基于条件生成对抗网络的图像翻译综述

屠杭垚[1,2], 王万良[3)*], 陈嘉诚[3)], 李国庆[3)], 吴菲[3)]

[1)] (浙大城市学院计算机与计算科学学院 杭州 310015)
[2)] (浙江大学计算机科学与技术学院 杭州 310058)
[3)] (浙江工业大学计算机科学与技术学院 杭州 310014)
(zjutwwl@zjut.edu.cn)

摘 要: 图像翻译旨在实现多组不同领域图像间的转换, 同时需要约束样本空间与目标空间分布的一致性. 文章旨在寻找条件生成对抗网络与图像翻译问题的结合点, 首先, 介绍了数据集的特点, 指出了不同数据集图像翻译难易程度; 其次, 从数学表达、性质以及目标函数设计方法得出算法实现的不同方式; 将现有图像翻译分成 3 种类别——匹配图像翻译、非匹配图像翻译和多领域图像翻译, 并得出不同应用场景所对应的图像翻译类别, 即高清任务对应匹配图像翻译, 低成本任务对应非匹配图像翻译, 多样化任务对应多领域图像翻译; 将图像质量评价方法分为主观图像质量评价与客观图像质量评价, 并分析客观图像质量评价中全参考图像与无参考图像质量评价的适用范围; 最后, 总结条件生成对抗网络在图像翻译中的进展, 并分析算法后指出了模式崩塌, 模型可解释性和少样本等未来所需解决的问题.

关键词: 条件生成对抗网络; 图像翻译; 生成模型; 图像质量评价
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.19807

A Survey of Image Translation Based on Conditional Generative Adversarial Networks

Tu Hangyao[1,2], Wang Wanliang[3)*], Chen Jiacheng[3)], Li Guoqing[3)], and Wu Fei[3)]

[1)] (*School of Computer and Computing Science, Hangzhou City University, Hangzhou* 310015)
[2)] (*School of Computer Science and Technology, Zhejiang University , Hangzhou* 310058)
[3)] (*School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou* 310014)

Abstract: Image translation aims to achieve conversion between multiple sets of images in different fields, and at the same time needs to constrain the consistency of the distribution of the sample space and the target space. The article aims to find the combination of conditional generative adversarial networks and image translation problems. Firstly, it introduces the characteristics of the datasets, pointed out the difficulty of image translation in different datasets. Secondly, derived different ways of algorithm implementation from mathematical expressions, properties and objective function design methods; divided existing image translation into three categories—paired image translation, unpaired image translation, and multi-domain image translation, and obtained the image translation categories corresponding to different application scenarios: that is, high-definition tasks correspond to paired image translation, low-cost tasks correspond to unpaired image translation, and diversified tasks correspond to multi-domain image translation. The image quality evaluation method is divided into subjective image quality evaluation and objective image quality evalua-

收稿日期: 2022-06-27; 修回日期: 2023-09-26. **基金项目**: 国家自然科学基金(61873240). **屠杭垚**(1995—), 男, 博士, 主要研究方向为计算机视觉、图像处理; **王万良**(1957—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 论文通信作者, 主要研究方向为人工智能; **陈嘉诚**(1997—), 男, 博士研究生, CCF 学生会员, 主要研究方向为计算机视觉、图像处理; **李国庆**(1994—), 男, 博士研究生, CCF 学生会员, 主要研究方向为多目标优化; **吴菲**(1995—), 女, 博士研究生, CCF 学生会员, 主要研究方向为动态多目标优化.

tion, and the applicable scope of full reference image and no reference image quality evaluation in objective image quality evaluation is analyzed. Finally, we summarized the progress of conditional generative adversarial networks in image translation, and analyzed the algorithm to point out the problems that need to be solved in the future such as mode collapse, model interpretability and few samples.

Key words: conditional generative adversarial networks; image translation; generative model; image quality evaluation

图像翻译是计算机视觉中常见的任务，旨在将一组领域的图像转换为另一组领域的图像，同时在翻译的过程中尽可能地约束样本和目标样本，使其具有一致的分布性. 图像翻译最早可以追溯到 Resales 等[1]提出使用贝叶斯理论得到样本类似的分布，通过推测图像类似分布空间，获得翻译图像. 随着神经网络的发展，以卷积神经网络(convolutional neural networks, CNN)为首的深度学习方法拉开序幕. 例如，Gatys 等[2]率先通过深度 CNN 分离和重组图像实现了图像翻译，随后 Taigman 等[3]提出了域传输网络(domain transfer networks, DTN)，首次提出了 2 个不同领域间图像转换的方法，实现了新的实体图像的翻译. 扩散模型(diffusion model, DM)定义了一个概率分布转换模型，它的前向传播过程，可以将一个复杂的分布转换为一个标准正态分布，实现了图像翻译，是一个较为前沿的发展方向. 受博弈论中零和博弈的思想启发，Goodfellow 等[4]提出了一种生成对抗网络(generative adversarial networks, GAN). 其广泛应用于图像合成、语义分割、语言处理、语音识别和医学图像处理等领域. 但在实际运用中，GAN 相对 CNN[5]不易收敛，容易出现模式崩塌且在不平衡数据集中更容易陷入过拟合的状态. 因此，现阶段出现了各类 GAN 衍生模型[6-8]，用于解决因原始 GAN 的缺陷导致的各种问题. GAN 由生成器和判别器组成. 生成器需要训练生成逼真的样本使判别器不能区分；而判别器需要训练区分真实的样本和生成的样本. 生成器和判别器相互对抗训练提高各自的能力达到纳什均衡状态，进而完成网络训练的过程. 由于图像翻译的特性，即需要以一张图片作为先验信息，因此生成器的输入不再是单一的随机向量，而是某一领域的图像. Mirza 等[9]提出了条件生成对抗网络(conditional generative adversarial network, CGAN)，它使生成器的输入不再是一组随机向量，而可以是特定的文本和图像，为 GAN 实现图像翻译提供了重要的理论基础. 综上，通过分析 GAN 的理论和应用发展状况，将图像翻译主要方法总结如图 1 所示.

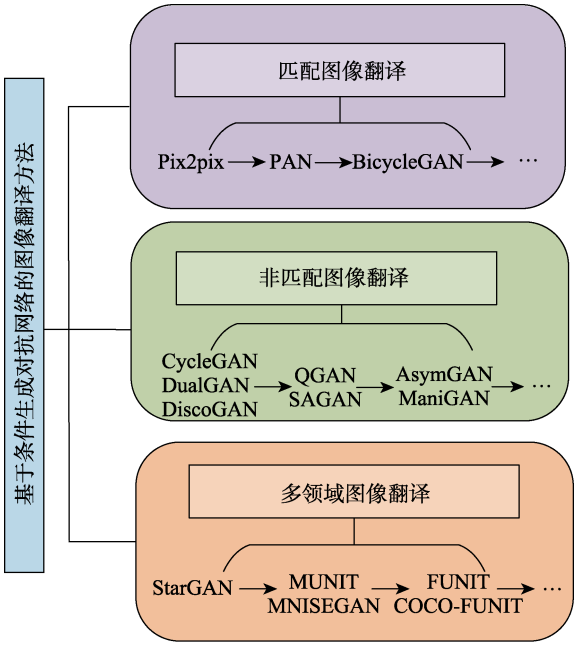


图 1　图像翻译主要方法总结

(1) 匹配图像翻译[10]. 匹配图像翻译是指数据集中输入与输出图像相互对应关系，即一对一映射. 由于输入与输出图像间关联度较高，网络训练较易，因此模型容易拟合所需的图像，但也因此使模型的输出较为单一，输出图像缺少多样性. 且在实际应用中，由于数据集具有采集数据难度较高、成本较大的特殊性，因此，该方法仅适用于实现高精度的任务.

(2) 非匹配图像翻译[11]. 非匹配图像翻译是指数据集中输入与输出图像不存在结构关联度，其使用目标域中任意图像作为源域对应的标签，降低了数据获取的难度. 不同于匹配图像翻译，其输入与输出图像间不存在对应关系，因此模型产生多样化的图像概率较大，但会存在一定的结构失真等情况，该方法适用于低成本的任务.

(3) 多领域图像翻译[12]. 匹配图像与非匹配图像翻译是 2 个领域间的相互转换, 但在实际任务中存在输出图像缺乏多样性的问题, 因此, 需要有多领域图像翻译的方法. 与上述 2 个方法相比, 多领域图像翻译需要增加网络输入噪声等一系列外部因素, 这将降低网络训练的稳定性. 因此多领域图像翻译存在着一定的实现难度.

由于图像翻译的类型的多样性, 因此对其有多种评价形式. 根据是否存在源域图像可以分为全参考图像质量评价与无参考图像质量评价. 全参考图像质量评价指现实中存在标准的答案, 即存在已知图像评判生成图像. 无参考图像质量评价指现实中不存在标准答案, 只能依靠像素、特征等判断图像的好坏.

1 数据集介绍

近些年出现了多种经典的图像翻译数据集, 为研究提供不同类型、不同大小、不同场景的应用数据, 可根据数据的对应方式将其大体分为匹配数据与非匹配数据. 下面将选取部分经典的数据集进行介绍.

(1) MNIST 数据集[13]. 是美国国家标准与技术研究所(national institute of standards and technology, NIST)数据集的一个子集, MNIST 数据集可在 http://yann.lecun.com/exdb/mnist 获取. 它是图像翻译入门级数据集, 由不同人手写数字构成. 该数据集包含 60 000 幅图像的训练数据集和 10 000 幅图像的测试数据集, 每幅图像都被裁剪为 28×28 像素, 采用人工标注各图像数字.

(2) Facades 数据集[14]. 是匹配数据集, 它收集了来自世界各个城市的高度结构化的建筑正面结构外观, 并针对每一幅图像进行了结构标注. 其翻译任务为将标注好的结构性图像翻译成现实中纹理丰富的建筑图像. 该数据集分别包含训练数据和测试数据 400 幅和 106 幅, 每幅图像都被裁剪为 512×256 像素, 是小样本数据集.

(3) Cityscapes 数据集[15]. 是匹配数据集, 收集了分别来自德国的 16 个城市、法国的一个城市和瑞士的一个城市的街道图像. 数据分为图像数据和注释数据, 它们一一对应. 其翻译任务是将注释数据翻译为现实中真实的街道数据. 该数据集分别包含训练数据、验证数据和测试数据 500 幅、1 525 幅和 2 975 幅; 每幅图像都被裁剪为 2 048×1 024 像素, 具有较强的实用性.

(4) Maps API 数据集. 是匹配数据集, 是以谷歌地图为基础, 选取各类城市的卫星地图和对应的二维地图. 卫星地图与二维地图相互对应, 该翻译任务将二维地图翻译成卫星地图. 该数据集分别包含训练数据和测试数据 1 096 幅和 500 幅, 每幅图像被裁剪为 1 200×600 像素, 也具有较强的实用性.

(5) Apple2orange 数据集[16]. 是非匹配数据集, 其将收集到的数据分为苹果数据和橘子数据 2 种. 翻译任务将苹果与橘子的图像相互转换. 其中, 苹果和橘子训练集分别包含 995 幅和 1 019 幅不同类型的图像, 苹果和橘子测试集分别包含 266 幅和 248 幅图像; 每幅图像被裁剪为 256×256 像素. 该数据集是非匹配转换方法中较为简单的一个数据集, 适合新手入门与练习.

(6) Summer2winter 数据集. 是非匹配数据集, 它将收集到的数据分为夏天的景色数据和冬天的景色数据 2 种. 其中, 夏天和冬天训练集分别包含 1 231 幅图像和 962 幅图像, 夏天和冬天测试集分别包含 243 幅和 190 幅图像; 每幅图像被裁剪为 256×256 像素. 夏天和冬天部分图像区分界限较为模糊, 加大了网络学习的难度, 因此, 该数据集更符合对网络性能的考察.

