计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(4):283−292 [doi: 10.15888/j.cnki.csa.009059]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合生成对抗网络及多角度注意力的图像翻译模型①

杨百冰, 陈泯融, 叶勇森

(华南师范大学 计算机学院, 广州 510631)
通信作者: 陈泯融, E-mail: chenminrong@scnu.edu.cn

摘　要: 本文提出一个新的无监督图像翻译模型, 该模型结合了生成对抗网络和多角度注意力, 称为 MAGAN. 多角度注意力引导翻译模型将注意力集中在不同域间最具有判别性的区域. 与现存的注意力方法不同的是, 空间激活映射一方面捕获通道间的依赖, 减少翻译图像的特征扭曲; 另一方面决定网络对最具判别性区域的空间位置的关注程度, 使翻译的图像更具有目标域风格. 在空间激活映射的基础上, 结合类激活映射, 可以获得图像的全局语义信息. 此外, 根据空间激活程度对图像特征信息的影响, 设计不同的注意力结构分别训练生成器和判别器. 实验结果表明, 本文模型在 selfie2anime、cat2dog、horse2zebra 和 vangogh2photo 这 4 个数据集上的 KID 分数分别达到 9.48、6.32、6.42 和 4.28, 性能优于大部分主流模型, 并且与基线模型 UGATIT 相比, 在 selfie2anime、cat2dog 和 horse2zebra 这 3 个数据集上的距离值分别减少了 2.13、0.75 和 0.64, 具有明显的性能优势.

关键词: 生成对抗网络; 图像翻译; 图像风格迁移; 多角度注意力; 无监督网络; 图像生成

引用格式: 杨百冰,陈泯融,叶勇森.结合生成对抗网络及多角度注意力的图像翻译模型.计算机系统应用,2023,32(4):283–292. http://www.c-s-a.org.cn/1003-3254/9059.html

Image-to-image Translation Model Combining GAN and Multi-angle Attention

YANG Bai-Bing, CHEN Min-Rong, YE Yong-Sen

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: This study proposes a new unsupervised image-to-image translation model that combines generative adversarial networks (GAN) and multi-angle attention, and it is called MAGAN for short. The multi-angle attention guides the translation model to focus its attention on the most discriminative regions among different domains. Unlike the existing attention-based methods, spatial activation mapping (SAM) not only captures the dependencies among channels to reduce the feature distortion of the translated image but also determines the extent to which the network focuses on the spatial location of the most discriminative regions so that the translated image is more in the style of the target domain. On the basis of SAM, the global semantic information of the image can be obtained by class activation mapping (CAM). In addition, different attention structures are designed to train the generator and the discriminator, respectively, according to the influence of spatial activation degree on the feature information of the image. Experimental results show that the model proposed in this study outperforms most mainstream models with kernel inception distance (KID) scores of 9.48, 6.32, 6.42, and 4.28 on the four datasets selfie2anime, cat2dog, horse2zebra, and vangogh2photo, respectively. Moreover, compared with the baseline model, namely, unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation (UGATIT), the proposed model has significant performance advantages in that it reduces the distances on the selfie2anime, cat2dog, and horse2zebra datasets by 2.13, 0.75, and 0.64, respectively.

① 基金项目: 国家自然科学基金 (61872153, 61972288)
收稿时间: 2022-09-19; 修改时间: 2022-10-19; 采用时间: 2022-11-16; csa 在线出版时间: 2023-03-01
CNKI 网络首发时间: 2023-03-02

Key words: generative adversarial network (GAN); image-to-image translation; image style transfer; multi-angle attention (MA); unsupervised network; image generation

图像翻译旨在将图像从一个域转换到另一个域. 随着生成对抗网络 (generative adversarial network, GAN)[1] 的兴起, 越来越多的研究将其应用于两个域之间的图像翻译, 并取得了较好的效果. 它们可以在差异较小的域 (如马和斑马) 之间进行有效翻译, 却很难在差异较大的域 (如人脸和动画) 之间高质量地翻译图像. 例如, 用于多域图像翻译的 StarGAN v2[2] 只需要映射纹理和颜色样式就可以轻松翻译面部风格 (如面部表情、肤色和性别). 然而, StarGAN v2 在两个差距较大域之间的图像翻译任务上表现一般. 一个合理的解释是这些翻译任务中存在复杂的外观变化: 即除纹理变换之外, 网络还需要夸大图像的局部结构, 将局部形状转化为目标域的相应样式.

注意力机制促使图像翻译网络在翻译过程中更加关注影响判断的区域, 因此一些研究[3–5] 将注意力机制集成到模型中. ContrastGAN[3] 通过调整对象实例的掩码实现对完整图像的注意力语义操作. UGATIT[5] 将 CAM[6] 作为辅助分类器来获取注意力图, 引导模型专注于类别判断区域. 然而, 这种注意力方法忽略了像素之间的关联, 从而导致生成的图像出现明显的特征失真, 如图 1 所示.

