# 生成对抗网络

## 生成对抗网络

生成对抗网络（Generative Adversarial Network，简称GAN）是非监督式学习的一种深度学习方法，通过让两个神经网络相互博弈的方式进行学习。该方法最初由 lan·Goodfellow 等人于2014年提出[1]，目前，GAN的变种更是有上千种，2019年计算机界的诺贝尔奖“图灵奖”得主，深度学习先驱之一的Yann LeCun也曾说：“GAN及其变种是数十年来机器学习领域最有趣的想法。”

我们先从感性上理解什么是GAN。举个例子：警察与假钞的故事，二者满足两个对抗条件：

1. 罪犯不停的更新制造假钞的技术以防警察识别。
2. 警察不停的提高对假钞的鉴别能力来识别假钞。

罪犯要想假钞不被识别出来就要不断提升制造假钞的技术达到以假乱真的水平，而警察想要鉴别假钞就要尽可能的提升自身鉴别假钞的能力，尽最大可能识别出假钞，以阻止假钞在市面上流通。两者在博弈过程中不断汲取总结经验，从而稳步提升各自水平，这就是对抗。而这个过程是一个交替的过程，也就是说两者是交替提升的，两者水平相当，互相博弈提升各自水平。

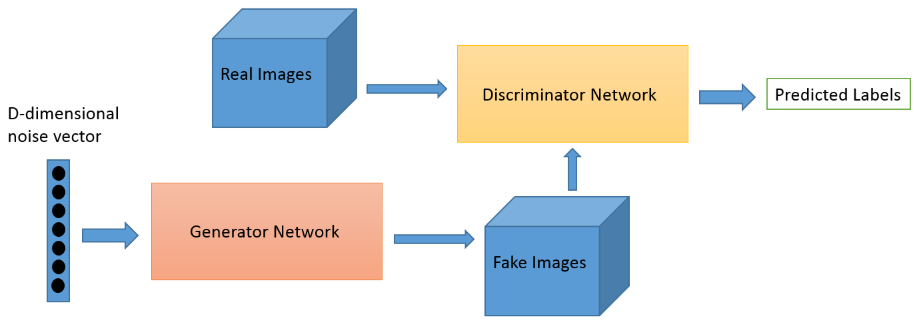


图 13‑1 GAN网络模型

GAN的学习训练过程也是一个如上的动态博弈过程。生成对抗网络由一个生成网络（Generator Network，G）与一个判别网络(Discriminator Network，D)组成。生成网络从潜在空间（latent space）中随机采样作为输入，其输出结果需要尽量模仿训练集中的真实样本。判别网络的输入为真实样本或生成网络的输出，其目的是将生成网络的输出从真实样本中尽可能分辨出来。而生成网络则要尽可能地欺骗判别网络。两个网络相互对抗、不断调整参数，其目的是将生成网络生成的样本和真实样本尽可能的区分开，使得生成器和判别器都能达到纳什均衡。

令x为代表图像的数据，D(x)是判别器网络，输出的概率为x是来自真实的训练数据还是生成器所生成的图像的概率。在训练过程中使得x来自训练数据时D(x)尽量接近1，x来自生成器时D(x)尽量接近0。D(x)也可以被认为是传统的二进制分类器。

对于生成器网络，z为从标准正态分布采样的潜在空间向量。G(z)表示生成器函数，该函数将矢量z映射到数据空间。生成器的目标是拟合训练数据

的分布，以便可以从该估计分布中生成假样本。

所以，D(G(z))是生成器*G*输出是真实的图像的概率。如Goodfellow的论文所述，*D*和*G*玩一个minmax游戏，其中*D*尝试最大化其正确分类真假的可能性*logD*(*x*)，以及*G*试图最小化以下可能性*D*会预测其输出是假的*log*(1−*D*(*G*(*x*)))。GAN的损失函数可表示为：