(7) Cezanne2photo 数据集. 是非匹配数据集, 这是深度学习和艺术领域结合的一个数据集. 其中, 训练集中包含了 6 287 幅普通景色的图像和 525 幅塞尚风格的图像, 测试集中包含 751 幅普通景色图像和 58 幅塞尚风格图像, 每幅图像被裁剪成 256×256 像素. 其翻译任务将普通景色图像转化成塞尚风格的图像, 由于塞尚风格图像存世量较少, 使此数据集被定义为不平衡数据集. 因此, 需要学者在提升翻译精确度的同时, 降低不平衡数据所带来的扰动.

(8) Face2anime 数据集[17]. 是非匹配数据集, 其翻译任务将真实图像风格人脸图像转换成日本动漫大师宫崎骏动漫风格人脸图像. 与普通人脸图像不同, 动漫图像具有清晰的界线感, 翻译至人脸图像需要有无明显界线感的动漫图像辅助. 因此, 该数据集中包含了 3 400 幅宫崎骏动漫人脸图像、对应的无明显界线感的动漫图像和普通人脸图像. 动漫人脸图像具有艺术图像不具有的清晰纹理和特有的分布, 使真实人脸图像转换为动漫图像具有一定的挑战性.

(9) CelebA 人脸数据集[18]. 是多领域人脸图像

数据集，分为图像数据集和特征数据集 2 部分. 其中，图像数据集包含了 202 599 幅人脸图像；特征数据集包含了不同人脸图像的属性，包括胡子、柳叶眉、大鼻子等 40 个类别属性.

2　GAN 和 CGAN 原理

GAN[4]中，G 为关于参数 θ 的生成器，$\boldsymbol{Z}$ 为随机向量；则生成器可表示为 $G(z)=X_{\text{fake}}\approx X_{\text{real}}$，即生成器生成的样本与真实样本服从同一分布. 在数学理论上，假设真实训练样本概率分布为 $P_{\text{data}}(x)$，则生成器输出的假样本概率分布为 $P_{\text{net}}(x;\theta)$，网络期望改变 θ 可使假样本逼近真实样本. 实际运算中并未能得到 $P_{\text{net}}(x)$ 的真实分布，因此采用采集大量真实样本来计算生成器的概率分布 $P_{\text{net}}(x;\theta)$ 是唯一有效的办法. 将生成器中的概率密度相乘得到，即 $L=\prod_{i=1}^{m}P_{\text{net}}(x_i;\theta)$. 其中，$m$ 为采样数量；x_i 为单一样本. 通过最大似然函数推导参数 θ^* 使 L 最大化. 通过公式变形导出 θ^* 与 KL 散度关系，即

$$\theta^*=\arg\min_{\theta}\text{KL}\left(P_{\text{data}}(x)\|P_{\text{net}}(x;\theta)\right).$$

设生成网络和判别网络的集合的价值函数表示为

$$V\left(\theta^{\text{D}},\theta^{\text{G}}\right)=E_{x\sim P_{\text{data}}}\left[\lg D(x)\right]+E_{x\sim P_z}\left[\lg(1-D(G(z)))\right]\tag{1}$$

固定生成器 G 为最优解，则令 $G(z)=x$ 对 $D(x)$ 求导得当 $D^*(x)=\dfrac{P_{\text{data}}(x)}{P_{\text{data}}(x)+P_{\text{net}}(x)}$ 时，V 最大. 将 $D^*(x)$ 代入原价值函数 V 中得到 P_{data} 与 P_{net} 的分布关系，即

$$V\left(G^*,D^*\right)=\text{KL}\left(P_{\text{data}}\left\|\frac{P_{\text{data}}+P_{\text{net}}}{2}\right.\right)+\text{KL}\left(P_{\text{net}}\left\|\frac{P_{\text{data}}+P_{\text{net}}}{2}\right.\right)-\lg(4)=2\times\text{JSD}\left(P_{\text{data}}\|P_{\text{net}}\right)-\lg(4)\tag{2}$$

其中，JS 散度(Jensen–Shannon divergence)具有非负性，当且仅当 $P_{\text{data}}=P_{\text{net}}$ 时，式(2)取得最小值 lg(4). 即最优化 G^*. 因此，式(1)的价值函数诠释了最大化判别器和最小化生成器的理念，完整的 GAN 的目标函数为

$$\min_G\max_D V(G,D)=\min_G\max_D E_{x\sim P_{\text{data}}}\left[\lg D(x)\right]+E_{z\sim P_z}\left[\lg(1-D(G(z)))\right].$$

GAN 通过 2 个网络的对抗训练解决一个网络下难以计算生成空间 $P_{\text{data}}(x)$ 的分布，因此在各类生成任务中有着良好的实用性.

监督学习[19-23]指通过有标签数据的数据集训练模型的一种训练方式，可以对未标签数据进行分类. 将监督学习的思想用于生成模型上，其期望结果为根据网络的标签输入生成与其对应的输出样本. 因此，为解决带标签数据的生成问题，CGAN[5]应运而生. 图 2 所示为 CGAN 的框架，它是对传统 GAN 的一个扩充. 它在 GAN 原有的模型上对判别器和生成器都加入一个辅助信息 y；y 可以是一项分类标签，一段语音或者一幅图像. 在图像翻译中，辅助信息 y 为某一领域的图像. CGAN 的目标函数在原有的基础上加入了条件 y，即

$$\min_G\max_D V(D,G)=E_{x\sim P_{\text{data}}}\left[\lg D\left(x|y\right)\right]+E_{z\sim P_z}\left[\lg\left(1-D\left(G\left(z|y\right)\right)\right)\right].$$

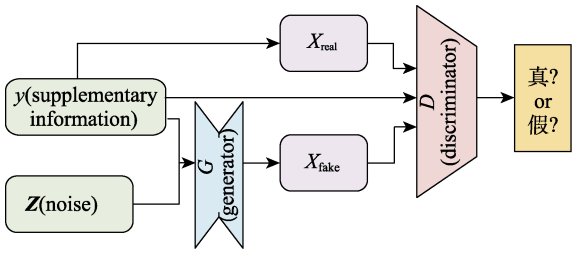


图 2　CGAN 框架图

在生成器中，将随机输入 $\boldsymbol{Z}$ 与条件输入 y 进行组合，形成一个全新的隐含表示. 而在判别器中真实数据 X_{real} 和生成数据 X_{fake} 都与条件 y 一起输入判别网络中，实现网络对抗训练.

此外，根据网络训练方式不同，多种模型也被提出. 例如，比较被大众广泛接受的 Wasserstein 生成对抗网络[24](Wasserstein GAN, WGAN)，带有梯度惩罚的 WGAN[25](Wasserstein GAN with gradient penalty, WGAN-GP)等.

WGAN 利用 EM 距离代替 JS 和 KL(Kullback-Leibler)散度，解决了生成与真实分布的距离衡量，从而改进了原始 GAN 存在的 2 类问题. WGAN 训练方式为 $W\left(P_{\text{r}},P_{\text{g}}\right)=\inf_{r\sim\Pi\left(P_{\text{r}},P_{\text{g}}\right)}E_{(x,y)\sim r}\left[\|x-y\|\right]$. 其中，$r\sim\Pi\left(P_{\text{r}},P_{\text{g}}\right)$ 是 P_{r} 和 P_{g} 组合的所有联合分布的集合. 对于每一个可能的联合分布 r，都可以从中

采样 $(x,y)\sim r$ 得到一个真实样本 x 和生成样本 y，并计算 2 个样本间 L_1 距离的期望的下界.

WGAN-GP 认为，虽然 WGAN 稳定了训练过程，但容易出现难收敛的情况，其原因为 WGAN 在训练时并未满足拉普拉斯连续，即定义域内每点的梯度恒定不超过某个常数.

WGAN-GP 增加了一个正则项来解决拉普拉斯连续问题，即对于原始输入的梯度的 L_2 范数约束在 1 附近，目的是将生成分布推向真实分布. WGAN-GP 公式为

$$L=\underset{\tilde{x}\sim P_g}{E}\left[D(\tilde{x})\right]-\underset{x\sim P_r}{E}\left[D(x)\right]+$$
$$\lambda\underset{\hat{x}\sim P_{\hat{x}}}{E}\left[\left(\left\|\nabla_{\hat{x}}D(\hat{x})\right\|_2-1\right)^2\right].$$

其中，$\hat{x}=\epsilon\tilde{x}+(1-\epsilon)x,\epsilon\sim U[0,1]$.

3 匹配数据集图像翻译

匹配数据集指在训练集中源域图像和目标域图像一一对应. 其应用领域为给手绘图像上色、黑白图像翻译为彩色图像以及图像去雾去雨等，包含了通用模型和指定任务的模型等.

3.1 图像翻译算法介绍

端对端图像翻译(image-to-image translation, Pix2pix)[26]是一个通用图像翻译框架，其以深度卷积生成对抗网络(deep convolution generative adversarial networks, DCGAN)[27]为基础，采用卷积、批量归一和 ReLU 激活函数来设计网络. 如图 3 所示，Pix2pix 的生成器 G 为学习源域图像和目标域图像的映射. 判别器 D 判断输入图像是真是假. 其中，生成器采用 U-Net 结构，以保证网络丢失更少原始信息；判别器采用 patchGAN[26]结构，通过局部判别提高其鲁棒性. Isola 等[26]为增强生成图像的真实性，在计算网络目标函数时，加入 L_1 损失对传统损失函数进行优化. 最终生成器目标为

$$G^*=\arg\min_G\max_D E_{x,y}\left[\lg D(x,y)\right]+$$
$$E_{x,z}\left[\lg(1-D(x,G(x,z)))\right]+$$
$$\lambda E_{x,y,z}\left[\left\|y-G(x,z)\right\|_1\right].$$

其中，λ 为超参数，可以根据情况调节. 当 λ=0 时，表示不采用 L_1 损失.

鉴于 Pix2pix 的良好性能，基于 Pix2pix 的应用算法广泛被专家和学者提出. 为了提高深度图像的质量和分辨率，Wang 等[28]提出了 Pix2pixHD 框架，实现了能够得到分辨率为 2048×1024 大小的具有视觉吸引力的结果. 该框架中将生成器中 G 分为 2 个子网络 G_1 和 G_2，其中，G_1 用于训练分辨率较低的图像，而 G_2 在外嵌入 G_1 中用于训练分辨率较高的图像，G_1 和 G_2 均由多个残差块组成. 鉴于 Pix2pixHD 的特殊结构，其 3 个尺度判别器用于鉴别真实图像和生成图像的 3 个尺度，对每层图像进行训练鉴别，3 个尺度判别器结构相似，但输入尺寸不同.

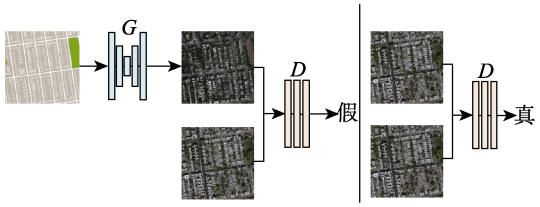


图 3 端对端图像翻译算法框架图

此外，Wang[29]提出了感知对抗网络(perceptual adversarial networks, PAN)，在 Pix2pix 基础上通过加入感知损失，实现了通用像素至像素的图像转换. Wang 等[30]还通过寻找图像中最不真实的区域进行重点判别，使用对抗训练实现图像高质量翻译. 针对 Pix2pix 对噪声的不敏感导致多样性差的问题，Zhu 等[31]提出匹配多模态图像翻译方法(toward multimodal image-to-image translation, BicycleGAN)，方法通过 cVAE-GAN[32]和 cLR-GAN[33] 2 种模型的组合，强制生成器不忽略噪声，使用噪声生成多样性结果. 自 Pix2pix 提出后，其变形式产生了井喷式的发展[34-39]，并在风格转换[26,29]、图像去噪[40-42]、医学图像处理[43-44]等应用中均取得了优秀的结果.

3.2 匹配数据图像翻译应用

第 3.1 节介绍了一对一映射框架及其各类变式，但方法均为探究匹配图像翻译的通用框架. 根据 Adam 提出的没有免费午餐(no free lunch)定理[42]，即所有问题出现的机会相同，或者说所有的问题同等重要时，若学习算法 L_A 在某些问题上比学习算法 L_B 要好，那么必然存在另一些问题，在这些问题中 L_B 比 L_A 表现更好. 因此在实际应用时，需要根据问题的情况提出相对应的算法解决问题. 图像翻译应用范围很广，下面列举几项较为常见的图像翻译应用方向.