图 1 UGATIT 生成的人脸扭曲图像

针对上述问题, 本文提出了一种用于图像翻译任务的多角度注意力 (multi-angle attention, MA), 从多个角度考虑了注意力机制的聚焦能力. 与 UGATIT 基于通道维度提取图像特征的方法不同, 本文否定了特征通道之间不相关的假设, 并强调隐藏激活神经元的权重与目标对象的空间相关性. 通过比较同一空间位置神经元的激活情况发现, 激活程度越高, 对最具判别力的区域分配的权重越大, 即网络对敏感区域的关注度越高. 本文还探索出相同的结构接受不同程度的空间激活, 性能存在明显差异. 对于生成器来说, 主要任务是捕获两个域之间最具判别性的区域, 必须更高程度地激活最大视差区域的特征, 以便获得足够的关注. 但是判别器需要从整体上判断生成图像的真实性, 而不是只关注对象的局部区域, 因此需要对判别对象的所有特征信息使用相同程度的激活. 为了使模型翻译出来的图像更多地具有目标域图像风格, 同时保留原始图像的内容, 本文采用 VGG19 作为特征提取网络, 以提取输入图像和生成图像的深度感知特征, 同时引入风格损失函数和内容重建损失函数, 并在风格损失函数中, 计算所提取特征的 Gram 矩阵[7] 来衡量特征间的关系. 最终通过定性和定量分析证明了本文设计的模型能够在多个数据集上有效地翻译图像.

1 相关工作

1.1 生成对抗网络

生成对抗网络 (GAN) 凭借其巧妙的网络结构和损失函数, 在各种图像处理任务中取得了卓越的成绩, 包括图像生成[8–11], 图像翻译[12–15], 多模态图像合成[16,17] 等. GAN 由生成器和判别器两个部分组成, 二者对抗训练, 将生成图像的分布逼近到与其对应的真实图像的分布. 为了改善生成图像的质量, 出现了一系列 GAN 的变体. 一方面, StyleGAN[9] 能够控制所生成图像的高层次属性. 其升级版 StyleGAN v2 修复了 StyleGAN 生成图像中的特征伪影, 从而提高了生成图像的质量. 另一方面, StarGAN[18] 超越了一对一翻译的限制, 以统一框架实现了一对多的转换.

1.2 图像翻译

图像翻译任务使用的数据集包括配对数据集和非配对数据集. 对于配对数据集, pix2pix[19] 通过对抗损失学习两个域之间的映射. 之后 Wang 等人[20] 提出它的改进版, 用来完成高分辨率图像的翻译. 由于数据样本的配对是一项非常繁琐的工作, 因此使用非配对数据集训练网络的方式愈加受欢迎. CycleGAN[21] 提出循环一致性损失, 首次实现以无监督的方式训练图像翻译网络. 为了实现多模态图像翻译, MUNIT[17] 采用 AdaIN[22] 来合并内容和样式的编码.

1.3 注意力

注意力可以促进网络感知影响性能的重要特征信

息. SAGAN[23] 将自注意力机制引入到图像生成网络, 提高了生成图像的质量. UGATIT 使用 CAM 作为辅助分类器, 帮助图像翻译网络区分不同域图像的判别性区域. CAM 通过全局平均池化操作, 使网络在分类训练过程中无需任何边界框标注即可完成目标定位. 为了更好地理解定位信息, CAM 将输入图像的预测类别分数可视化, 突出显示网络捕获到的目标. 虽然该方法对完整物体的识别有积极的效果, 但它是基于每个通道提取特征, 并假设通道之间不相关. 此外, Zagoruyko 等人[24] 对决定网络判断的输入空间区域进行了编码, 该方法将网络的注意力集中在某一层次的特征上, 但未能成功定位完整的对象.

2 网络结构

本文提出了结合生成对抗网络及多角度注意力的图像翻译模型 (MAGAN), 其目标是训练生成器来学习具有未配对数据的两个域之间的相互映射. 具体来说, 设 X 和 Y 是两个不同的域, 给定$x \in X$, 学习一个映射函数将 x 转换到域 Y, 即: $G_Y(x) \to y'$. 类似地, 给定$y \in Y$, 学习一个逆向映射函数$G_X(y) \to x'$, 用于将转 y 换到域 X. 考虑到两个不同域之间的风格差异, 使用两个判别器D_X和D_Y分别判断给定图像是对应域下的真实图像还是生成的假图像. 本文只解释从域 X 到域 Y 的转换 (见图 2), 反之亦然.