从理论上讲，此minimax游戏的解决方案是，鉴别者会盲目猜测输入是真实的还是假的。

以MNIST手写数据集为例,GAN网络工作的流程如下：

（1）生成网络从随机数据开始，生成一张图像；

（2）生成的图像被输入到鉴别器中、鉴别器判断它与ground truth数据之间的差异；

（3）鉴别网络分别考虑他们真假的可能性；

得到两个网络的反馈：

（1）鉴别网络循环反馈数据与ground truth之间的差异；

（2）生成网络持续接受鉴别网络反馈，不断优化生成器网络。

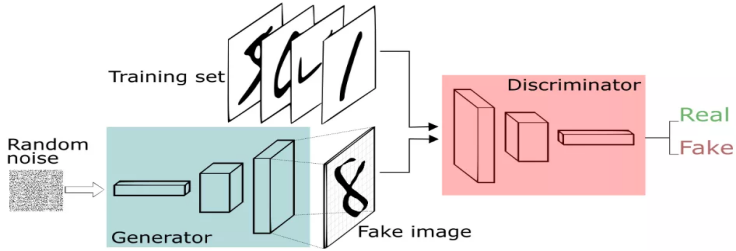


图 13‑2 MNIST数据集在GAN网络中的表现

在GAN的实际训练中，GAN的收敛理论仍在积极研究中，实际上GAN的训练过程是不稳定的，生成器和判别器之间的对抗性训练可能导致训练过程中的震荡和不稳定性，使得模型难以收敛或生成低质量的样本，常常会遇到梯度消失/爆炸问题。此外当生成器和判别器的能力差距较大时，生成器可能会陷入模式崩溃（Mode Collapse）的情况，只生成局限的样本模式而无法涵盖整个数据分布，这导致生成的样本缺乏多样性和创造力。

## GAN的应用

### 图像生成

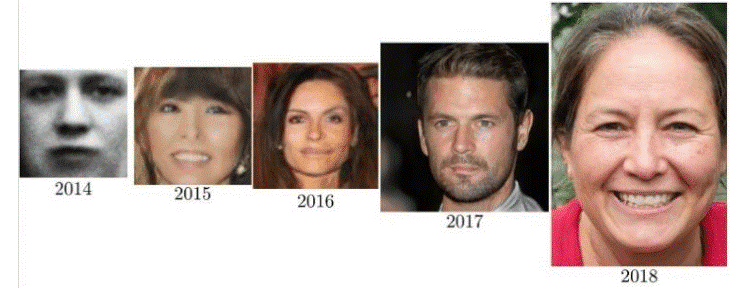


图 13‑3 GAN生成人脸

生成对抗网络最早应用可以追溯到2014年由lan Goodfellow等人提出[1]，生成器通过学习从随机噪声中生成图像样本，而判别器则学习区分生成器生成的样本和真实图像。GAN的初期应用主要集中在生成逼真的图像样本，例如生成手写数字、人脸图像等。在2014年，只能用于生成低分辨率的、粗糙的黑白人脸照片或手写的数字照片，而卷积神经网络网络的出现，使得GAN模型在图像领域涌现了一大批非常优秀的结果。