3.2.1 图像去雾

图像去雾是计算机视觉中常见的任务[45-46]，它是将一幅有雾霾的图像转换为一幅无雾霾的图像，其本质与图像翻译相近. 目前主流的采用深度学习解决图像去雾问题分为 2 种：直接使用端到端模

型实现图像去雾，以及结合大气散射模型的深度学习方法实现图像去雾；它们各有优势，本节将简单介绍部分算法.

(1) 端到端模型. Qu 等[47]提出了增强 Pix2pix 去雾网络(enhanced Pix2pix dehazing network, EPDN)，其为生成器被判别器引导生成无雾图像后，使用增强器产生真实的高清无雾图像，增强器采用金字塔池化模块. 其中，对抗网络部分类似于 Pix2pixHD 结构，使用多尺度网络生成高清图像. 基于 EPDN，Bu 等[48]通过加入了多个引导过滤层来提取图像高频特征，将高频特征与生成器级联得到无雾图像. Li 等[49]提出 BPFD-Net，其生成器使用快捷模型将图像的上下层、颜色和纹理间的特征联系起来，以加强网络非线性特性，使得生成图像和输入图像在相同未知像素上保持一致，以确保图像的真实性. 从整体看，直接使用端到端模型略逊色于结合大气散射模型的深度学习算法.

(2) 结合大气散射模型的深度学习算法. 以纯深度学习为代表的图像去雾模型也存在缺陷. 深度学习采用卷积提取图像中的特征来解决一系列任务，如分类、分割和翻译等，但传统方法也具备优秀的提取特征的能力. 因此，考虑深度学习结合传统方法是一项十分有必要的工作. 传统去雾算法依靠估计透射率 $t(x)$ 通过大气散射模型得到无雾图像，但在出现含有天空的图像中效果并不理想. Jiao 等[50]证明透射率是可微的，并结合大气散射模型采用 Pix2pix 网络生成透射率图，得到了较为清晰的、命名为 Guided-Pix2pix 无雾图像. 屠杭垚等[51]采用图 4 所示双解码器结构生成无雾图像和并估计率图 $t(x)$，它们经过大气散射模型可还原有雾图像. 通过还原图像与原图像间的约束，使该算法产生高质量的无雾图像.

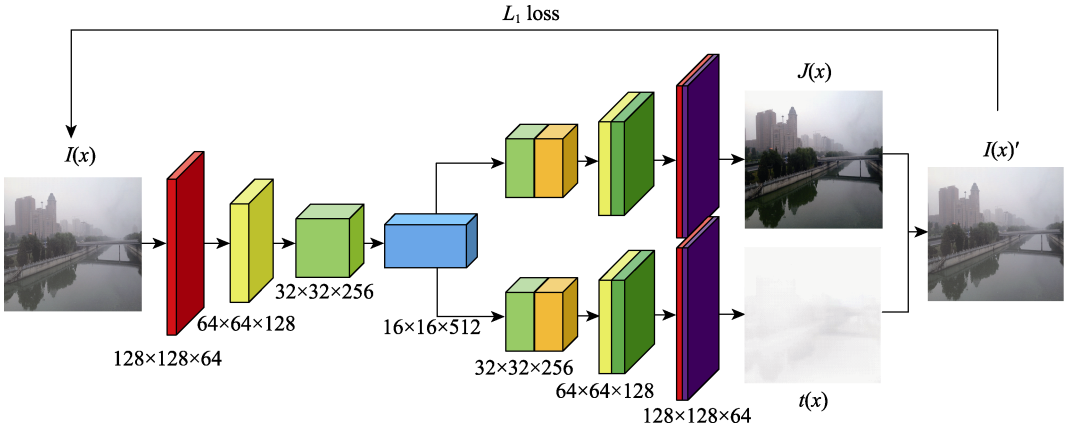


图 4　文献[51]生成器结构图

3.2.2　图像着色

将人类绘制的素描图像合成彩色图像是一个重要的问题[52-55]，例如，按照特定风格为素描图像着色，可以减少大量人工重复性工作. 随着计算机性能的提高，将深度学习应用于艺术领域取得了丰硕的成果，尤其是集中于将素描图像合成逼真图像. 根据所知论文的应用场景和难易程度不同，本节将其大体分为人脸素描图像着色和风景素描图像着色.

(1) 人脸素描着色. 由于真实场景中的光照变化和复杂的背景，且人脸须呈现出较清晰的边界和纹理，因此使用素描图生成逼真的人脸图像是一项具有挑战性的任务. Qi 等[56]提出了一种新颖的语义确定生成对抗网络(semantic-driven generative adversarial network, SDGAN). 其原理为对人脸照片进行面部显著性检测，以便提供全面的面部纹理结构；将面部纹理结构作为先验信息输入网络中，得到逼真的人脸图像. 随着 Transformer 框架[57]的发展，Zhu 等[58]在 GAN 中首次加入了 Transformer，提出轮廓转换(sketch-transformer, ST)方法，它可以捕捉远程空间依赖性，并生成合理的细节纹理.

(2) 风景素描着色. 与人脸素描着色不同，由于物体与相机之间的距离发生了变化，参考图像与目标域图像仅可能包含不同甚至多样性的图像内容占比，通过 Pix2pix 算法求解结果并不令人满意. 因此，各类风景图像着色方法被提出. 李洪安等[59]在 Pix2pix 基础上改进生成器的 U-Net[60]结构，使用 8 个上采样和下采样提取图像特征和预测结构颜色，并参考 cartoonGAN[61]，在损失中加入光滑 L_1 损失(smooth L_1 loss)[61]，弱化因上色而导致的纹理不清晰的缺点. 为了合成符合真实场景的颜色，Sketch-to-Art[62]引入双掩码机制，通过素描图

对特征图整形并翻译图像, 以更好地与参考图像风格保持一致. 为了实现稳定(克服上色时颜色过渡和边界模糊)的图像上色, Mehrotra 等[63]采用 Gabor 滤波器保留多尺度和多方向的细节特征, 以避免在网络下采样时丢失图像特性信息; 并在 Pix2pix 损失函数中加入惩罚项稳定网络训练过程, 在一定程度上降低了噪声对生成图像的影响.

3.2.3 医学图像处理

医学图像处理大体分为医学图像分割[43-44]和医学影像转换[64-65], 医学图像分割需要使用机器学习算法提取病灶, 以辅助医生决策. 在医学影像中, CT 图像和 MRI 图像特点不同, 其作用也不同, 但拍摄不同的影像会大大增加时间成本和物质成本, 因此采用算法转换各类图像信息是研究的主体, 本节介绍 Pix2pix 算法在医学中的应用.

(1) 医学图像分割. Pix2pix 是一种可用于图像分割的 GAN, 但是由于人体细胞类别较多, 因此 Tsuda 等[66]集成多个 Pix2pix 网络分割图像, 使各生成器和判别器分配到不同的任务, 达到高效学习的效果. Sato 等[67]认为, 在网络训练过程中, 生成器能力是弱于判别器的, 因此为了提高生成器的能力, 将判别器的特征图级联至生成器中, 使生成器了解判别器的分类准则, 进而提高生成器分割效率. Zhao 等[68]为实现 3D 细胞核分割, 采用定位细胞核位置、模拟细胞核间的相互作用和模拟细胞核的纹理各个阶段, 由粗到细地分割具体图像, 大大提高了分割性能.

(2) 医学影像转换. 医学影像转换被认为是医学图像分析的一个新前沿, 具有非常多的潜在应用. 其中最著名的是 Armanious 等[69]提出的 MedGAN. 由于医学影像的特殊性, MedGAN 通过一种新的生成器架构 CasNet 来增强图像转换后的清晰度. 且其在 PET 转 CT 图像、MR 运动伪影矫正和 PET 图像去噪数据集上均取得较优秀的效果, 打开了医学影像转换之路. 由于拍摄部位可能存在金属植入物或存在有限视野等因素, 医疗模式中局部变形是常见现象, ipA-MedGAN[70]在转换影像时, 通过修复变形区域是较为有效的方法. 为了保证转换图像的高保真度, Upadhyay 等[71]采用了如图 5 所示一种不确定性引导渐进方式训练 GAN 的注意力图. Primary GAN(图 5 中第 1 个网络)从输入图像域中获取输入图像, 获取生成的图像 $b_{[0]i}$, 以及不确定性引导图 $\alpha_{[0]i}$ 和 $\beta_{[0]i}$, 不确定性引导图可迫使网络关注图像中生成质量较差的不确定性区域, 从而提高图像生成质量. 注意力图由不确定性引导图获得, 它和生成图像级联后作为下一网络 Subsequent GAN-M 的输入(图 5 中第 M 个网络).

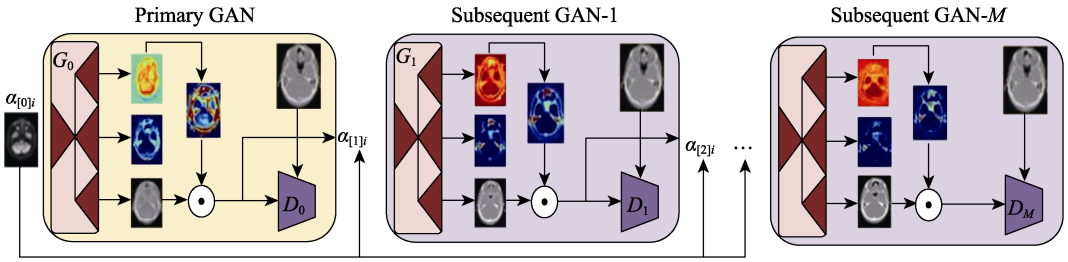


图 5 结合注意力机制的不确定性引导方式模型

4 非匹配数据集图像翻译

自 Isola 等[26]提出了 Pix2pix 算法, 掀起了对图像翻译研究的热潮. 但由于匹配图像的特殊性, 在现实生活中, 采集匹配图像成本高且难度大, 因此需要提出一种非匹配数据集的图像翻译算法.

4.1 非匹配数据图像翻译算法介绍

循环一致性生成对抗网络(cycle-consistent generative adversarial networks, CycleGAN)[72], DualGAN[73], DiscoGAN 均发现了匹配图像的缺陷, 几乎在同一时间, 三者均提出了相似的网络框架结构. 三者均采用循环一致性结构构建源域和目标域之间的映射. 循环一致结构如图 6 所示, 领域 X 的输入图像 1 通过生成器 1 翻译为生成图像 1, 生成图像 1 通过生成器 2 翻译为重建图像 1. 翻译为重建图像的目的是构造生成图像与输入图像间的一致结构, 同理, 领域 Y 也进行对偶操作. 判别器鉴别生成图像和对偶输入图像是否为同一领域, 实现网络的非匹配图像转换. 即

$$x' = G_2\left(G_1\left(x\right)\right) \approx x,\ y' = G_4\left(G_3\left(y\right)\right) \approx y;$$

$$\left(x_1, x_2, \cdots, x_n\right) \in X, \left(y_1, y_2, \cdots, y_n\right) \in Y.$$

其中, x 和 x' 分别为输入图像 1 和重建图像 1; y 和 y' 分别为输入图像 2 和重建图像 2.

