比（Peak Signal-to-Noise Ratio, PSNR）、感知损失（Perceptual Loss）、对抗损失（GAN Loss）和多尺度结构相似性（Multi-Scale Structural Similarity, MS-SSIM）。

MAE 又称为 L1 损失函数（L1 LOSS），表达式如式 2-8 所示：

$$MAE = \frac{\sum_i^n |y_i - y_i^p|}{n} \tag{2-8}$$

其中y_i为真实值，y_i^p为预测值，n为样本数量。显然 L1 LOSS 的导数为常数函数，因此其梯度稳定，不会产生梯度爆炸的问题，同时对异常值更加鲁棒。但 L1 LOSS 在 0 点处为拐点，此时不可导。如果使用梯度下降法进行学习，可能会错过最小值。

MSE 又称为 L2 损失函数（L2 LOSS），表达式如式 2-9 所示：

$$MSE = \frac{\sum_i^n (y_i - y_i^p)^2}{n} \tag{2-9}$$

与 MAE 相比，MSE 处处可导，便于使用梯度下降求解。但由于存在平方项，MSE 更容易

受到偏离值的影响。此外，由于 MSE 的导数在距离 0 点处较远时较大，在靠近 0 点时又非常小，因此可能会产生梯度爆炸的问题；训练速度也会在接近 0 点时变得非常慢。根据 L1 LOSS 和 L2 LOSS 的定义，其值越小则说明从像素角度考虑，生成的图像越接近真值。

PSNR 基于 MSE 计算得到，定义为峰值信号的能量与噪声的平均能量之比，其表达式如式 2-10 所示：

$$PSNR = 10log_{10}\frac{MaxValue^2}{MSE} \tag{2-10}$$

其中，$MaxValue$为图像单一颜色通道的最大强度。通常情况下，个人计算机中的绝大部分图片为 8 位色深，即每个颜色通道有$2^8 = 256$种强度，最大值为$256 - 1 = 255$。PSNR 越大，意味着生成图像中噪音的影响越小，因此图片观感更佳。

与 MAE 和 MSE 相比，感知损失[27]更注重图像的感知质量，符合人眼对图像的视觉感知。感知损失的计算在形式上与 MSE 基本一致，区别在于从图像空间计算转变为了特征空间计算，公式如式 2-11 所示：

$$L_{\text{perceptual}} = \sum_{i=1}^{N}\frac{1}{C_iH_iW_i}\sum_{j=1}^{C_i}\sum_{k=1}^{H_i}\sum_{l=1}^{W_i}(F_{ijkl} - P_{ijkl})^2 \tag{2-11}$$

其中N为特征图的数量；C_i、H_i和W_i为第i个特征图的通道数、高度和宽度；F_{ijkl}和P_{ijkl}分别为生成图像和真实图像在第i个特征图上位置为 (j,k,l)的像素值。类似地，感知损失越小说明图像越符合人眼的直觉和感观。

对抗损失实际上是本章第 2 节中介绍的价值函数，即式 2-1，此处不再赘述。

MS-SSIM 由结构相似性（Structural Similarity, SSIM）改进而来。SSIM 是一种用于比较两幅图像之间结构相似性的指标，考虑了亮度、对比度和结构三个方面的差异；而 MS-SSIM 则先对图像进行多次降采样，在不同尺度上计算 SSIM，最后进行加权平均，因此与 SSIM 相比能更好地捕捉图像的全局和局部结构信息。其计算方式如下式 2-12 所示：

$$\text{MS-SSIM}(X,Y) = [l_M(X,Y)]^{\alpha_M}\cdot\prod_{j=1}^{M}[c_j(X,Y)\cdot s_j(X,Y)]^{\alpha_j} \tag{2-12}$$

其中，$l_M(X,Y)$表示亮度相似性，$c_j(X,Y)$表示第j个尺度下的对比度相似性，$s_j(X,Y)$表示第j个尺度下的结构相似性，α_M表示亮度相似性的权重，α_j表示第j个尺度下的权重。SSIM 和 MS-SSIM 越接近 1，则生成的图像从结构上越接近真值。

5.2 损失函数

在 GAN 的训练中，本文综合式 2-1 的对抗损失、式 2-8 的 L1 损失和式 2-11 的感知损失作为整个网络的损失函数。若三者的权重分别为W_{GAN}、W_{L1}和$W_{perceptual}$，则综合后的损失函数如式 2-13 所示：

$$L_{training} = W_{GAN}\times L_{GAN} + W_{L1}\times L1 + W_{perceptual}\times L_{perceptual} \tag{2-13}$$

其中L_{GAN}表示 G 网络或 D 网络的 GAN 损失。综合后的损失函数既能维持 L1 损失的稳定性，又能针对人眼感知进行结果优化，使得模型在真实场景中能取得较好的结果。

第三章　实验过程与结果分析

1 模型训练

本文中模型训练的最终目标为优化后的 GAN 模型，称为“真实场景优化超分辨率生成对抗网络”（Optimized for Real-World Super Resolution GAN, ORSRGAN）。为了区分加入人工退化数据加强训练前后的 GAN 模型，本文将直接得到的 GAN 模型记作“ORSRGAN”，将加强训练后的 GAN 模型记作“finetune-ORSRGAN”