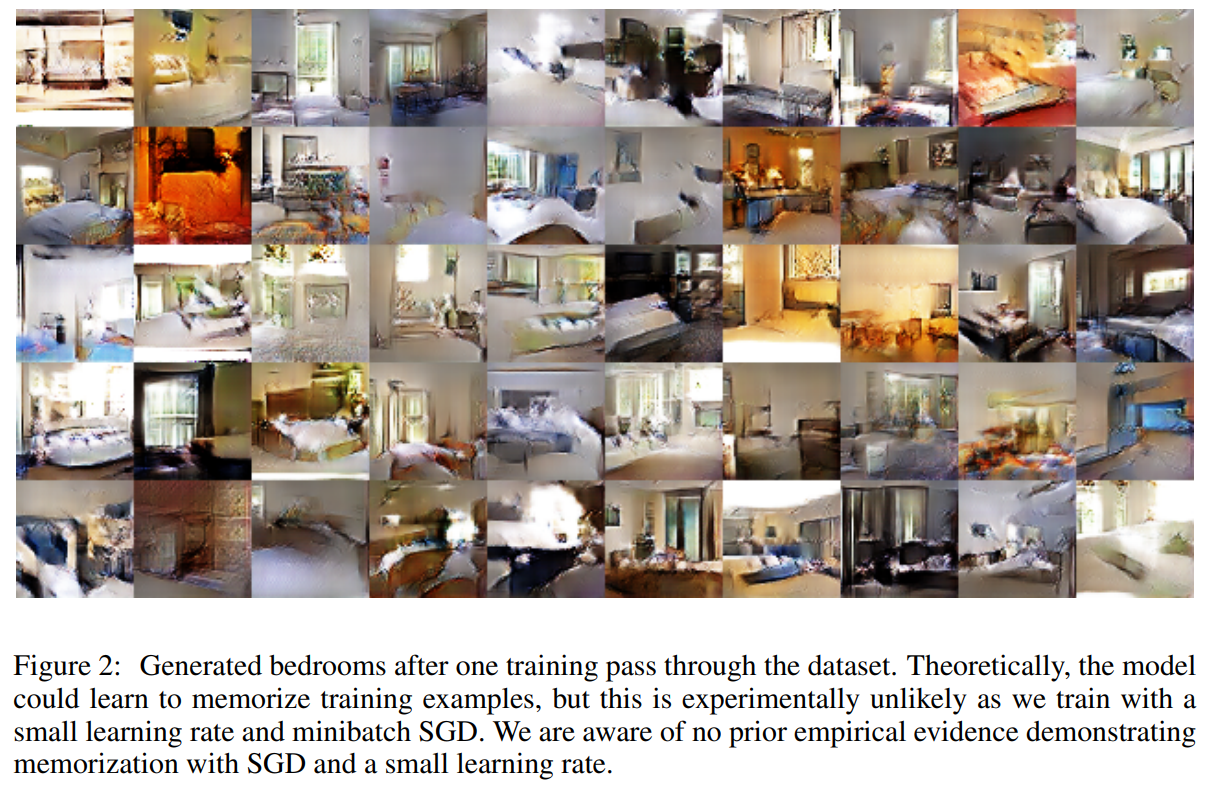


图 13‑4 DCGAN生成卧室图片

2015年Alec Radford等人提出Deep Convolutional Generative Adversarial Networks（DCGAN）[2]，第一次将卷积神经网络成功实现GAN模型并取得了非常好的结果。DCGAN将深度卷积神经网络CNN与生成对抗网络GAN结合用于无监督学习领域。最早的时候，GAN采用的是全连接层。而后也有一些基于卷积神经网络的GAN实现，但是相对来说，DCGAN的最终表现与同期的模型相比更优秀。这是GAN研究的一个重要里程碑，因为它提出了一个重要的架构变化来解决训练不稳定，模式崩溃和内部协变量转换等问题，为之后GAN在深度学习领域中打下了基础。

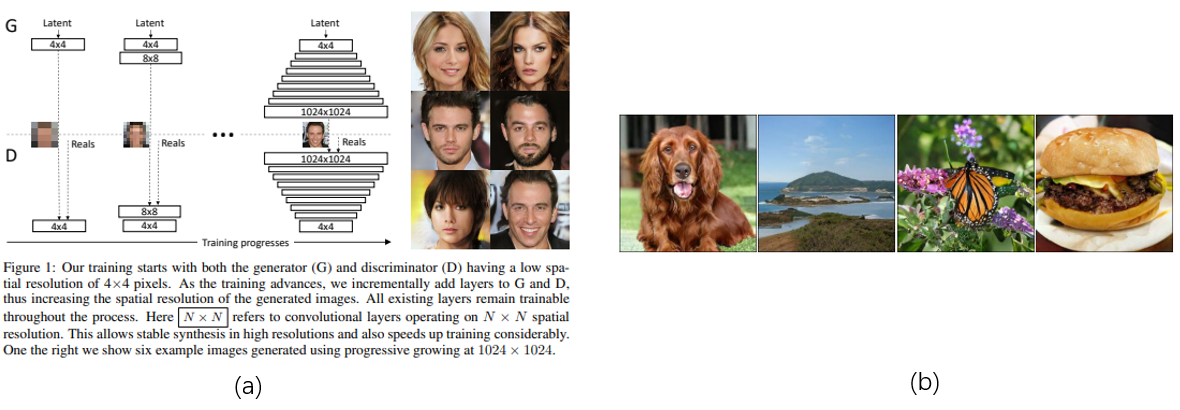


图 13‑5 (a) PGGAN (b) BigGAN图像生成

在DCGAN之后，涌现了一大批GAN在深度学习领域出色的成果。Tero Karras等人在2016年提出的Progressively Growing of GANs（PGGAN）[3]，采用了一种多尺度的架构对图像从4\*4到8\*8一直提升到1024\*1024的分辨率，极大的提高了GAN模型图像生成的质量。Andrew Brock等人在2018年提出的BigGAN[4]采用了自注意力机制（Self-Attention）和谱归一化（Spectral Normalization）等技术，在ImageNet上足以生成以假乱真的图片。