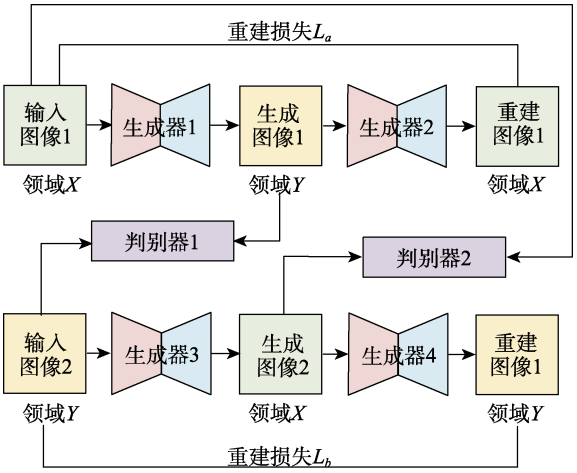


图 6　循环一致性结构训练方式

CycleGAN 采用了 Johnson 等[74]的风格转移架构，并在网络中加入残差网络来稳定网络训练，以提高生成图像质量. 其对抗损失函数为

$$L_{\mathrm{GAN}}\left(G_2,G_4,D_X,X,Y\right)=E_{x\sim P_{\mathrm{data}(x)}}\left[\lg D_X\left(x\right)\right]+$$
$$E_{y\sim P_{\mathrm{data}(y)}}\left[\lg\left(1-D_X\left(G_X\left(y\right)\right)\right)\right],$$
$$L_{\mathrm{GAN}}\left(G_1,G_3,D_Y,X,Y\right)=E_{y\sim P(y)}\left[\lg D_Y\left(y\right)\right]+$$
$$E_{x\sim P(x)}\left[\lg\left(1-D_Y\left(G_Y\left(X\right)\right)\right)\right].$$

除了对抗损失，CycleGAN 在重建图像和输入图像间加入了循环一致损失，目的是使生成图像与输入图像具有一致的结构. 因此循环一致损失为

$$L_{\mathrm{cycle}}\left(G_1,G_2\right)=E_{x\sim P_{\mathrm{data}}(x)}\left[\left\|G_2\left(G_1\left(x\right)\right)-x\right\|\right],$$
$$L_{\mathrm{cycle}}\left(G_3,G_4\right)=E_{x\sim P_{\mathrm{data}}(x)}\left[\left\|G_4\left(G_3\left(y\right)\right)-y\right\|\right].$$

其中，G_1 和 G_3 将领域 X 图像翻译为领域 Y 图像；而 G_2 和 G_4 将领域 Y 图像翻译为领域 X 图像. 因此，完整的目标函数为

$$L\left(G_1,G_2,G_3,G_4,D_X,D_Y\right)=$$
$$L_{\mathrm{GAN}}\left(G_1,G_3,D_Y,X,Y\right)+$$
$$L_{\mathrm{GAN}}\left(G_2,G_4,D_Y,X,Y\right)+$$
$$\lambda\left(L_{\mathrm{cycle}}\left(G_1,G_2\right)+L_{\mathrm{cycle}}\left(G_3,G_4\right)\right).$$

DiscoGAN 借鉴了 Pix2pix 的生成器，采用 U-Net 对称结构加入残差网络，它与 CycleGAN 不同，只能支持 64×64×3 的图像尺寸，但 CycleGAN 则可支持 256×256×3 的图像尺寸. 其完成目标函数与 CycleGAN 类似，均为对偶重建损失+GAN 损失. 而 DualGAN 在 CycleGAN 的基础上采用前 WGAN[75]的权重剪裁，由于 WGAN 在颜色映射方面有很大的优势，因此，DualGAN 在上色、颜色转换等任务中表现出出色的能力. DualGAN 的目标函数也与 CycleGAN 类似，均为重建损失和 GAN 损失. 自此三者为非匹配图像翻译打下良好基础.

4.1.1　基于增加约束的非匹配图像翻译方法

由于仅使用循环一致损失，约束效力较弱，会出现一定的重影或结构失真等情况. 因此，StyleGAN[76]根据图像质量评价方法来构造质量感知框架，优化源域图像和重建图像，从而减少生成图像的伪影. 考虑循环一致约束效力较弱，Sem-GAN[77]搭建了语义一致框架，其本质是翻译图像时使用语义对齐帮助网络搭建目标域图像的外观信息. 具体步骤如图 7 所示.

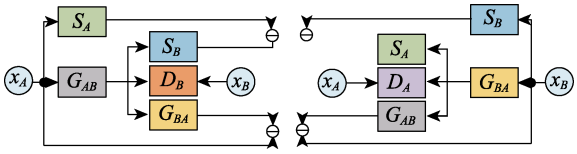


图 7　添加语义框架翻译方法

领域 X 图像利用 G_{AB} 和 G_{BA} 重建源域图像的同时，也利用 S_A 和 S_B 建立源域图像和重建图像的语义图像. 通过寻找重建图像中丢失的语义信息，强制生成图像得到源域图像结构信息是 Sem-GAN 的主要工作；同理，领域 Y 图像也如此. Sem-GAN 建立了循环一致约束和语义约束，为基于增加约束的非匹配图像翻译方法做出了一定的贡献.

4.1.2　结合注意力机制的非匹配图像翻译方法

注意力机制[78-80]是被广泛应用于自然语言处理、图像处理、语音识别的核心技术之一，它可以从大量信息中快速地筛选出高价值的信息帮助网络完成任务. 最早引入注意力机制的是自注意 GAN(self-attention generative adversarial networks, SAGAN)[81]，它把注意力机制引入 GAN 中，用于解决在生成图像中参数不规范细节的问题. 但 SAGAN 解决的是 GAN 中生成问题，并非图像翻译问题. 因此，根据注意力机制的特性，Tang 等[82]提出了使用注意力引导机制的 AttentionGAN，其在生成器上进行了改进：源域图像经过双生成器得到内容图像和注意力掩码，注意力掩码与源域图像相乘得到初步翻译图像，将其与内容图像相加得到最终翻译图像. AttentionGAN 解决了以往算法未能翻译图像高级的语义信息的短板，采用输出和注意力掩码结合的方式，在翻译图像的同时最小化背景变化. 但 AttentionGAN 仅在生成器中引入注意力机制. SPA-GAN[83]在判别器中引入了空间注意力机制，加强了判别器的能力. 由于 GAN 的最终训练任务是达到纳什均衡，即判别器的能力与生成器的能力均衡，因此提高判别器能力可以

从侧面促使生成器产生更逼真的翻译图像.

4.1.3 基于潜在空间的非匹配图像翻译方法

图像翻译的实质是在一个潜在联合分布中学习到边缘分布, 联合分布是指 2 组风格图像联合分布空间, 而边缘分布是指每组图像存在的分布空间. Liu 等[84]提出了无监督的图像到图像翻译网络(unsupervised image-to-image translation networks, UNIT), 它将 2 组源域图像编码至同一特征空间 $\boldsymbol{Z}$, 其使用生成器翻译对应风格的图像. 为保持翻译图像结构相似性, UNIT 也使用 CycleGAN 的循环一致损失, 其提出的领域分布的思想为后续算法提供了新的思路. Lee 等[85]认为, 来自单幅输入图像的多种可能输出使 2 个领域间的映射面临着挑战, 因此, 其提出一种基于解耦的表示方法(diverse image-to-image translation via disentangled representations, DRIT), 将图像嵌入到 2 个域空间中, 即捕获跨领域的共享信息的域不变的内容空间, 以及捕捉独立的领域信息的特定域属性空间. 翻译图像即为共享信息 C 和独立的领域信息 A. 图 8 所示为 CycleGAN, UNIT 和 DRIT 算法结构示意图. 其中, $E(x)$ 为编码器, $G(x)$ 为生成器, C 为共享信息, A 为独立的领域信息.

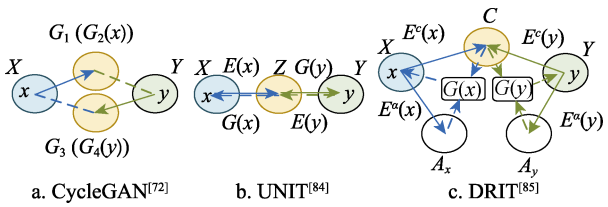


图 8 不同潜在空间非匹配算法结构示意图

4.1.4 非对称数据的非匹配图像翻译方法

CycleGAN 取得了非匹配图像翻译的重大进展, 其采用数据集 Summer2Winter 和 Apple2Orange 翻译的图像复杂程度相似, 取得了良好的翻译结果. 但将算法应用于 label2photo 等一系列复杂的图像翻译任务中, 就会出现质量差、映射歧义等问题. 因此, 需要非对称数据的非匹配图像翻译方法来解决问题. Li 等[86]率先提出非匹配非对称 GAN (asymmetric GAN for unpaired image-to-image translation, AsymGAN), 通过映入辅助变量来适应非对称数据的学习. 辅助变量的学习从简单域到复杂域的额外信息, 这样工作具有 2 个作用: (1) 平衡了 2 个域间的信息; (2) 调整辅助变量便于清晰映射建模, 避免信息不平衡导致的 2 个域间的映射模糊. 鉴于 CycleGAN 翻译非对称性人物中对结构的不敏感性, Dou 等[87]引入了源域图像和生成图像的边缘损失, 通过缩小边缘损失提高网络效率, 实现了近红外图像与彩色图像间非对称数据的学习. Sheng[88]为了在非对称数据下增强翻译图像的真实性, 通过加入双注意力机制模型来加强网络的颜色识别能力. 领域适应是一项典型且重要的图像翻译技术, 其旨在减少不同领域之间的差异, 通常用于匹配图像翻译方法. Hoffman 等[89]将领域适应与非匹配领域相结合, 实现了领域适应的非匹配图像翻译方法. Pizzati 等[90]通过域桥训练翻译网络, 将翻译工作扩展到无监督域适应, 并达到无监督域适应的先进性能. Tu 等[91]使用了卷积 2D 化结合亚像素卷积, 实现了非匹配图像翻译, 达到了多样性的翻译效果.

4.1.5 多模态数据引导的非匹配图像翻译

多模态数据引导指采用文本、视频等模态数据辅助图像完成翻译任务. Xu 等[92]采用文本引导图像模式, 使用图像-文本特征共享空间, 提高了图像翻译的质量. Tumanyan 等[93]利用预训练的文本到图像扩散模型的强大功能来生成符合目标文本的新图像, 同时保留引导图像的语义布局完成图像翻译任务. ManiGAN[94]包含文本图像仿射组合模块和细节校正模块 2 个关键组件, 文本图像仿射组合模块选择与给定文本相关的图像区域, 细节校正模块纠正不匹配的属性并补全合成图像的缺失内容, 完成图像翻译任务.

4.2 非匹配数据集算法的应用场景

与匹配数据集类似, 非匹配数据集应用场景也十分广泛, 其与匹配数据集应用场景存在重合, 本节重点介绍较为常见且与第 3 节未重复的应用方向.

4.2.1 艺术图像风格转换

随着深度学习的发展, 艺术与深度学习结合的例子屡见不鲜. 例如, 风景图像转换为梵高风格图像、风格图像转换为莫奈风格图像和现实图像转换为卡通图像等, 这一系列的工作可以为艺术家提供灵感. 但是, 与现实生活中应用的图像不同, 各类艺术图像有着特定的内容和表现形式, 需要加入不同的机制解决艺术图像的转换问题.

(1) 手绘图像风格转换. 手绘图像指梵高、莫奈和塞尚等风格的画像, 由于画家各自的风格较为鲜明, 因此网络容易识别属于画家的独特风格和共性风格. Fu[95]借鉴 DRIT 在生成器的反卷积模块中, 使用的风格编码器和内容编码器提取风格和内容特征. 且使用多尺度判别器, 辅助生成器完成翻译任务. Lin 等[96]认为, 在绘画领域中存在颜色和笔触间的差异, 但现有方法[97-98]大多侧重提高风格化结果的质量, 忽略了颜色与笔触细节的

差异. 因此, 多属性绘画生成(multi-attribute guided painting generation, MGPG)方法采用多层感知网络解析变量, 并聚合多属性向量完成图像翻译, 且为了增强判别器的能力, 其利用多任务判别器和样式保持约束来扩大不同域间的颜色和纹理差距.

(2) 卡通图像风格转换. 图像卡通化是近年来新兴的图像风格方式, 可以帮助设计师设计现实世界的卡通画像. 但由于卡通画像具有高度简化和抽象的独立特征, 以及清晰的边缘、纹理相对简单的特点, 现有方法描述纹理细节的损失并不适用. cartoonGAN[61]基于以上问题提出 2 种适合卡通化的损失: (1) 使用 VGG[99]网络特征图的稀疏正则化, 建立语义损失; (2) 加入对抗边缘损失, 建立边缘损失. cartoonLossGAN[100]通过模仿人工素描过程学习, 构建清晰的边缘, 并模仿人工着色过程为卡通图像着色, 实现了图像卡通化. 根据卡通图像翻译的特性, cartoonLossGAN[100]提出了一种初始化策略, 使模型更容易趋于稳定. AnimeGAN[17]重点关注动漫图像与真实场景图像间的转换, 提出灰度样式损失、灰度对抗损失和颜色重建损失 3 种损失, 解决了传统算法生成的动漫图像清晰度不足的问题, 同时保留了大量的原始图像内容. 从卡通图像的应用看出, 实现卡通图像风格转换的重点在于建立新的损失函数, 这为相关学者的研究提供了有效的思路.