1.1 模型预训练

在训练 ORSRGAN 模型之前，首先要准备预训练（Pre-trained）模型。预训练模型的网络结构与 GAN 模型基本一致，唯一的不同之处在于预训练模型缺少判别器，只有生成网络，是一种 Real-ESRNet[8]模型，本文中记作“ORSRNet”。因此，预训练网络不加入对抗损失和感知损失作为损失函数的一部分，而是只使用 L1 损失函数。缺少对抗损失虽然会造成超分性能的下降，但也不会生成 GAN 模型特有的伪影，故可利用这种特性检测 GAN 模型超分后产生的伪影[19]。当 L1 损失近似收敛时，L2 损失和峰值信噪比也不会很大。因此 Real-ESRNet 可被看作是以峰值信噪比为导向的网络结构。

ORSRNet 同样也需要预训练模型，本文使用放大倍数为 4 的 ESRGAN[16]模型的 G 网络作为 ORSRNet 的预训练模型；数据集使用 DIV2K 和 Flickr2K 组成的高分图像数据集，G 网络结构为 RRDB，优化器为 Adam 型，学习率为 10^{-4}，迭代 100000 次。训练平台使用英伟达 RTX3060 12G 显卡，大约用时 5 天完成。

1.2 正式训练

得到 ORSRNet 后，将其作为 ORSRGAN 的预训练模型，仍使用相同的数据集进行训练。损失函数中 L1 损失和感知损失的权重为 1，对抗损失的权重为 0.1。感知损失的实际计算依赖于预训练的 VGG19 网络作为特征提取器；VGG19 也是一种卷积神经网络，因包含 16 个卷积层和 3 个全连接层共计 19 个隐藏层而得名。在将 GT 图片送入网络之前，可以选用非锐化掩蔽（Unsharp Masking, USM）算法对 GT 值进行锐化，超分结果的观感也会更好， 不过这种做法也会在颜色突变处引入更多的伪影。本文的目标是尽可能追求更高的还原度，因此没有采用这种做法。设定 G 网络和 D 网络的优化器为 Adam 型，学习率为 10^{-4} ，迭代 20000 次，在相同的硬件平台上大约耗时 8 天完成训练。正式训练得到的 G 网络即为 ORSRGAN 模型。

1.3 加强训练

得到 ORSRGAN 后，使用手动退化的配对数据集对其进行微调（finetune）。与正式训练过程相比，进行微调时不需要加入退化模型；因为已经有了成对的 GT 值和 LQ 值，只

需要将事先准备好的 LQ 图像输入生成器即可送入鉴别器与 GT 图像进行对比。在其他的网络结构与设置上二者完全一致。由于训练数据集的样本不够丰富，为了避免出现过拟合的现象，导致微调后的网络在实际应用中效果反而更差，故只进行 10000 次迭代，在相同的硬件平台上大约耗时 4 天半。微调后得到的 G 网络为 finetune-ORSRGAN 模型。