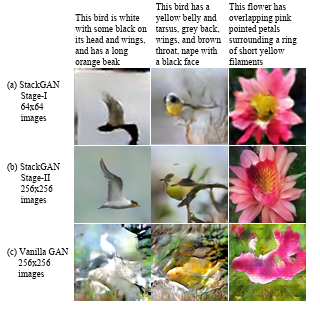


图 13‑6 StackGAN通过文本指导图像生成

而由Han Zhang等人在2016年提出的StackGAN[5]打通了文本到图像的通道，探索了使用文本指导生成图像的合成，得到了非常好的结果，通过提供文本描述，可以生成逼真的符合文本信息的图像。

图像生成在生成式AI（Artificial Intelligence Generated Content，AIGC）领域备受欢迎，近期的图像生成模型如DALL·E2[6]、Imagen[7]、Stable Diffusion[8]等，在图像生成上实现了前所未有的图像质量和模型灵活性水平，扩散模型也占据了主导地位。但由于其依赖迭代推理模式，实现简单目标稳定训练的代价就是推理过程计算成本的增加。

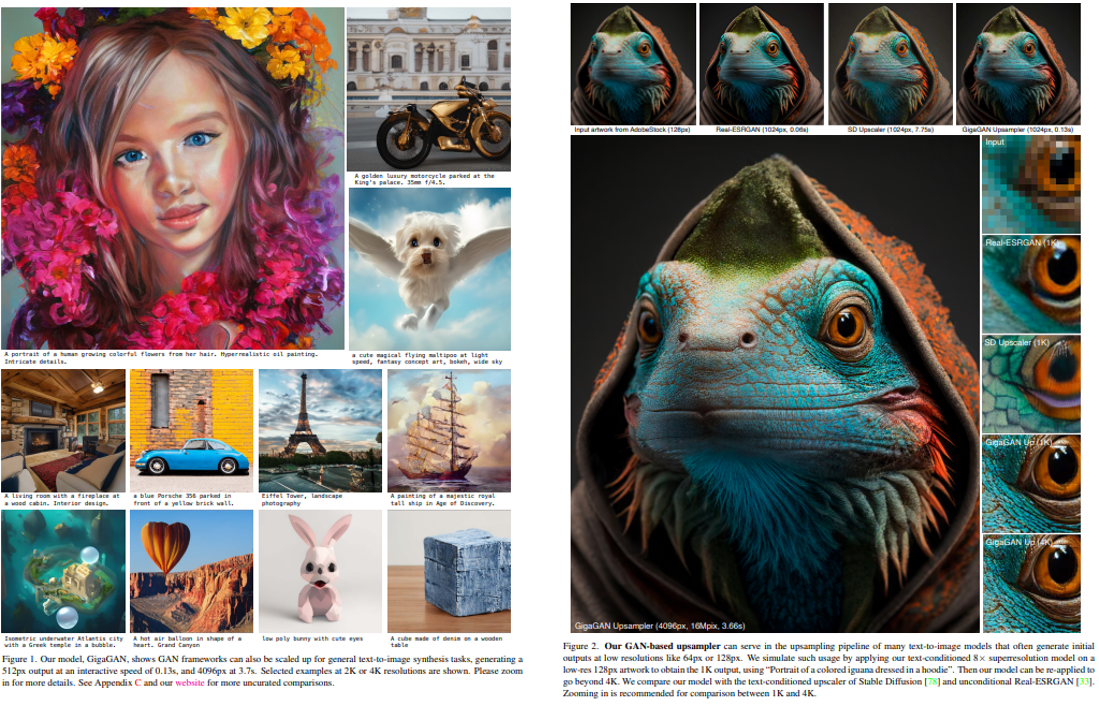


图 13‑7 GigaGAN图像生成

GAN一向是图像生成领域常用的基础架构，通过单向前项传递生成图像类比于扩散模型更显高效，但训练过程的不稳定使得GAN在复杂数据集上训练极具挑战性。现在，在一篇CVPR 2023论文中，来自POSTECH、卡耐基梅隆大学和Adobe研究院的研究者们联合提出具有十亿参数的新模型GigaGAN[9]，在大型数据集（例如LAION2B-en[10]）上实现了稳定和可扩展的训练，并采用多阶段方法足以生成4K分辨率的图片。类比于扩散模型和自回归模型，速度上提升了数十倍，能够在0.13秒内生成512像素的图像，其次可以在3.66秒内合成4k分辨率的超高分辨率图像。