4.2.2　图像超分辨率

图像分辨率的高低会影响人或者计算机的决策, 因此对低分辨率图像, 通常采取一定的方法来提高其分辨率. 传统一对一映射的方法需要真实的高分辨率图像[101-103], 并非盲超分辨率. 使用非盲超分辨率的方法只能在已知高分辨率(super resolution, SR)图像存在的情况下才能达到预期的效果, 这使图像超分辨率方法的意义大大降低. 因此, 非匹配算法的提出为盲超分辨率[104-107]提供了一种有效方式.

(1) 基于增加约束的超分辨率. 由于超分辨率数据集与传统非匹配数据集的略微差异(高分辨率图像与低分辨率图像大小不一致)存在未知退化的 SR 的问题. 与卡通图像风格转换相似, 增加损失约束可以提升超分辨率. Kim 等[108]通过加入逐像素损失、VGG 损失和 SSIM 损失, 引导生成图像遵循目标域分布, 获得了符合目标域分布的高分辨率图像. Wu 等[109]认为, 低频内容的不逼真和高频纹理的不精确, 是导致生成高分辨率图像结果不佳的原因. 因此, 采用数据约束的方法从损失函数、训练数据和后期处理 3 个方面去解决无监督学习存在的像素监督成效低、抑制伪影效果差和存在噪声等问题, 并取得了较好的效果.

(2) 基于领域差异的超分辨率. 由于训练数据和测试数据间有着一定的域差距限制, 仅增加约束的提升效果有限. 因此, Wei 等[110]提出通过域差距感知训练实现无监督真实场景图像超分辨率(unsupervised real-world image super resolution via domain-distance aware training, DASR), 通过域差距感知训练数据与测试数据间领域差, 采用域距离加权监督策略解决问题. 域差距感知训练可以从目标域中的真实数据得到额外的信息, 而域距离加权可以更合理地使用标记数据.

5　多领域图像翻译

5.1　多领域图像翻译算法

Pix2pix 模型解决了匹配图像翻译的问题, CycleGAN 模型解决了非匹配图像翻译的问题. 但它们都是解决一对一映射问题的, 即从一个领域翻译至另一个领域, 而不同领域中存在多种模态的图像, 存在多领域翻译的情况, 因此需要有针对多领域图像翻译的算法.

5.1.1　基于 StarGAN 的多领域图像翻译算法

Choi 等[111]首先发现多领域图像翻译任务所存在的问题(需要训练多个网络翻译), 并提出 StarGAN, 实现了单个网络中同时翻译不同领域的数据集. 为了让生成器学习多个领域间转换的能力, StarGAN 在生成器中添加目标领域信息, 即通知生成器需要翻译哪个领域, 而判别器除了需判断图像是否真实以外, 还需要判断图像属于哪个类别, 保证输出图像随目标领域的不同呈现不同效果.

为了生成更逼真的图像, StarGAN 在损失函数的设计上结合了对抗损失、领域分类损失和重建损失. 其中, 对抗损失为

$$L_{\text{adv}} = E_x\left[\lg D_{\text{src}}\left(x\right)\right] + E_{x,c}\left[\lg\left(1 - D_{\text{src}}\left(G\left(x,c\right)\right)\right)\right].$$

其中, G 尽可能地在标签 c 的条件下生成图像, 同时 D 尽可能地区分真实图像和生成图像.

为了区分生成图像属于哪一类别, StarGAN 建立了领域分类损失. 它由 2 部分组成: 用于帮助优化判别器的真实图像的领域分类损失, 以及用于优化生成器的生成图像的领域分类损失. 前者定义为 $L_{\text{cls}}^{\text{r}} = E_{x,c'}\left[-\lg D_{\text{cls}}\left(c'|x\right)\right]$, 其中, $D_{\text{cls}}\left(c'|x\right)$ 表示由 D 计算的域标签的概率分布; 后者定义为

$L_{\text{cls}}^{\text{f}}=E_{x,c}\left[-\lg D_{\text{cls}}\left(c\middle|G(x,c)\right)\right]$, 其中, G 用于尽力缩小生成图像与目标域图像的类别距离. 为了保证翻译后的图像保留其输入图像的内容, 且同时只改变与域相关的输入图像部分. StarGAN 通过重建损失达到重建图像内容的目的. 重建损失定义为 $L_{\text{rec}}=E_{x,c,c'}\left[\left\|x-G\left(G\left(x,c\right),c'\right)\right\|_1\right]$. 其中, c' 为网络输出标签; G 将翻译后的图像 $G(x,\ c)$和原始域图像标签 c 作为输入, 尝试重建原始图像 x.

综上, 生成网络和判别网络的整体目标函数分别为

$$L_{\text{D}}=-L_{\text{adv}}+\lambda_{\text{cls}}L_{\text{cls}}^{\text{r}},$$

$$L_{\text{G}}=L_{\text{adv}}+\lambda_{\text{cls}}L_{\text{cls}}^{\text{f}}+\lambda_{\text{rec}}L_{\text{rec}}.$$

其中, λ_{cls} 和 λ_{rec} 是超参数, 分别控制领域分类损失和重建损失.

StarGAN 虽解决了多领域翻译的问题, 但存在生成图像多样性不足、模态可扩展性不高等现象. Choi 等[112]针对 StarGAN 的不足, 提出了 StarGAN v2 框架, 以提高生成图像的多样性和多领域的可扩展性. 在 StarGAN 的基础上, StarGAN v2 引入映射网络和样式编码器, 映射网络学习将随机高斯噪声转换为样式代码, 而编码器学习从给定的参考图像中提取样式代码; 通过样式代码的方式将随机高斯噪声的多样性和扩展性附赠于 StarGAN v2, 解决 StarGAN 存在的 2 个问题.

StarGAN 在完成翻译任务会留下一些高频信息, 其侧重低分辨率的图像合成, 不适用于高清图像的翻译任务. 因此,在 StarGAN 的基础上, EStarGAN[113]加入 CSRIR 模块, CSRIR 模块本质为空间注意力和通道注意力的结合机制, 侧重于图像细节纹理的特征的提取; 可以更有效地学习深度网络中高频信息完成高清图像翻译.

5.1.2 基于风格编码器的多领域图像翻译算法

多领域图像的翻译框架可以保存源域图像的结构信息, 但不能很好地迁移翻译图像的风格. 因此, 研究人员提出加入风格编码器, 以解决图像风格翻译的问题. Sun 等[114]提出了无需独立风格编码器的多模态无监督图像到图像转换(multimodal unsupervised image-to-image translation without independent style encoder, MNISE-GAN), 重新使用判别器的前部网络作为风格编码器, 其优势在于简化了整体网络结构, 重新使用判别器减少了网络参数量, 使多领域图像翻译更有效; 增强了判别器分类性能, 不同模态图像的风格有所区别, 风格编码器和判别器的共用加强了判别器的能力, 提高了分类效率. Liu 等[115]提出了少镜头无监督图像到图像翻译(few-shot unsupervised image-to-image translation, FUNIT), 当前的算法虽取得了巨大的成功, 但在多类目标域中使用大量的特征数据会极大地限制数据的应用, 而 FUNIT 适用于少量、无监督的图像到图像转换算法, 具有实现小样本生成的能力. COCO-FUNIT[116]加入了条件风格编码器, 用来阻止与任务无关的外观信息至翻译图像中. 条件风格编码器通过调节输入内容图像帮助网络完成样式代码的变化, 在小样本数据下, 有效地减少了无关信息在网络翻译中的扰动.

5.1.3 基于解耦表示的多领域图像翻译算法

多模态无监督图像到图像的翻译(multimodal unsupervised image-to-image translation, MUNIT)[117]假设图像可以分解为域不变的内容代码和捕获特定属性的样式代码, 若需要将图像翻译为其他领域图像, 只需将内容代码和捕获特定属性的样式代码重新组合, 即可产生所需要的领域的图像. MUNIT 通俗易懂地解决了图像多领域的翻译问题, 但解耦方式较为粗糙. Hu 等[118]认为, 风格信息解耦后的代码是随机噪声, 因此会产生次优的多风格表示. 为实现更有效的多领域图像翻译, 其在原有的 MUNIT 的基础上加入了自注意力、跳跃连接等机制, 使网络更注重图像的全局和细节的信息, 实现了高质量的图像多领域翻译. Yu 等[119]建立多个编码器用于提取图像中领域、风格内容等信息, 使用先验信息与解耦代码作为特定损失, 以提高网络解耦效率, 取得了较为优秀的效果. 类似地, TransGaGa[120]引入几何先验损失和条件编码器的损失, 将图像空间分解为外观和几何空间的笛卡儿积, 以鼓励网络学习独立但互补的表示. 综上所述, 基于解耦表示的多领域图像翻译算法注重通过提升解耦内容信息来实现图像翻译.

5.2 多领域图像翻译的应用场景

5.2.1 人脸多领域翻译

模块化生成对抗网络(modular generative adversarial networks, ModularGAN)[121]认为, 在训练特定的映射中无法利用来自其他模态的数据, 因此其参考模块网络将 ModularGAN 设定为多个可组合的模块, 各模块具有不同的功能. 在训练数据时, 利用所有的数据训练所有模块; 在测试时, 根据特定的翻译任务构建特定的 GAN, 人脸图像的可区分性使网络生成多种颜色头发和多表情状态的图像, 进而实现人脸图像的多领域翻译.

Zhao 等[122]认为, 在非匹配图像翻译中, 使用

循环一致损失产生了严格的像素级约束，可以较好地还原源域图像的轮廓的信息．但在多领域图像翻译中，循环一致损失导致网络无法执行形状或者大对象的移除，如人脸图像中的墨镜．因此，Zhao 等[122]提出了 ACL-GAN，在网络中加入对抗一致损失，ACL 所代表的对抗一致损失的目的是最大化目标域和源域共享信息，即鼓励生成图像包含来自源域图像的更多特征，而不是像素级的一致性，较好地解决了现阶段存在的问题．

Karras 等[76]认为，在人脸图像翻译时重点需要关注人脸风格，因此，算法使用 AdaIN 模块、根据条件去改变 StyleGAN 的特征空间，以完成不同风格的人脸翻译；对每个特征图进行归一化后，可能会破坏特征间的信息，因此，需去除归一化以完成高质量的特征图像翻译．无锯齿生成对抗网络[123](alias-free generative adversarial networks, StyleGAN3)和大型多样化数据的样式生成对抗网络[124](scaling StyleGAN to large diverse datasets, StyleGAN-XL)在人脸图像多领域翻译上具有更高的精度和更快的速度，实现了人脸多领域的场景翻译．

上述方法分别解决了人脸的微观翻译和宏观翻译，为人脸多领域翻译提供了良好的思路．

5.2.2　服饰多领域翻译

随着互联网的普及和发展，在网络购物时需要对拍摄服饰进行相似检索．由于拍摄服饰时存在不同角度，以及人体因素的干扰，不能将拍摄图像直接作为数据进行匹配．因此，需要将拍摄图像翻译为计算机能识别的平面服装图像，即多对一图像翻译．Zhang 等[125]根据平面服饰的特性设计网络级联结构命名为 CascadeGAN，通过级联结构可以给平面图像提供更多的细节纹理，实现了不同模态的输入图像翻译为单一图像的任务．为了让生成的平面服饰更具有真实性，Triple-GAN[126]建立了生成器、判别器和分类器 3 个网络，可生成更真实的平面图像．其中，生成器用于生成目标域图像，分类器任务为预测生成图像的标签，判别器用于选择生成图像、真实平面图像和真实非平面图像．