2 模型性能对比

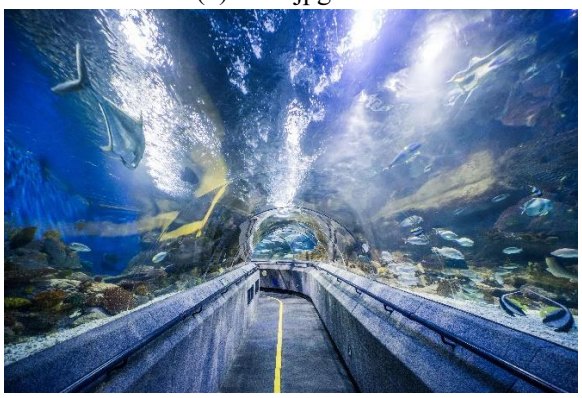
2.1 测试数据集

测试数据集中的 GT 图像如下图 3-1 所示。

(a) test1.jpg

(b) test2.jpg

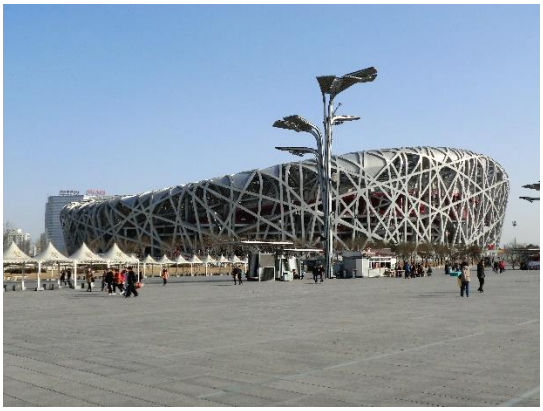
(c) test3.jpg



(d) test4.jpg

(e) test5.jpg

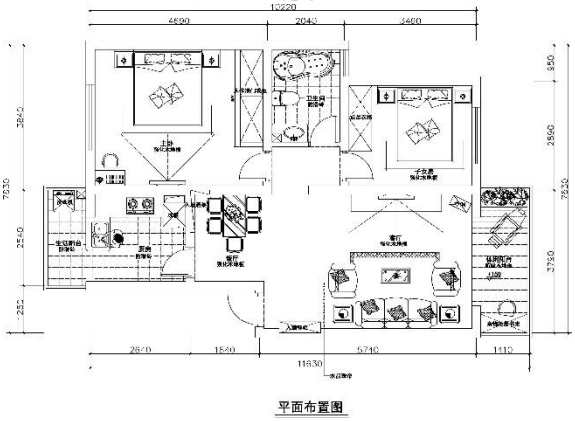
(f) test6.jpg



(g) test7.jpg

(h) test8.jpg

(i) test9.jpg



(j) test10.jpg

图 3-1　测试数据集

Fig.3-1　Datasets for testing

本文共选取了十组图片作为测试数据集，这些图片中都包含有大量的纹理细节，便于区分各模型的性能。其中前九张图片均为真实场景下的照片，包括室内与室外、动物与人物、人工建筑和自然景色，模拟了现实生活中多数照片的拍摄情况；最后一张图片为图纸，因为对图纸和文档进行 SR 也是现实生活中的常见需求。例如坐在教室后排的同学想要拍摄黑板上的板书，但由于手机摄像头焦距较短，只能进行数码放大，无法捕捉到足够清晰的图像；抑或是使用传感器尺寸较大的相机拍摄桌面上的纸质文件，由于广角镜头存在边缘畸变，加之厂商为了提高在极暗环境下的拍摄能力，倾向于使用大光圈，导致近距离对焦时边缘画质急剧下降，拍摄的纸质文件也会变得模糊。在加强训练时，本文也加入了文档扫描件和具有复杂细节的地图作为训练集，从而提高 ORSRGAN 对文字内容的 SR 能力。

2.2 超分辨率重建结果与分析

本文分别对测试数据集进行了最近邻插值、双线性插值、双三次插值（取 a 为-1）、ORSRGAN 超分和 finetune-ORSRGAN 超分，统计每张图片使用每种超分方法后的 L1 损失、L2 损失、峰值信噪比和多尺度结构相似性。统计结果如下表 3-1 所示。

表 3-1　各超分方法性能测试结果

Tab. 3-1　Benchmark results of 5 SR methods

图片序号	测试指标	NN	Bilinear	Bicubic	ORSRGAN	Finetune
1	L1 LOSS	12.9271	14.2201	8.8578	15.2154	14.8429
	L2 LOSS	424.2164	491.8552	179.2957	333.4836	283.674
	PSNR	21.8549	21.2124	25.5951	20.2428	21.2017
	MS-SSIM	0.9027	0.852	0.9352	0.8614	0.8952
2	L1 LOSS	11.0691	10.608	10.1208	10.9117	10.1587
	L2 LOSS	342.9072	288.4034	277.8202	333.4836	283.674
	PSNR	22.779	23.5307	23.6931	22.9	23.6026
	MS-SSIM	0.8184	0.7909	0.8264	0.7964	0.8187
3	L1 LOSS	11.5634	12.6883	10.6213	13.6826	12.9275
	L2 LOSS	529.7047	580.7437	427.5051	711.5721	597.8763
	PSNR	20.8904	20.4909	21.8213	19.6086	20.3646
	MS-SSIM	0.8654	0.8145	0.8706	0.8237	0.8493
4	L1 LOSS	11.5384	10.8216	10.7423	10.6495	10.2409
	L2 LOSS	325.4752	280.8297	271.3284	281.2509	245.5368
	PSNR	23.0056	23.6463	23.7958	23.6398	24.2296
	MS-SSIM	0.7966	0.7791	0.7878	0.8066	0.8183
5	L1 LOSS	17.2405	18.0636	15.1809	18.1017	17.0739
	L2 LOSS	970.0914	955.2209	665.4183	1075.6605	867.4456
	PSNR	18.2626	18.3297	19.8998	17.814	18.7483
	MS-SSIM	0.7781	0.7293	0.7879	0.7683	0.7817
6	L1 LOSS	5.937	6.226	5.2323	5.9539	5.7723
	L2 LOSS	169.0105	142.9919	131.4214	157.6335	115.613
	PSNR	25.8516	26.5776	26.9441	26.1543	27.5007
	MS-SSIM	0.934	0.9198	0.9432	0.9367	0.9391
7	L1 LOSS	7.3993	7.7829	7.221	7.8533	8.0308
	L2 LOSS	356.625	359.6731	301.8113	398.8802	358.2976
	PSNR	22.6086	22.5717	23.3334	22.1223	22.5883
	MS-SSIM	0.8936	0.8714	0.8907	0.8888	0.9023
8	L1 LOSS	4.7827	4.5179	4.6182	4.8953	4.6891

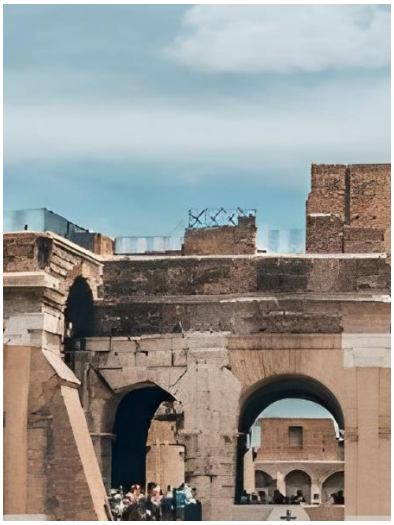
9	L2 LOSS	93.5483	67.3046	84.7818	89.8471	55.3316
	PSNR	28.4204	29.8503	28.8477	28.5957	30.701
	MS-SSIM	0.8997	0.898	0.8961	0.9096	0.9092
	L1 LOSS	20.3392	20.886	20.6364	23.5392	22.484
	L2 LOSS	1154.7783	1031.6289	1086.7415	1492.7288	1297.41
	PSNR	17.5058	17.9955	17.7695	16.3909	17
	MS-SSIM	0.6193	0.567	0.6054	0.5555	0.5999
10	L1 LOSS	13.6431	17.8806	14.9841	15.265	16.7883
	L2 LOSS	3042.0488	3006.5151	2571.9878	3242.6091	3686.627
	PSNR	13.2991	13.3501	14.0281	13.0218	12.4645
	MS-SSIM	0.7908	0.7669	0.7922	0.7687	0.7949