GAN第一次从实验室走向公众一项标志性事件就是《Edmond de Belamy》的拍卖[11]。2018年美国当地时间10月25日早晨，一场不同寻常的拍卖会在纽约佳士得拍卖行举行——史上首次公开竞拍由人工智能创作的绘画。由GAN创作的Edmond de Belamy在佳士得拍卖会上被一位通过电话参与竞拍的神秘买家以43.25万美元（约合300万人民币）的价格成交。这幅画是一个法国巴黎艺术组合Obvious利用GAN生成的，连左下角的签名也是GAN的经典损失函数。



图 13‑8 《Edmond de Belamy》

### 风格迁移

风格迁移任务旨在将一个图像的风格应用到另一个图像上，从而实现风格的转换。GAN在风格迁移中的应用通常使用变体的生成对抗网络，如CycleGAN[12]、Pix2Pix[13]等。这些模型通过对抗训练学习到两个领域之间的映射关系，使得输入图像的内容保持不变，但风格被转换为目标风格。这种方法使得艺术创作、图像编辑等方面具有了更大的创造力和灵活性。

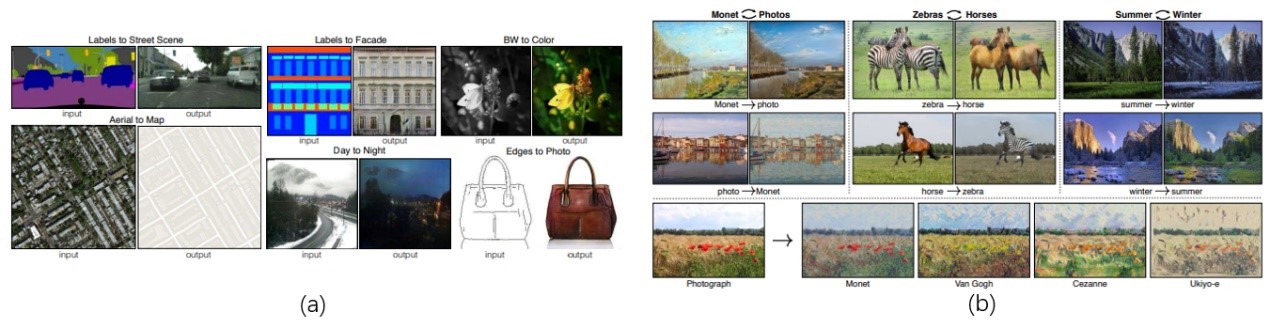


图 13‑9 (a) Pix2Pix, (b) CycleGANs

由Phillip Isola等人在2016年提出的Pix2Pix模型[13]，旨在将输入的图像从像素级别实现图像翻译（图13-9（a））。Pix2Pix模型为“Image-to-Image translation”这一类问题使用GAN提供了一个通用基础框架，采用了U-Net网络提升细节，并且使用PatchGAN来处理图像的高频部分。由于Pix2Pix模型的训练数据必须要求成对出现，而在现实生活中找到两个画风成对出现的图片是相当困难的。CycleGANs[12]应运而生，它只需要两种不同绘画风格的图片，而不需要严格的对应关系就可以实现更为广泛的图像风格转换，为GAN在风格迁移领域提供了更多有趣的可能（图13-9（b））。