虚拟试穿是一种将一份真实服装图像叠加至图像中目标人物身上的技术，即多对一翻译．Tsunashima 等[127]使用分离表示的图像翻译算法，分别从人物图像和服饰图像中提取人物和服饰特征，并交换衣服和服饰特征．最后使用领域生成器实现虚拟试穿．Xie 等[128]提出的路由空间自适应的可扩展的不配对虚拟试穿网络(towards scalable unpaired virtual try-on via patch-routed spatially-adaptive GAN, PASTA-GAN)认为，现阶段的虚拟试穿存在服饰纹理保留不清晰、服饰细节不逼真等情况．因此，PASTA-GAN 使用一个路由补丁模块保存服饰和纹理特征，其将服饰特征解耦至标准化的 patch 块，消除服饰的固有空间信息，再将标准化的 patch 块重构为符合人物姿势的图像．考虑到服饰扭曲等问题，PASTA-GAN 引入新的空间自适应残差模块，以引导生成器生成更逼真的细节．服饰图像多领域翻译的多对一和一对多的方法，对电商匹配商品和选择商品具有更好的效果．

6　图像翻译的质量评价

对于图像质量的判断，不同的方法都会存在一定的差异性[129-130]．因此，图像质量的度量着重对图像某一特点的评价．在目前的研究中，生成图像质量的评价主要分为主观评估和客观指标评估 2 大类．多位专家综合评估是主观评估的一种方式，而客观指标用于评价图像特定方面．

6.1　主观评估

主观评估方法需要人工判断图像是否存在瑕疵，即通过专家判断图像的真实程度，根据图像质量给出相应的分数．但主观评估也可能出现对生成图像质量的误判，如图 9 和表 1 所示，专家打分为 0~4 分，分数越高，表示图像翻译质量越高．对于相同的输入图像，输出图像 1 和输出图像 2 都属于高质量的图像翻译结果，大多数专家的主观评估结果为输出图像 2 优于输出图像 1；但在 FID 指标上，输出图像 1 优于输出图像 2．

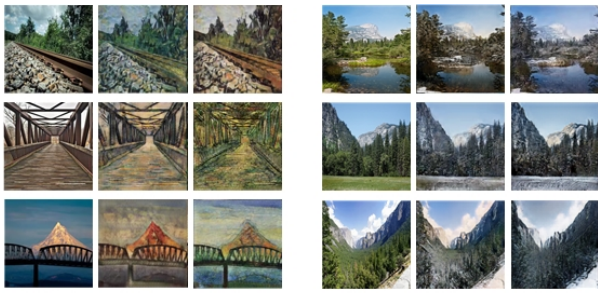


图 9　高质量图像评估差异

表 1　输出图像 1 和输出图像 2 得分均值

评价方法	输出图像 1	输出图像 2
FID	**51.64**	55.70
专家打分	3.45	**3.94**

注．粗体字表示最优值．

6.2 客观评估

6.2.1 全参考图像质量评价

全参考图像质量评价方法通过比较生成图像与对应的参考图像的相似度进行, 得到生成图像的质量评价指标.

(1) 结构相似度(structrual similarity index metric, SSIM)[131]. SSIM 从图像的组成角度出发, 将亮度、对比度和结构定义成主要信息, 其定义为

$$\mathrm{SSIM}=\frac{(2\overline{x}\cdot\overline{y}+C_1)(2\sigma_{xy}+C_2)}{(\overline{x}^2+\overline{y}+C_1)(\sigma_x^2+\sigma_y^2+C_2)}.$$

其中, x 和 y 分别表示图像 A 和 B; $\overline{x}$ 和 $\overline{y}$ 分别表示 x 和 y 的平均值; σ_x^2 和 σ_y^2 分别表示 x 和 y 的方差, σ_{xy} 表示 x 和 y 的协方差; $C_1=(k_1,L)^2$ 和 $C_2=(k_2,L)^2$ 均为常数; L 为像素的动态范围, k_1=0.01, k_2=0.03. SSIM 的取值范围为 0~1. 若输入 2 幅相同图像, 则 SSIM≡1.

(2) 峰值信噪比(peak signal to noise ration, PSNR). 其计算公式为

$$\mathrm{PSNR}=10\times\ln\frac{255}{\mathrm{MSE}}.$$

其中, 源域图像与生成图像的均方误差为

$$\mathrm{MSE}(x,y)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\big(x(i,j)-y(i,j)\big)^2.$$

其中, x 表示源域图像; y 为生成图像; H 和 W 分别表示图像的高度和宽度.

(3) 学习感知图像块相似度(learned perceptual image patch similarity, LPIPS)[132]. LPIPS 用于度量 2 幅图像之间的差别, 该度量标准学习生成图像到真实图像的反向映射, 强制生成器学习从假图像中重构真实图像的反向映射, 并优先处理它们之间的感知相似度. LPIPS 比传统方法更符合人类的感知情况, 其值越低, 表示 2 幅图像越相似; 反之, 图像的差异越大. LPIPS 计算公式为

$$d(x,y)=\sum_t\frac{1}{H_lW_l}\sum_{h,w}\left\|w_l\odot\left(x_{hw}^l-y_{hw}^l\right)\right\|_2^2.$$

其中, x 表示源域图像; y 表示生成图像; x_{hw}^l 与 y_{hw}^l 分别表示源域图像与生成图像第 l 层中的参数, x_{hw}^l与$y_{hw}^l\in\mathbb{R}^{H_l\times W_l\times C_l}$, $w_l\in\mathbb{R}^{C_l}$.

(4) 图像通用质量指标(universal quality index, UQI). 图像通用质量指标能够很容易地计算并且应用到各种图像处理的应用中, 它主要结合相关性损失、亮度失真、对比度失真 3 个方面计算. 计算公式为

$$Q=\frac{4\sigma_{xy}\overline{x}\cdot\overline{y}}{(\sigma_x^2+\sigma_y^2)\left[(\overline{x})^2+(\overline{y})^2\right]}.$$

其中, $\sigma_{xy},\sigma_x^2,\sigma_y^2$ 分别为

$$\sigma_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-x)(y_i-y),$$

$$\overline{x}=\frac{1}{N}\sum_{i=1}^{N}x_i,\ \overline{y}=\frac{1}{N}\sum_{i=1}^{N}y_i,$$

$$\sigma_x^2=\frac{1}{N-1}\sum_{i=1}^{N}(\overline{x}-x_i)^2,$$

$$\sigma_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{y})^2.$$

其中, x 表示源域图像; y 表示生成图像; x_i 和 y_i 分别表示 x 和 y 中的信号; $\overline{x}$与$\overline{y}$ 分别表示 x 与 y 中信号的均值.

(5) 视觉信息保真度(visual quality fidelity, VIF)[133]. VIF 评价复杂变换后图像的失真程度, 通过 2 种共同信息来评价图像的质量, 如图 10 所示.

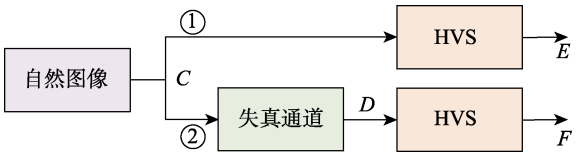


图 10 VIF 自然场景分析流程图

自然图像由通道①(无失真通道)经过 HVS 得到的输入图像与输出图像的共同信息——参考图像信息; 自然图像通过通道②(失真通道)得到输入和输出的共同信息——失真图像信息. VIF 即为失真图像信息与参考图像信息的比值. 具体提取信息推导公式参见文献[133].

(6) 特征相似度(feature similarity index mersure, FSIM). SSIM 是对生成图像与参考图像间结构相似度的评价, 而 FSIM[134]是对图像特定特征间相似度的评价. 通过提取相位一致性特征和梯度特征, 将它们进行融合得到最终的 FISM. 其中, 相位(phase congruency, PC)相似度的计算公式为

$$S_{\mathrm{PC}}=\frac{2\mathrm{PC}_1(x)\cdot\mathrm{PC}_2(x)+T_1}{\mathrm{PC}_1^2(x)+\mathrm{PC}_2^2(x)+T_1}.$$

梯度相似度计算公式为

$$S_G(x)=\frac{2G_1(x)\cdot G_2(x)+T_2}{G_1^2(x)+G_2^2(x)+T_2}.$$

相位和梯度融合相似度计算公式为

$$S_L=\left[S_{\mathrm{PC}}(x)\right]^{\alpha}\left[S_G(x)\right]^{\beta}.$$

其中，α 与 β 为超参数. 最终 FISM 计算公式为

$$\mathrm{FSIM}=\frac{\sum\limits_{x\in\Omega}S_L(x)\cdot\mathrm{PC}(x)}{\sum\limits_{x\in\Omega}\mathrm{PC}(x)}.$$

6.2.2　无参考图像质量评价

(1) 初始得分(inception score, IS)[135]. IS 是以 Inception 网络为模型的评价方法. Inception 网络是已在 ImageNet 数据集上训练好的分类模型，当输入生成图像 x'，网络会给出条件分布概率 $P(y|x')$. 图像质量越高，$P(y|x')$ 的熵也就越低，即分类器将较高的置信度分给了某一类，代表图像清晰度越高. $P(y|x')$ 熵的高低只能评判图像清晰度，但图像的多样性也是重要的指标，边缘分布 $P(y)$ 的熵越高代表图像多样性越多. IS 指标的计算公式为

$$\mathrm{IS}=\mathrm{e}^{E_{x\sim p_z}D_{\mathrm{KL}}\left(P(y|x')\|P(y)\right)}.$$

其中，E 表示计算期望；D_{KL} 表示计算 $P(y|x')$ 和 $P(y)$ 间的 KL 散度.

(2) Fréchet 距离(Fréchet inception distance, FID)[136]. 与 IS 相比，FID 提取 Inception 网络的特征提取层作为特征提取函数，采用特征提取函数计算真实分布 P_{data} 和生成分布 $P_{(z)}$ 间的分布均值和方差等信息，并据此来计算真实分布与生成分布间的距离. 其计算公式为

$$\mathrm{FID}(x,z)=\|\mu_x-\mu_z\|_2^2+\mathrm{tr}\left(\Sigma_x+\Sigma_z-2(\Sigma_x\Sigma_z)^{\frac{1}{2}}\right).$$

其中，μ_x 和 μ_z 分别表示真实分布和生成分布的均值；Σ_x 和 Σ_z 分别表示真实分布和生成分布的方差.

综上所述，FID 数值越小，代表生成分布越接近真实分布，则生成图像越接近真实图像.

7　结　语

图像翻译作为计算机视觉中十分重要的任务之一，已得到广泛的研究. 本文首先从任务的分类、数据集的分类、实现任务的算法等对任务存在的问题和应对策略进行了详细的讨论，其中详细介绍了 GAN 和 CGAN 原理. 其次，按照数据的类别将图像翻译任务定义为匹配图像翻译、非匹配图像翻译和多领域图像翻译 3 种，并分别介绍了它们的算法原理和具体应用. 最后，对目前工作中的生成图像的质量评估方法进行了详细的汇总和分析.

通过以上论述，基于 CGAN 完成图像翻译任务的方法主要改进包括生成器与判别器的神经网络结构、网络训练方式和网络损失函数，以及平衡数据域间差异.

虽然上述改进的模型在测试和实践中取得了一定的进展，但是图像翻译问题中还存在着一定的挑战，初步分为以下 3 个方面:

(1) 模式崩塌问题. 模式崩塌是指在网络训练中只能生成几种风格或者有限的模式. 在现实应用中体现为真实图像训练后的网络模型涵盖所有模式，但是同时生成一些没有意义或者现实中不可能出现的图像. 目前缓解此类问题的方法采用修改目标函数、改变训练方式，以及建立最优传输等方法等. 例如, MGGAN 利用现有 GAN 架构上的引导网络来诱导生成器学习数据分布的整体模式，将图像转换为学习的流形空间，缓解了模式崩塌问题.

(2) 模型的可解释性问题. 基于CGAN实现图像翻译任务展现出的良好性能仍归因于神经网络优秀的拟合能力. 翻译图像的成功与否，与网络结构的好坏有着强关联性，因此，弱化了理论研究对算法的支持. 模型缺少理论研究对算法的支持，使网络在应对具体问题时只能人为或使用优化算法寻找良好的网络架构以及较为合适的模型，导致提出算法的性能未能令人信服.