从各性能指标来看，finetune-ORSRGAN 与双三次插值的理论效果基本持平；最近邻插值和未进行强化训练的 ORSRGAN 引入了最多的噪声；双线性插值对原图像结构的破坏最明显。微调前的 ORSRGAN 在性能指标上的表现明显弱于双三次插值，与训练使用的 HR 数据集以及迭代次数有关。DIV2K 和 Flickr2K 组成的 HR 训练数据集仅包含照片，并且多为风景照和人物照，同时绝大多数照片都是在明亮环境下拍摄的。而测试使用的真实退化数据集图片类型更加丰富，例如大量的文本内容，或复杂光源环境下的照片；因此训练数据集未能充分考虑所有在真实场景中可能存在的 SR 需求。此外，由于硬件平台限制，神经网络每批处理的图片数量不能太多，否则会爆显存；而为了能尽快地得到模型，迭代次数也无法设置的更高，因此训练的周期数受限，影响了模型对训练集的拟合程度。而加入了 100 组配对数据进行优化的 finetune-ORSRGAN 由于接收了类型更丰富、与测试集也更贴近的训练集，与优化前相比各项性能指标得到了大幅提升。图 3-2 至 3-5 分别展示了部分测试图片得到的超分结果。

(a) SR result with NN

(b) SR result with Bilinear

(c) SR result with Bicubic



(d) SR result with ORSRGAN

(e) SR result with Finetune

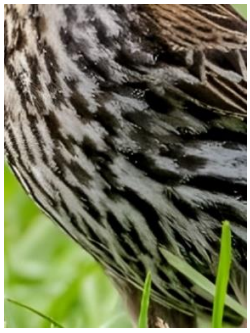
(f) GT

图 3-2　五种超分方法在 3 号测试图片上的结果与真值的对比

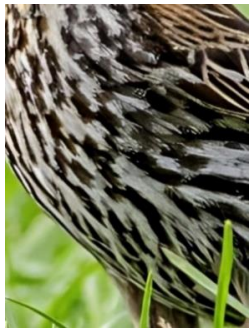
Fig.3-2　Comparison of GT and SR results of 5 methods in “test3.jpg”

(a) LQ

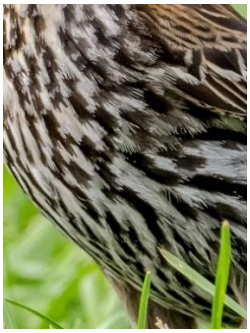
(b) Bicubic



(c) ORSRGAN



(d) Finetune



(e) GT

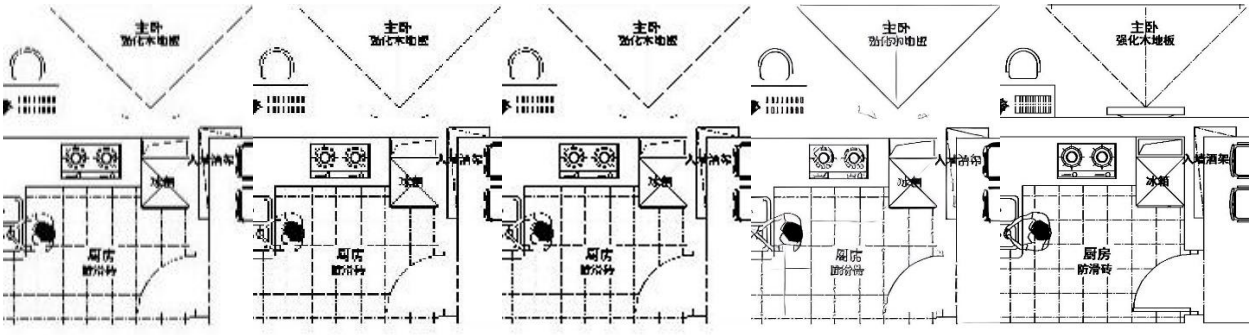
图 3-3　双三次插值与两种 GAN 模型在 6 号测试图片上的结果与低分图片和真值的对比

Fig.3-3　Comparison of LQ, GT and SR results of bicubic interpolation and 2 GAN models in “test6.jpg”

(a) Bicubic (b) Finetune (c) GT

图 3-4 双三次插值与 finetune-ORSRGAN 模型在 8 号测试图片上的结果与低分图片和真值的对比

Fig.3-4 Comparison of GT and SR results of bicubic interpolation and finetune-ORSRGAN in “test8.jpg”