图 13‑10 StyleGAN

由Tero Karras等人提出的StyleGAN[14]以及StyleGAN2[15]一直都是Image Generation领域的经典模型，通过设置一连串的全连接层来解决图像特征中的解耦问题（feature disentanglement），划分生成图片的latent space，将图片清晰的划分成为若干线性子空间，每个子空间代表特定的图像特征，如面部朝向，头发长短，发色等特征。同时StyleGAN还为每个线性子空间添加了一定的随机噪声，以获得种类更加多样的生成图片。StyleGANs的图片生成质量在当时几乎超过了所有的SOTA算法，能够生成清晰稳定的图像，并且还可以生成同一张人脸在不同年龄，不同性别，不同姿势，不同发型等特征下的表现情况（图13-10）。

### 图像超分辨率

所有会使用电脑的人在查看图片的时候都可能会对图片进行放大操作，想要观察地更加仔细，但是图片经过放大后会变得很模糊而导致看不太清。传统的图片放大方式通常采用线性插值的方式，虽然这些插值方法很快，但是往往会对图像质量产生非常大的损伤。由港中大多媒体实验室在2015年提出SRCNN[16]第一次将卷积神经网络应用到图像超分领域，起损失函数采用的为均方误差（MSE），根据构建出的超分图像与真实的高分辨率图像的差距去更新卷积神经网络的参数。但是这种方法采用的MSE损失，在训练过程中尽管MSE损失、PSNR（峰值信噪比）表现出非常不错的成绩，但往往会导致生成过于平滑的图像（图13-11）。

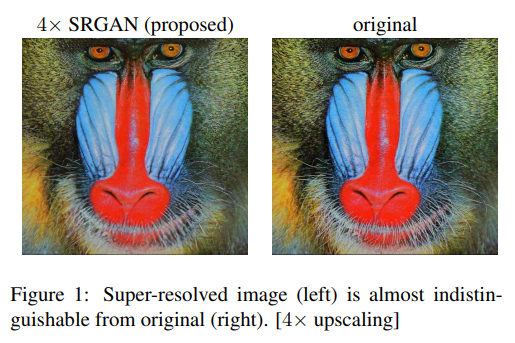


图 13‑11 采用MSE导致图像过于平滑

由Christian Ledig等人提出的SRGAN[17]第一次将GAN应用到图像超分领域，与普通的GAN模型的区别仅仅在于其生成器的输入数据不再是随机噪声，而是低分辨率的图像数据。SRGAN本质上提供了一种新的损失函数——感知损失（perceptual loss），同时旨在PSNR并不高的情况下，生成让人眼感知更加舒适的高分辨率图像，主要利用了GAN网络能够产生具有高感知质量的图像，使用感知损失去训练一个生成网络G。此外该网络主要由ResNet和GAN组成，所以整体的网络模型称之为SRGAN。总的来说SRGAN为GAN在图像超分领域开辟了一个先河，打下了良好的基础，从而引出了一系列GAN在图像超分领域非常优秀的研究成果。