(3) 少样本学习问题. 数据集的大小影响网络模型训练的需求，虽然部分数据集采集较容易获取，但存在特殊图像获取难度较高、成本高的问题. 因此，在图像翻译领域需提出基于先验知识和基于潜在、共享空间的少样本模型. 目前存在解决类似问题的模型[115,137]，其中 Liu 等[115]认为，网络能够从少量示例中提取新颖对象的本质并进行归纳，期望解决少样本学习问题；但存在领域泛化性能不高、生成图像质量低等问题，这是未来需研究的方向之一.

参考文献(References):

[1] Resales R, Achan K, Frey B J. Unsupervised image translation[C] //Proceedings of the Ninth IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2003: 472-478

[2] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2414-2423

[3] Taigman Y, Polyak A, Wolf L. Unsupervised cross-domain image generation[OL]. [2022-06-27]. https://doi.org/10.48550/

arXiv.1611.02200

[4] Goodfellow I J, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680

[5] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C] //Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2012: 1097-1105

[6] Zhu Min, Ming Zhangqiang, Yan Jianrong, *et al*. A survey on generative adversarial network based person re-identification method[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(2): 163-179(in Chinese)
(朱敏, 明章强, 闫建荣, 等. 基于生成对抗网络的行人重识别方法研究综述[J]. 计算机辅助设计与图形学学报, 2022, 34(2): 163-179)

[7] Fan Xinnan, Yang Xin, Shi Pengfei, *et al*. Underwater image enhancement based on feature fusion generative adversaral networks[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(2): 264-272(in Chinese)
(范新南, 杨鑫, 史朋飞, 等. 特征融合生成对抗网络的水下图像增强[J]. 计算机辅助设计与图形学学报, 2022, 34(2): 264-272)

[8] Zhou Xun, Yang Degang, Feng Ji. Single image super-resolution reconstruction algorithm with generative adversarial network based on multi-size convolution and Laplacian filtering[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(10): 1504-1513(in Chinese)
(周勋, 杨德刚, 冯骥. 基于多感受野拉普拉斯生成对抗网络的单幅图像超分辨率重建算法[J]. 计算机辅助设计与图形学学报, 2021, 33(10): 1504-1513)

[9] Mirza M, Osindero S. Conditional generative adversarial nets[OL]. [2022-06-27]. https://arxiv.org/abs/1411.1784

[10] Wu J W, Yin F, Zhang Y M, *et al*. Image-to-markup generation via paired adversarial learning[C] //Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Heidelberg: Springer, 2018: 18-34

[11] Huang J L, Liao J, Kwong S. Unsupervised image-to-image translation via pre-trained stylegan2 network[J]. IEEE Transactions on Multimedia, 2021, 24: 1435-1448

[12] Lin C T, Wu Y Y, Hsu P H, *et al*. Multimodal structure-consistent image-to-image translation[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 11490-11498

[13] Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[OL]. [2022-06-27]. https://arxiv.org/abs/1708.07747

[14] Tyleček R, Šára R. Spatial pattern templates for recognition of objects with regular structure[C] //Proceedings of the 35th German Conference on Pattern Recognition. Heidelberg: Springer, 2013: 364-374

[15] Cordts M, Omran M, Ramos S, *et al*. The Cityscapes dataset for semantic urban scene understanding[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 3213-3223

[16] Kim T, Cha M, Kim H, *et al*. Learning to discover cross-domain relations with generative adversarial networks[C] //Proceedings of the 34th International Conference on Machine Learning. New York: ACM, 2017: 1857-1865

[17] Chen J, Liu G, Chen X. AnimeGAN: a novel lightweight GAN for photo animation[C] //Proceedings of 11th International Symposium on Intelligence Computation and Applications. Heidelberg: Springer, 2019: 242-256

[18] Liu Z W, Luo P, Wang X G, *et al*. Deep learning face attributes in the wild[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 3730-3738

[19] Żurański A M, Martinez Alvarado J I, Shields B J, *et al*. Predicting reaction yields via supervised learning[J]. Accounts of Chemical Research, 2021, 54(8): 1856-1865

[20] Burkart N, Huber M F. A survey on the explainability of supervised machine learning[J]. Journal of Artificial Intelligence Research, 2021, 70: 245-317

[21] Sun H, Zheng X T, Lu X Q. A supervised segmentation network for hyperspectral image classification[J]. IEEE Transactions on Image Processing, 2021, 30: 2810-2825

[22] Sharma S, Singh G, Sharma M. A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in humans[J]. Computers in Biology and Medicine, 2021, 134: Article No.104450

[23] Kapucu C, Cubukcu M. A supervised ensemble learning method for fault diagnosis in photovoltaic strings[J]. Energy, 2021, 227: Article No.120463

[24] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C] //Proceedings of the 34th International Conference on Machine Learning. New York: ACM, 2017: 214-223

[25] Gulrajani I, Ahmed F, Arjovsky M, *et al*. Improved training of Wasserstein GANs[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 5769-5779

[26] Isola P, Zhu J Y, Zhou T H, *et al*. Image-to-image translation with conditional adversarial networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5967-5976

[27] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[OL]. [2022-06-27]. https://arxiv.org/abs/1511.06434

[28] Wang T C, Liu M Y, Zhu J Y, *et al*. High-resolution image synthesis and semantic manipulation with conditional GANs[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 8798-8807

[29] Wang C Y, Xu C, Wang C H, *et al*. Perceptual adversarial networks for image-to-image transformation[J]. IEEE Transactions on Image Processing, 2018, 27(8): 4066-4079

[30] Wang C, Zheng H Y, Yu Z B, *et al*. Discriminative region proposal adversarial networks for high-quality image-to-image translation[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 796-812

[31] Zhu J Y, Zhang R, Pathak D, *et al*. Toward multimodal image-to-image translation[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems.

Red Hook: Curran Associates Inc., 2017: 465-476

[32] Bao J M, Chen D, Wen F, *et al*. CVAE-GAN: fine-grained image generation through asymmetric training[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2764-2773

[33] Yang J W, Kannan A, Batra D, *et al*. LR-GAN: layered recursive generative adversarial networks for image generation[OL]. [2022-06-27]. https://arxiv.org/abs/1703.01560

[34] Wang X Y, Yan H P, Huo C L, *et al*. Enhancing Pix2Pix for remote sensing image classification[C] //Proceedings of the 24th International Conference on Pattern Recognition (ICPR). Los Alamitos: IEEE Computer Society Press, 2018: 2332-2336

[35] Li H A, Zhang M, Yu Z H, *et al*. An improved pix2pix model based on Gabor filter for robust color image rendering[J]. Mathematical Biosciences and Engineering, 2022, 19(1): 86-101

[36] Zuo Z C, Li Y X. A SAR-to-optical image translation method based on PIX2PIX[C] //Proceedings of the IEEE International Geoscience and Remote Sensing Symposium IGARSS. Los Alamitos: IEEE Computer Society Press, 2021: 3026-3029

[37] Balachandran V, Sarath S. A novel approach to detect unmanned aerial vehicle using Pix2Pix generative adversarial network[C] //Proceedings of the Second International Conference on Artificial Intelligence and Smart Energy (ICAIS). Los Alamitos: IEEE Computer Society Press, 2022: 1368-1373

[38] Christovam L E, Shimabukuro M H, Galo M L B T, *et al*. Pix2pix conditional generative adversarial network with MLP loss function for cloud removal in a cropland time series[J]. Remote Sensing, 2021, 14(1): Article No.144

[39] Berardi G, Salti S, Di Stefano L. SketchyDepth: from scene sketches to RGB-D images[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 2414-2423

[40] Fan L W, Zhang F, Fan H, *et al*. Brief review of image denoising techniques[J]. Visual Computing for Industry, Biomedicine, and Art, 2019, 2(1): Article No.7

[41] Chen Z L, Zeng Z Y, Shen H L, *et al*. DN-GAN: denoising generative adversarial networks for speckle noise reduction in optical coherence tomography images[J]. Biomedical Signal Processing and Control, 2020, 55: Article No.101632

[42] Adam S P, Alexandropoulos S A N, Pardalos P M, *et al*. No free lunch theorem: a review[M] //Demetriou I C, Pardalos P M. Approximation and Optimization. Heidelberg: Springer, 2019: 57-82

[43] Hesamian M H, Jia W J, He X J, *et al*. Deep learning techniques for medical image segmentation: achievements and challenges[J]. Journal of Digital Imaging, 2019, 32(4): 582-596

[44] Chen X, Williams B M, Vallabhaneni S R, *et al*. Learning active contour models for medical image segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 11624-11632

[45] He K M, Sun J, Tang X O. Single image haze removal using dark channel prior[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(12): 2341-2353

[46] Jin Xianli, Zhang Wei, Liu Linfeng. Image defogging algorithm based on guided filtering and adaptive tolerance[J]. Journal on Communications, 2020, 41(5): 27-36(in Chinese)
(金仙力，张威，刘林峰. 基于引导滤波和自适应容差的图像去雾算法[J]. 通信学报, 2020, 41(5): 27-36)

[47] Qu Y Y, Chen Y Z, Huang J Y, *et al*. Enhanced pix2pix dehazing network[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 8152-8160

[48] Bu Q R, Luo J, Ma K, *et al*. An enhanced pix2pix dehazing network with guided filter layer[J]. Applied Sciences, 2020, 10(17): Article No.5898

[49] Li S Y, Lin J, Yang X, *et al*. BPFD-Net: enhanced dehazing model based on Pix2pix framework for single image[J]. Machine Vision and Applications, 2021, 32(6): Article No.124

[50] Jiao L B, Hu C M, Huo L Z, *et al*. Guided-Pix2Pix: end-to-end inference and refinement network for image dehazing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 3052-3069

[51] Tu Hangyao, Wang Wanliang, Chen Jiacheng, *et al*. Dehazing algorithm combined with atmospheric scattering model based on generative adversarial network[J]. Journal of Zhejiang University (Engineering Science), 2022, 56(2): 225-235(in Chinese)
(屠杭垚，王万良，陈嘉诚，等. 结合大气散射模型的生成对抗网络去雾算法[J]. 浙江大学学报(工学版)，2022，56(2): 225-235)

[52] Heo H, Hwang Y. Automatic sketch colorization using DCGAN[C] //Proceedings of the 18th International Conference on Control, Automation and Systems (ICCAS). Los Alamitos: IEEE Computer Society Press, 2018: 1316-1318

[53] Kim H, Jhoo H Y, Park E, *et al*. Tag2pix: line art colorization using text tag with SECat and changing loss[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 9056-9065

[54] Zou C Q, Mo H R, Gao C Y, *et al*. Language-based colorization of scene sketches[J]. ACM Transactions on Graphics, 2019, 38(6): Article No.233

[55] Zhang Q, Wang B, Wen W, *et al*. Line art correlation matching feature transfer network for automatic animation colorization[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 3871-3880

[56] Qi X Q, Sun M Y, Wang W N, *et al*. Face sketch synthesis via semantic-driven generative adversarial network[C] //Proceedings of the IEEE International Joint Conference on Biometrics (IJCB). Los Alamitos: IEEE Computer Society Press, 2021: 1-8

[57] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010

[58] Zhu M R, Liang C C, Wang N N, *et al*. A sketch-transformer network for face photo-sketch synthesis[C] //Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. Freiburg: IJCAI, 2021: 1352-1358

[59] Li Hong'an, Zheng Qiaoxue, Zhang Jing, *et al*. Pix2Pix-based grayscale image coloring method[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(6): 929-938(in Chinese)

(李洪安, 郑峭雪, 张婧, 等. 结合 Pix2Pix 生成对抗网络的灰度图像着色方法[J]. 计算机辅助设计与图形学学报, 2021, 33(6): 929-938)

[60] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241

[61] Chen Y, Lai Y K, Liu Y J. CartoonGAN: generative adversarial networks for photo cartoonization[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 9465-9474

[62] Liu B C, Song K P, Zhu Y Z, *et al*. Sketch-to-art: synthesizing stylized art images from sketches[C] //Proceedings of the 15th Asian Conference on Computer Vision. Heidelberg: Springer, 2020