(a) LQ (b) NN (c) Bicubic (d) Finetune (e) GT

图 3-5 最近邻插值、双三次插值与两种 GAN 模型在 10 号测试图片上的结果与真值的对比

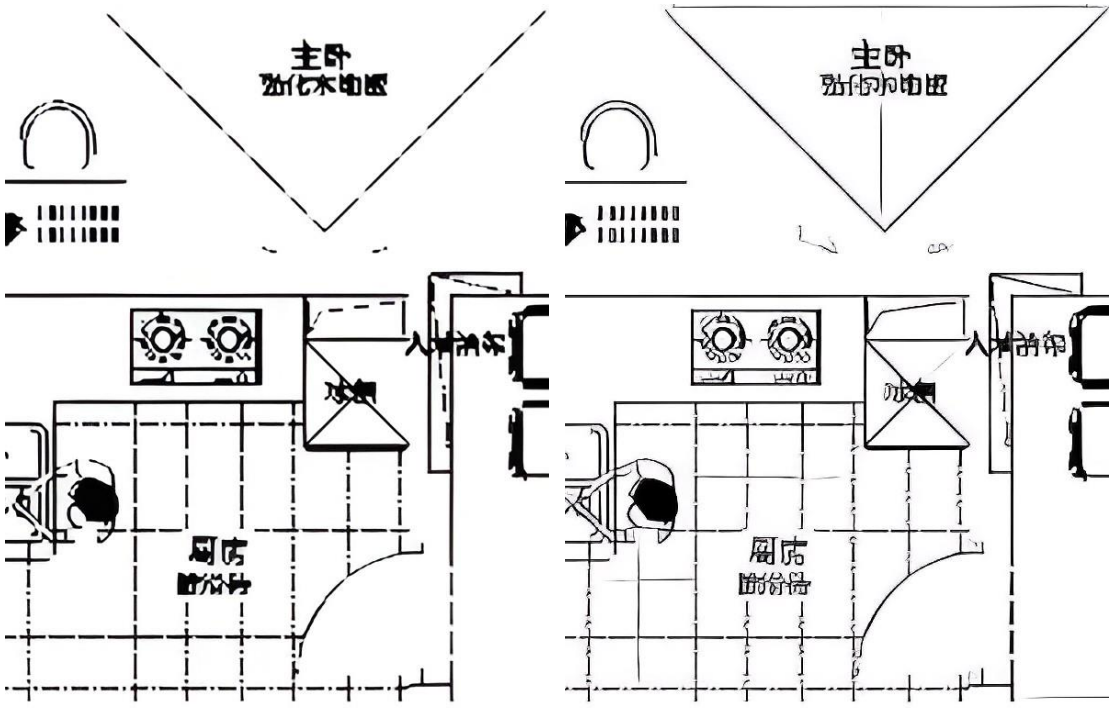
Fig.3-5 Comparison of LQ, GT and SR results of NN, bicubic and 2 GAN methods in “test10.jpg”

在图 3-2 对应的测试中，双三次插值表现出了最优秀的理论性能，而未经过优化的 ORSRGAN 理论性能则最弱；但从实际观感的角度出发，双三次插值的结果并没有完全胜过 ORSRGAN。后者与真值相比丢失了部分颜色信息，画面显得更加寡淡，物体的立体感也有所下降，导致各项评判指标偏低，但却成功消除了栏杆处的图像噪声；而双三次插值在栏杆和天空的分界处则有着明显的噪声。对于从来没见过真值图片的人而言，画面色彩更淡的问题显然没有图像噪声带来的负面影响更大；因此仅使用当前的性能指标来评价超分效果好坏仍是不够全面的。并且，双三次插值的计算量极大，耗时极长；与神经网络不同，神经网络只有在训练时需要消耗大量时间，推理速度要快得多；而双三次插值对每一张图片进行处理都需要经过漫长的等待，在实际应用中受限更多。此外，在不同的应用场景下 SR 任务的目标也不同。对于遥感图像而言，图像的准确性尤其是结构相似度比其他

参数更为重要；而在对漫画图片进行 SR 时图像观感则更加重要。因此实际选用哪种 SR 方法需要结合任务目标考虑。

类似的问题也体现在图 3-3 中，在这轮测试中双三次插值具有更小的 L1 损失和更优秀的 MS-SSIM；而 finetune-ORSRGAN 具有更小的 L2 损失和更高的峰值信噪比，理论上二者的效果应该旗鼓相当。但考虑到真值图像中的羽毛非常顺滑，而双三次插值生成的羽毛带有更多的白色噪点，没有 finetune-ORSRGAN 平滑，因此实际观感上后者的结果明显更加接近真值。若优先追求图像观感，则应当将 PSNR 作为更重要的参考指标。

当然，过度平滑也意味着信息的丢失。在图 3-4 对应的测试中，不论是理论数据还是实际观感，finetune-ORSRGAN 都要远胜于双三次插值。双三次插值的 SR 结果中包含大量的人工痕迹，而 finetune-ORSRGAN 的结果则要平滑和自然得多。但与真值图片相比，GAN 模型生成的面部丢失了大量的皱纹信息，使得图片上的人物显得年轻得多。这种“涂抹感”的问题在处理文字信息时会更加严重。在图 3-5 中，GAN 网络生成的文字被完全扭曲了，导致可读性甚至劣于三种插值算法，但 finetune-ORSRGAN 却仍具有最好的多尺度结构相似性。上文中提到了，在加强训练时，本文加入了专门的文本图像进行训练，这种做法的确起到了一定的效果，如图 3-6 所示。