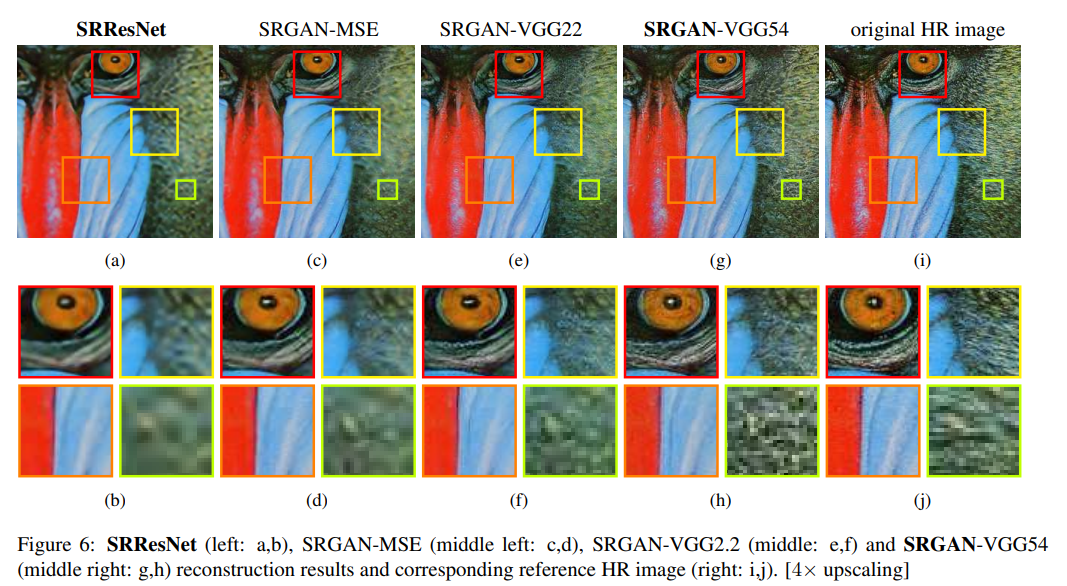


图 13‑12 SRGAN超分

## 参考文献

[1] I. Goodfellow *et al.*, “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[2] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.” arXiv, Jan. 07, 2016. doi: 10.48550/arXiv.1511.06434.

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation.” arXiv, Feb. 26, 2018. doi: 10.48550/arXiv.1710.10196.

[4] A. Brock, J. Donahue, and K. Simonyan, “Large Scale GAN Training for High Fidelity Natural Image Synthesis,” arXiv.org. Accessed: Oct. 09, 2023. [Online]. Available: https://arxiv.org/abs/1809.11096v2

[5] H. Zhang *et al.*, “StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks,” presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5907–5915. Accessed: Oct. 09, 2023. [Online]. Available: https://openaccess.thecvf.com/content\_iccv\_2017/html/Zhang\_StackGAN\_Text\_to\_ICCV\_2017\_paper.html

[6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical Text-Conditional Image Generation with CLIP Latents.” arXiv, Apr. 12, 2022. doi: 10.48550/arXiv.2204.06125.

[7] C. Saharia *et al.*, “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.” arXiv, May 23, 2022. doi: 10.48550/arXiv.2205.11487.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 10674–10685. doi: 10.1109/CVPR52688.2022.01042.

[9] M. Kang *et al.*, “Scaling Up GANs for Text-to-Image Synthesis,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10124–10134. Accessed: Oct. 09, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Kang\_Scaling\_Up\_GANs\_for\_Text-to-Image\_Synthesis\_CVPR\_2023\_paper.html

[10] C. Schuhmann *et al.*, “LAION-5B: An open large-scale dataset for training next generation image-text models”.

[11] A. Solea and G. Prezioso, *Global Arts Leadership in the Digital Age: Voices from the World’s Major Art Industries*. Cambridge Scholars Publishing, 2022.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks,” presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232. Accessed: Oct. 12, 2023. [Online]. Available: https://openaccess.thecvf.com/content\_iccv\_2017/html/Zhu\_Unpaired\_Image-To-Image\_Translation\_ICCV\_2017\_paper.html

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-To-Image Translation With Conditional Adversarial Networks,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134. Accessed: Oct. 12, 2023. [Online]. Available: https://openaccess.thecvf.com/content\_cvpr\_2017/html/Isola\_Image-To-Image\_Translation\_With\_CVPR\_2017\_paper.html

[14] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410. Accessed: Oct. 12, 2023. [Online]. Available: https://openaccess.thecvf.com/content\_CVPR\_2019/html/Karras\_A\_Style-Based\_Generator\_Architecture\_for\_Generative\_Adversarial\_Networks\_CVPR\_2019\_paper.html

[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8110–8119. Accessed: Oct. 12, 2023. [Online]. Available: https://openaccess.thecvf.com/content\_CVPR\_2020/html/Karras\_Analyzing\_and\_Improving\_the\_Image\_Quality\_of\_StyleGAN\_CVPR\_2020\_paper.html

[16] “Image Super-Resolution Using Deep Convolutional Networks | IEEE Journals & Magazine | IEEE Xplore.” Accessed: Oct. 12, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7115171

[17] C. Ledig *et al.*, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690. Accessed: Oct. 12, 2023. [Online]. Available: https://openaccess.thecvf.com/content\_cvpr\_2017/html/Ledig\_Photo-Realistic\_Single\_Image\_CVPR\_2017\_paper.html