[63] Mehrotra R, Namuduri K R, Ranganathan N. Gabor filter-based edge detection[J]. Pattern Recognition, 1992, 25(12): 1479-1494

[64] Ben-Cohen A, Klang E, Raskin S P, *et al.* Virtual PET images from CT data using deep convolutional networks: initial results[C] //Proceedings of the Second International Workshop on Simulation and Synthesis in Medical Imaging. Heidelberg: Springer, 2017: 49-57

[65] Hussein R, Zhao MY, Shin D, *et al.* Multi-task deep learning for cerebrovascular disease classification and MRI-to-PET translation[C] //Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Los Alamitos: IEEE Computer Society Press, 2022

[66] Tsuda H, Hotta K. Cell image segmentation by integrating pix2pixs for each class[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2019: 1065-1073

[67] Sato M, Hotta K, Imanishi A, *et al*. Segmentation of cell membrane and nucleus by improving Pix2pix[C] //Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies. Funchal: SciTePress, 2018: 216-220

[68] Zhao Z, Wang H X, Zhang Y Z, *et al*. A coarse-to-fine data generation method for 2D and 3D cell nucleus segmentation[C] //Proceedings of the IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). Los Alamitos: IEEE Computer Society Press, 2020: 41-46

[69] Armanious K, Jiang C M, Fischer M, *et al*. MedGAN: medical image translation using GANs[J]. Computerized Medical Imaging and Graphics, 2020, 79: Article No.101684

[70] Armanious K, Kumar V, Abdulatif S, *et al*. ipA-MedGAN: inpainting of arbitrary regions in medical imaging[C] //Proceedings of the IEEE International Conference on Image Processing (ICIP). Los Alamitos: IEEE Computer Society Press, 2020: 3005-3009

[71] Upadhyay U, Chen Y B, Hepp T, *et al*. Uncertainty-guided progressive GANs for medical image translation[C] //Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2021: 614-624

[72] Zhu J Y, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2242-2251

[73] Yi Z L, Zhang H, Tan P, *et al*. DualGAN: unsupervised dual learning for image-to-image translation[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2868-2876

[74] Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 694-711

[75] Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks[OL]. [2022-06-27]. https://arxiv.org/abs/1701.04862

[76] Karras T, Laine S, Aittala M, *et al*. Analyzing and improving the image quality of styleGAN[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8107-8116

[77] Cherian A, Sullivan A. Sem-GAN: semantically-consistent image-to-image translation[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Los Alamitos: IEEE Computer Society Press, 2019: 1797-1806

[78] Gao X Y, Shi F, Shen D G, *et al*. Task-induced pyramid and attention GAN for multimodal brain image imputation and classification in Alzheimer's disease[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(1): 36-43

[79] Galassi A, Lippi M, Torroni P. Attention in natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(10): 4291-4308

[80] Fu J, Liu J, Tian H J, *et al*. Dual attention network for scene segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3141-3149

[81] Zhang H, Goodfellow I, Metaxas D, *et al*. Self-attention generative adversarial networks[OL]. [2022-06-27]. https://arxiv.org/abs/1805.08318

[82] Tang H, Liu H, Xu D, *et al*. AttentionGAN: unpaired image-to-image translation using attention-guided generative adversarial networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(4): 1972-1987

[83] Emami H, Aliabadi M M, Dong M, *et al*. SPA-GAN: spatial attention GAN for image-to-image translation[J]. IEEE Transactions on Multimedia, 2021, 23: 391-401

[84] Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017

[85] Lee H Y, Tseng H Y, Huang J B, *et al*. Diverse image-to-image translation via disentangled representations[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 36-52

[86] Li Y, Tang S, Zhang R, *et al*. Asymmetric GAN for unpaired image-to-image translation[J]. IEEE Transactions on Image Processing, 2019, 28(12): 5881-5896

[87] Dou H, Chen C, Hu X Y, *et al*. Asymmetric CycleGan for un-

paired NIR-to-RGB face image translation[C] //Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Los Alamitos: IEEE Computer Society Press, 2019: 1757-1761

[88] Sheng Y W. Asymmetric CycleGAN for unpaired image-to-image translation based on dual attention module[C] //2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST). Los Alamitos: IEEE Computer Society Press, 2021: 726-730

[89] Hoffman J, Tzeng E, Park T, *et al*. Cycada: cycle-consistent adversarial domain adaptation[OL]. [2022-06-27]. https://arxiv.org/abs/1711.03213

[90] Pizzati F, De Charette R, Zaccaria M, *et al*. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2020: 2979-2987

[91] Tu H Y, Wang W L, Chen J C, *et al*. Unpaired image-to-image translation with improved two-dimensional feature[J]. Multimedia Tools and Applications, 2022, 81(30): 43851-43872

[92] Xu X G, Chen Y C, Tao X, *et al*. Text-guided human image manipulation via image-text shared space[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6486-6500

[93] Tumanyan N, Geyer M, Bagon S, *et al*. Plug-and-play diffusion features for text-driven image-to-image translation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 1921-1930

[94] Li B W, Qi X J, Lukasiewicz T, *et al*. ManiGAN: text-guided image manipulation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 7877-7886

[95] Fu X H. Digital image art style transfer algorithm based on CycleGAN[J]. Computational Intelligence and Neuroscience, 2022, 2022: Article No.6075398.

[96] Lin M X, Deng Y Y, Tang F, *et al*. Multi-attribute guided painting generation[C] //Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Los Alamitos: IEEE Computer Society Press, 2020: 400-403

[97] Park D Y, Lee K H. Arbitrary style transfer with style-attentional networks[C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Los Alamitos: IEEE Computer Society Press, 2019: 5880-5888

[98] Sanakoyeu A, Kotovenko D, Lang S, *et al*. A style-aware content loss for real-time HD style transfer[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 715-731

[99] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2022-06-27]. https://arxiv.org/abs/1409.1556

[100] Dong Y S, Tan W, Tao D C, *et al*. CartoonLossGAN: learning surface and coloring of images for cartoonization[J]. IEEE Transactions on Image Processing, 2022, 31: 485-498

[101] Wang Z H, Chen J, Hoi S C H. Deep learning for image super-resolution: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3365-3387

[102] Zhang K, Van Gool L, Timofte R. Deep unfolding network for image super-resolution[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 3214-3223

[103] Ma C, Rao Y M, Cheng Y A, *et al*. Structure-preserving super resolution with gradient guidance[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 7766-7775

[104] Ji X Z, Cao Y, Tai Y, *et al*. Real-world super-resolution via kernel estimation and noise injection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2020: 1914-1923

[105] Cai J R, Zeng H, Yong H W, *et al*. Toward real-world single image super-resolution: a new benchmark and a new model[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 3086-3095

[106] Lugmayr A, Danelljan M, Timofte R. Unsupervised learning for real-world super-resolution[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Los Alamitos: IEEE Computer Society Press, 2019: 3408-3416

[107] Bulat A, Yang J, Tzimiropoulos G. To learn image super-resolution, use a GAN to learn how to do image degradation first[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 187-202

[108] Kim G, Park J, Lee K, *et al*. Unsupervised real-world super resolution with cycle generative adversarial network and domain discriminator[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2020: 1862-1871

[109] Wu S X, Dong C, Qiao Y. Blind image restoration based on cycle-consistent network[J]. IEEE Transactions on Multimedia, 2022, 25: 1111-1124

[110] Wei Y X, Gu S H, Li Y W, *et al*. Unsupervised real-world image super resolution via domain-distance aware training[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 13380-13389

[111] Choi Y, Choi M, Kim M, *et al*. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 8789-8797

[112] Choi Y, Uh Y, Yoo J, *et al*. StarGAN v2: diverse image synthesis for multiple domains[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8185-8194

[113] Zhang X, Yang M N, Chai H Y. EStarGAN: enhanced StarGAN for multi-style high-definition image translation[C] //Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC). Los Alamitos: IEEE Computer Society Press, 2021: 1737-1744

[114] Sun Y B, Lu Y, Lu H W, *et al*. Multimodal unsupervised image-to-image translation without independent style encoder[C]

//Proceedings of the 28th International Conference on Multimedia Modeling. Heidelberg: Springer, 2022: 624-636

[115] Liu M Y, Huang X, Mallya A, *et al*. Few-shot unsupervised image-to-image translation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 10550-10559

[116] Saito K, Saenko K, Liu M Y. COCO-FUNIT: few-shot unsupervised image translation with a content conditioned style encoder[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 382-398

[117] Huang X, Liu M Y, Belongie S, *et al*. Multimodal unsupervised image-to-image translation[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 172-189

[118] Hu J F, Wu G, Wang H, *et al*. Latent Style: multi-style image transfer via latent style coding and skip connection[J]. Signal, Image and Video Processing, 2022, 16(2): 359-368

[119] Yu X M, Chen Y Q, Liu S, *et al*. Multi-mapping image-to- image translation via learning disentanglement[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2019: Article No.269

[120] Wu W, Cao K D, Li C, *et al*. TransGaGa: Geometry-aware unsupervised image-to-image translation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 8004-8013

[121] Zhao B, Chang B, Jie Z Q, *et al*. Modular generative adversarial networks[C] //Proceedings of the 15th European Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2018: 157-173

[122] Zhao Y H, Wu R H, Dong H. Unpaired image-to-image translation using adversarial consistency loss[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 800-815

[123] Karras T, Aittala M, Laine S, *et al*. Alias-free generative adversarial networks[OL]. [2022-06-27]. https://arxiv.org/abs/2106.12423

[124] Sauer A, Schwarz K, Geiger A. StyleGAN-XL: scaling stylegan to large diverse datasets[C] //Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings. New York: ACM Press, 2022: Article No.49

[125] Zhang H J, Sun Y F, Liu L L, *et al*. CascadeGAN: a category-supervised cascading generative adversarial network for clothes translation from the human body to tiled images[J]. Neurocomputing, 2020, 382: 148-161

[126] Fang H, Deng W H, Zhong Y Y, *et al*. Triple-GAN: progressive face aging with triple translation loss[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2020: 3500-3509

[127] Tsunashima H, Arase K, Lam A, *et al*. UVIRT—unsupervised virtual try-on using disentangled clothing and person features[J]. Sensors, 2020, 20(19): Article No.5647

[128] Xie Z Y, Huang Z Y, Zhao F W, *et al*. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive GAN [OL]. [2022-06-27]. https://arxiv.org/abs/2111.10544

[129] Chu Jiang, Chen Qiang, Yang Xichen. Review on full reference image quality assessment algorithms[J]. Application Research of Computers, 2014, 31(1): 13-22(in Chinese)
(褚江, 陈强, 杨曦晨. 全参考图像质量评价综述[J]. 计算机应用研究, 2014, 31(1): 13-22)

[130] Wang Zhiming. Review of no-reference image quality assessment[J]. Acta Automatica Sinica, 2015, 41(6): 1062-1079(in Chinese)
(王志明. 无参考图像质量评价综述[J]. 自动化学报, 2015, 41(6): 1062-1079)

[131] Wang L T, Hoover N E, Porter E H, *et al*. SSIM: a software levelized compiled-code simulator[C] //Proceedings of the 24th ACM/IEEE Design Automation Conference. Los Alamitos: IEEE Computer Society Press, 1987: 2-8.

[132] Zhang R, Isola P, Efros A A, *et al*. The unreasonable effectiveness of deep features as a perceptual metric[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 586-595

[133] Sheikh H R, Bovik A C. Image information and visual quality[J]. IEEE Transactions on Image Processing, 2006, 15(2): 430-444

[134] Zhang L, Zhang L, Mou X Q, *et al*. FSIM: a feature similarity index for image quality assessment[J]. IEEE Transactions on Image Processing, 2011, 20(8): 2378-2386

[135] Barratt S, Sharma R. A note on the inception score[OL]. [2022-06-27]. https://arxiv.org/abs/1801.01973

[136] Dowson D C, Landau B V. The Fréchet distance between multivariate normal distributions[J]. Journal of Multivariate Analysis, 1982, 12(3): 450-455

[137] Lin J X, Xia Y C, Liu S, *et al*. ZstGAN: an adversarial approach for unsupervised zero-shot image-to-image translation[J]. Neurocomputing, 2021, 461: 327-335