(a) ORSRGAN　　(b) Finetune

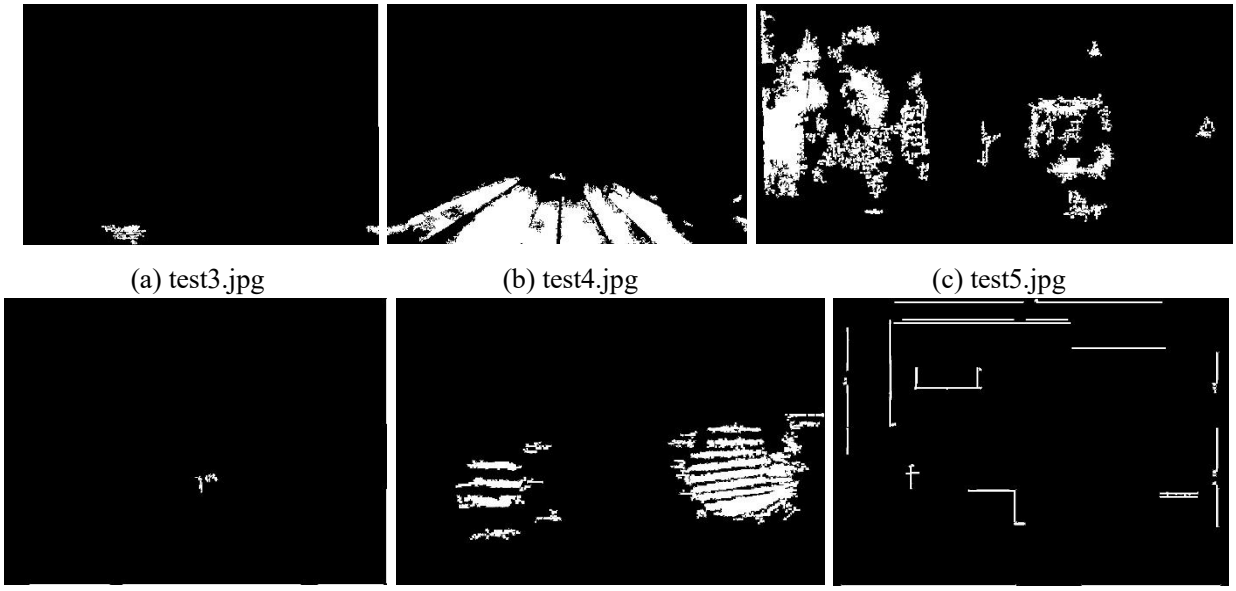
图 3-6　优化前后的 GAN 模型在 10 号测试图片上的结果对比

Fig.3-6　Comparison of SR results of 2 GAN methods in “test10.jpg”

在白底黑字的场景中，有损压缩后的文字笔划边缘变得模糊，图像原有的纹理被破坏，导致可读性下降。从观感的角度来看，这种信息损失如同水滴在了纸上，让笔墨发散。强化学习后的网络会尝试让笔划重新回到“发散”前的状态，其结果是文字笔划变得更细。这种做法有一定的合理性，但却破坏了笔划之间的关系，扭曲了原有的文字结构，不符合人类阅读时的原理，使得超分后的文本几乎完全不可读。同时，图 3-6 中的点划线和虚线也被进行了类似的处理，圆点与短线之间产生了原图中不存在的噪声，原图中的虚线也变成了实线。这类问题是由于神经网络将这些原本就存在的结构误判成了因退化而形成的模糊字体所导致的。现有的 GAN-SR 网络由于结构上存在限制，无法区分单幅图像中哪些是文本内容而哪些是常规内容。因此在处理这类图片时，需要有新的手段进行针对性的文本优化。

3 伪影检测

伪影是超分辨率技术中面临的主要困境之一。由于低分图像中的信息缺失难以原样复原，且各超分技术自身也存在局限性，生成的图像往往会存在各种瑕疵。传统插值算法中的伪影通常是由于下冲现象（Undershoot）导致细节丢失，表现为画面锐度不足；而 GAN 模型中的伪影则通常由过冲现象引发，因过度锐化图像导致产生了真值图像中原本没有的细节，甚至覆盖掉了本应还原的细节；这与 GAN 架构中 G 网络和 D 网络之间的竞争有关。如果在训练时只保留原 GAN 架构中 G 网络，去掉 D 网络，此时的网络不会产生 GAN 特有的伪影。利用这种手段，可以把只有 G 网络的模型与 GAN 模型进行对比，检测 GAN 模型产生的伪影。超分辨率伪影检测与删除技术（Detect and Delete Super Resolution Artifacts, DeSRA）[19]就采用了这个原理。在 ORSRGAN 的训练中，ORSRNet 与 ORSRGAN 相比就只缺少了 D 网络，因此可以利用 ORSRNet 的 SR 结果，通过 DeSRA 对 finetune-ORSRGAN 生成的超分结果进行伪影检测。最终，测试数据集中的 3 号、4 号、5 号、7 号、9 号与 10 号图片检测出了伪影，结果如图 3-7 所示。从检测结果中可以看出，伪影较为严重的图片为 4 号、5 号和 9 号。在 4 号图片中，水族馆的两侧墙壁和地面上存在大量因光线反射而产生的细节； 5 号图片中电路板上的芯片和底板颜色十分接近，神经网络可能将二者看作同一种颜色进行处理；9 号图片中每块砖头之间的纵向条纹颜色较浅，容易被当作噪声消除掉。由此可见，当原图像中存在大量复杂的纹理细节，且这类细节在退化后与噪声十分接近时，在 SR 的过程中 GAN 模型可能会直接丢失这些细节，或是虚构出与原图不符的细节，从而产生伪影。



(a) test3.jpg (b) test4.jpg (c) test5.jpg

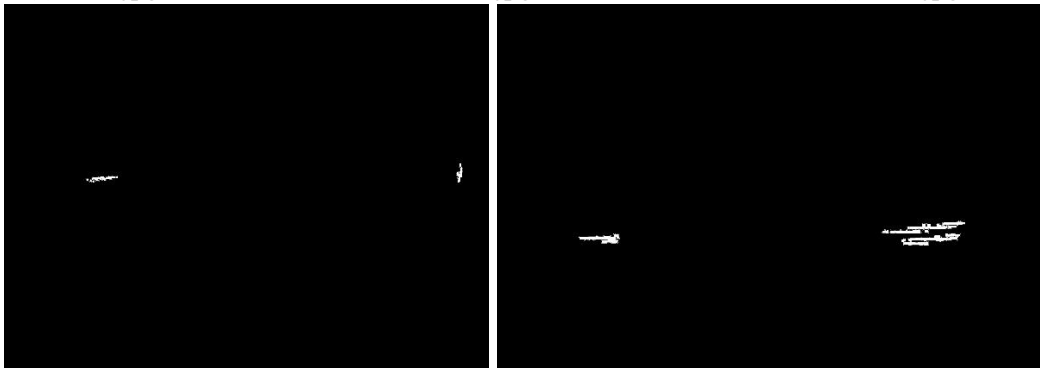
(d) test7.jpg (e) test9.jpg (f) test10.jpg

图 3-7 finetune-ORSRGAN 伪影检测结果

Fig.3-7 Artifact map of finetune-ORSRGAN

DeSRA 也提供了针对伪影问题优化后的超分模型 RealESRGAN-DeSRA。由于该模型仅进行了 5000 次迭代，评估指标表现不佳，但却具有强大的抗伪影性能。对 RealESRGAN-DeSRA 在相同测试集上的 SR 结果进行伪影检测，只有 3 号、4 号、5 号、7 号和 9 号检测出了非常轻微的伪影，结果如图 3-8 所示。可以观察到，与 finetune-ORSRGAN 相比，RealESRGAN-DeSRA 的 SR 结果中已经消除了多数的 GAN 伪影。以 9 号图片为例，如下图 3-9 所示，finetune-ORSRGAN 生成的图片抹平了砖块与砖块之间的纵向条纹间隙，因而有大量的 GAN 伪影被检测出，而 DeSRA 则更多地保留了这种细节；虽然与 GT 值仍有一定差距，但在 SR 任务中这类图片已经是较为苛刻的类型了。当然，对于 ORSRGAN 而言，其抗伪影性能仍有待提高。在第一次训练使用合成数据的模型时，可以尝试进一步改进退化模型，合成包含有类似图 3-9 中的缺陷的 LR 数据；在进行加强训练时，也可以向数据集中加入更多类似的图片进行专项训练。实际上，仅用人眼观察 9 号图片的 LQ 图像时，很容易就能辨别出砖块之间存在的规律性条纹间隙，因为噪声很难呈现出如此高度统一的形态。但在 GAN 网络中，这些条纹是否是噪声只能通过计算损失值得到。由于纹理细节中同样可能包含噪声，这些细节可能会被当成噪声一并抹去，这是 GAN 模型造成 SR 结果失真的根本原因之一。同样地，只有让网络能够更好地理解图像的构成，才能使 SR 结果在结构上更加还原 GT 值。

(a) test3.jpg　　(b) test4.jpg　　(c) test5.jpg



(d) test7.jpg　　(e) test9.jpg

图 3-8　RealESRGAN-DeSRA 伪影检测结果

Fig.3-8　Artifact map of RealESRGAN-DeSRA

(a) Finetune　　(b) DeSRA　　(c) GT

图 3-9　finetune-ORSRGAN 和 RealESRGAN-DeSRA 在 9 号测试图片上的结果与真值对比

Fig.3-9　Comparison of GT and SR results of finetune-ORSRGAN and RealESRGAN-DESRA in “test9.jpg”

第四章 结论与展望

1 研究结论

本文首先针对基于 ML 的 SR 方法严重依赖配对数据集的问题，提出了基于高阶退化模型由 GT 数据合成 LR 数据参与训练的方法，并对该方法得到的 ORSRGAN 模型进行性能评估；再针对该模型在部分性能指标上存在的缺陷，构建真实退化工作流程，引入真实退化数据，进行加强训练，得到 finetune-ORSRGAN 模型；最后对微调后的模型进行伪影检测。论文得到的主要研究结论如下：

(1) 使用纯合成 LR 数据构建面向真实场景的 LQ 图像 SR 网络，受限于 HR 数据集的丰富性以及退化模型的代表性不足的问题。虽然得到的图像更加平滑，但各项性能指标不尽如人意，不能满足真实场景下对图像保真度的高要求。

(2) 经过真实退化数据强化后的模型，其所有常用性能指标都得到了显著的提升，生成的 SR 结果更贴近 GT 值，证明了真实退化数据对于训练面向真实场景的 SR 网络具有重要价值。我们仍然需要开发一种快速得到真实退化数据的手段。

(3) 上述方法得到的网络仍然具有一定的局限性，在处理文本内容以及复杂纹理时可能会失效。有时图像中的部分纹理结构已经被破坏了，却不会通过现有的评价指标反映出来。这些问题都需要在未来提出更进一步的解决方案。

2 研究展望

综合前文所述，目前 GAN 超分辨率模型最亟待解决的问题是退化流程的问题，其次是文字结构破坏的问题。

虽然 Real-ESRGAN 提出了高阶退化，但实际操作中仍然是二阶退化；继续增加退化流程，或者对每个退化流程中的各项进行随机排列，在模型性能提升与训练难度提升之间能否达到平衡，目前仍是未知数，需要进行更多的研究。合成 LR 数据的思路为真实退化的配对数据难以获取的问题提供了另辟蹊径的解决方案，但这种手段依然受限于训练数据的规模。若无论如何都需要依赖强大的数据集去训练效果强大的网络，不如尝试构建自动化的数据采集工作流程，包括自动化的真实退化流程。因为现有的研究结果表明配对数据的重要性是不容忽视的。即便使用纯合成 LR 数据，也需要有更加广泛的训练数据。例如目前常用的 DIV2K 和 Flickr2K 数据集就完全没有涉及到文本内容，也很少有夜晚暗光环境下的照片，或是光照环境复杂的图片。如果不能正确地处理光线关系，超分后的图片很容易出现局部过曝的情况。因此不论采取何种技术路线，数据集的规模都很重要。

虽然目前已有专门用于文本超分的 TSRN（Scene Text Image Super-Resolution Recurrent Network）[28]等模型，但这些工作往往比较分散，与 GAN-SR 网络结合存在一定的障碍。

GAN-SR 模型在文本处理中的异常既源于上文所述的缺少相关训练样本，也体现在神经网络无法真正理解和分辨什么是文字。即使文字的结构已经在 SR 的过程中被扭曲了，在计算 MS-SSIM 时也不会出现异常。文字有着严格的结构规范，而这种结构在真实退化场景中很容易被破坏。人眼能够勉强分辨退化后的文本，很大程度上源于人脑中存在对文字的先验知识。例如，给一个从来没有见过中文的人看 LR 的中文，他也无法进行辨别；而现有的 GAN-SR 模型就存在这种问题，其在真实场景中的应用必然是受限的。此外，对于 LR 的纸质文档照片等内容，显然存在比直接将整张图片扔进神经网络进行 SR 更好的办法。对纸质文档而言，空白部分是不重要的，只有文字本身是重要的。即使把空白部分全部涂白，也不会影响人对剩余内容的阅读。如果直接对整张图片进行超分，由于照片中不可避免存在噪声和色块，部分噪声可能会被放大，混入超分后的图片中。实际上，如果有性能足够的语义分割模型，把文字内容单独分割出来进行处理，余下部分直接赋值为白色，效果都会好得多。目前在商用领域中，部分智能手机对纸质文件进行拍摄时，会启用所谓的文档模式，其基本思路与之相同：着重增强文字内容，其他区域则可以直接抹平。如果尝试将特定内容的语义分割引入通用超分网络，能弥补通用模型因缺乏对特定内容的先验知识而破坏其原有结构的问题，实现模型性能的进一步提升。

参考文献

[1] bilibili AI Lab. Real Cascade U-Nets for Anime Image Super Resolution [OL]. GitHub: bilibili,2022. GitHub.

[2] 张宁. 基于深度学习的无监督遥感图像超分辨率重构技术研究[D].中国科学院大学(中国科学院长春光学精密机械与物理研究所),2023.

[3] 邓聪,徐健.抑制人工痕迹全变分正则化的图像超分辨率[J].中国新通信,2020,22(14):94-97.

[4] 林莉,唐昌华,王岩,等.基于改进机器学习的超分辨率图像细节复原[J].计算机仿真,2024,41(04):210-213+288.

[5] Yang J, Wright J, Huang T S, et al. Image super-resolution via sparse representation[J]. IEEE transactions on image processing, 2010, 19(11): 2861-2873.

[6] Elad M, Feuer A. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images[J]. IEEE transactions on image processing, 1997, 6(12): 1646-1658.

[7] Shin R, Song D. Jpeg-resistant adversarial images[C]//NIPS 2017 workshop on machine learning and computer security. 2017, 1: 8.

[8] Wang X, Xie L, Dong C, et al. Real-esrgan: Training real-world blind super-resolution with pure synthetic data[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 1905-1914.

[9] Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 295-307.

[10] Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1874-1883.

[11] Zhang Y, Li K, Li K, et al. Image super-resolution using very deep residual channel attention networks[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 286-301.

[12] Wang Z, Gao G, Li J, et al. Lightweight image super-resolution with multi-scale feature interaction network[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1-6.

[13] Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017: 136-144.

[14] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[15] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[16] Wang X, Yu K, Wu S, et al. Esrgan: Enhanced super-resolution generative adversarial networks[C]//Proceedings of the European conference on computer vision (ECCV) workshops. 2018: 0-0.

[17] Zhang K, Liang J, Van Gool L, et al. Designing a practical degradation model for deep blind image super-resolution[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 4791-4800.

[18] 郭小萌. 基于 GAN 的人脸盲超分辨率的研究与实现[D].武汉邮电科学研究院,2023.

[19] Xie L, Wang X, Chen X, et al. Desra: detect and delete the artifacts of gan-based real-world super-resolution models[J]. arXiv preprint arXiv:2307.02457, 2023.

[20] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[21] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International conference on machine learning. pmlr, 2015: 448-456.

[22] 张鹏婴,张明,李建军,等.基于轻量化生成对抗网络的遥感图像超分辨率重建[J].激光杂志,2024,45(04):114-120.

[23] Zhang Y, Tian Y, Kong Y, et al. Residual dense network for image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2472-2481.

[24] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015: 234-241.

[25] Schonfeld E, Schiele B, Khoreva A. A u-net based discriminator for generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8207-8216.

[26] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks[J]. arXiv preprint arXiv:1802.05957, 2018.

[27] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 694-711.

[28] Wang W, Xie E, Liu X, et al. Scene text image super-resolution in the wild[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. Springer International Publishing, 2020: 650-666.

致　谢

在本文的结尾，我要向过去四年间所有支持着我的人表达最诚挚的感谢。

首先我要感谢王洁和李玉花老师。她们直接指导了本文的创作，提供专业建议，奠定了本文的基础。

同时，我也要感谢我的所有任课老师们；没有他们的教导，我也无法具备足够的专业素养完成这篇文章和整个毕业设计。

在我的人生遭遇迄今为止最为艰难的处境之时，人工智能学院和校方也给予了最大程度的帮助。因此我也要感谢南京农业大学，感谢我们的人工智能学院，以及我们学院的辅导员们。当然，我也不能忘了所有支持和帮助过我的同学们。

最后，我要感谢腾讯 Arc 实验室的研究员 Xintao Wang 和其他 GitHub 社区的贡献者们，感谢他们为开源社区环境添砖加瓦。