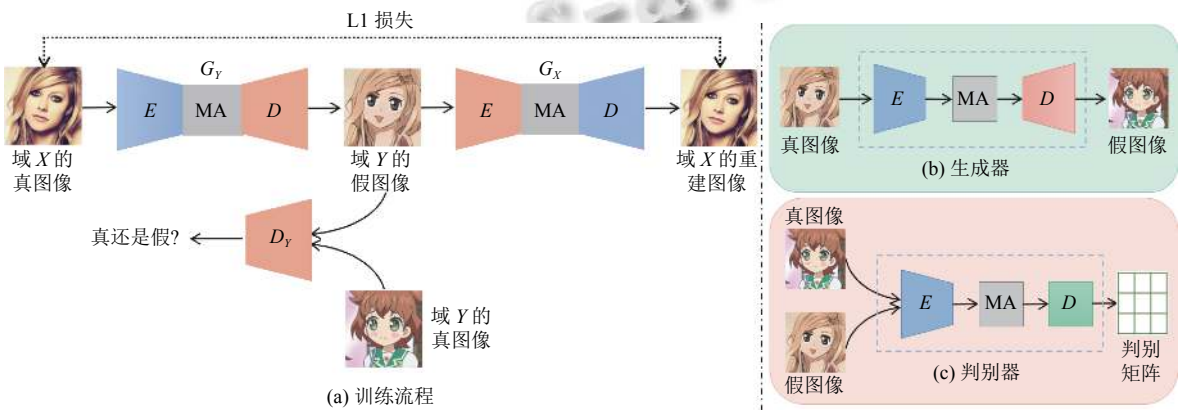


图 2 完整的网络架构

2.1 生成器

如图 2(b) 所示, 生成器由编码器 E、解码器 D 和多角度注意力 MA 组成. 编码器提取输入图像的特征信息; 解码器从样式代码构造图像; 多角度注意力预测图像的类别概率并获得注意力特征图.

编码器 E 由 3 个卷积层和 4 个残差层组成. 它的输入是源图像 x, 输出 x 的特征图θ. 编码器的功能表示为: $\theta^n = E^n(x)$, 其中θ^n是编码器输出的第 n 个特征图 $(1 \leqslant n \leqslant C)$, C 是编码器的输出通道.

不仅同一通道的特征具有特定的关系, 而且不同通道之间的特征也存在密切的相关性. 忽略特征的空间相关性是翻译图像中存在特征扭曲的关键原因. 为了解决上述问题, 本文设计了一个多角度注意力Γ_G, 其对特征图的详细处理过程如图 3 所示. 该过程由 3 部分组成: 全局最大池化 (GMP)、全局平均池化 (GAP) 和空间激活映射 (SAM), 分别用α、β和λ表示. 全局最大池化用于捕获前景对象的边缘特征, 即$\alpha^n = \max\{\theta^{n_{i,j}}\}$. 全局平均池化用于对物体进行模糊定位: $\beta^n = \dfrac{\sum_{i,j}\theta^{n_{i,j}}}{i \cdot j}$, 其中$\theta^{n_{i,j}}$表示第 n 个特征图上 (i, j) 位置的像素值. 空间激活映射考虑了不同通道特征的相关性, 公式表示为 $\gamma = \sum_{n=1}^{C}|\theta^n|^p$, 其中 p 为特征的激活程度, p 的值越大, 所对应的空间位置上的权重就越大, 表明该部分就会受到更多的关注. 对于生成器来说, 需要对最具判别力的区域赋予更大的权重, 故在生成器中取 p=2. 将上述操作的结果分别输入到相应的类别预测层预测类别概率. 首先, 使用全局最大池化的输出用于预测类别: $\xi_\alpha = \sum_n \omega_\alpha^n \cdot \alpha^{n_{i,j}}$. 之后, 将全局平均池化的结果输入全连接层, 得到$\xi_\beta = \sum_n \omega_\beta^n \cdot \beta^{n_{i,j}}$. 最后, 基于空间维度提取的特征, 得到全连接的结果是$\xi_\gamma = \sum_m \omega_\gamma^m \cdot \gamma^m$, 其中 m 是特征图θ的空间维度. 因此, 多角度注意力Γ_G预测的最终类别概率为:

$$\Gamma_G(x)=Concat(\xi_\alpha,\xi_\beta,\xi_\gamma)=\sigma\left(\sum\nolimits_n \omega^n \sum\nolimits_{i,j}\theta^{n_{i,j}}\right) \quad (1)$$

激活神经元的权重即为特征图的注意力权重. 一方面, 利用 CAM 分类器的权重信息对特征图θ进行加权, 得到$a_\alpha=\{\omega_\alpha\cdot\alpha^n \mid 1\leqslant n\leqslant C\}$. 另一方面, 将 SAM 的全连接层参数映射回特征图θ进行加权得到$a_\beta=\omega_\beta\cdot\beta$. 其中, ω_α和ω_β是两个预测层中激活的神经元的值. 考虑到通道间的相关性, 引入 1×1 卷积层得到注意力特征图$a_G(x)$:

$$a_G(x)=Conv(a_\alpha,a_\beta) \quad (2)$$

在图像翻译过程中, 解码器翻译前景对象的同时保留全局结构. 首先, 将多角度注意力输出的注意力特征图$a_G(x)$输入全连接层以提取风格信息. 之后, 将样式代码注入到解码器 D 的残差块中进行样式迁移, 以生成目标域图像y'.

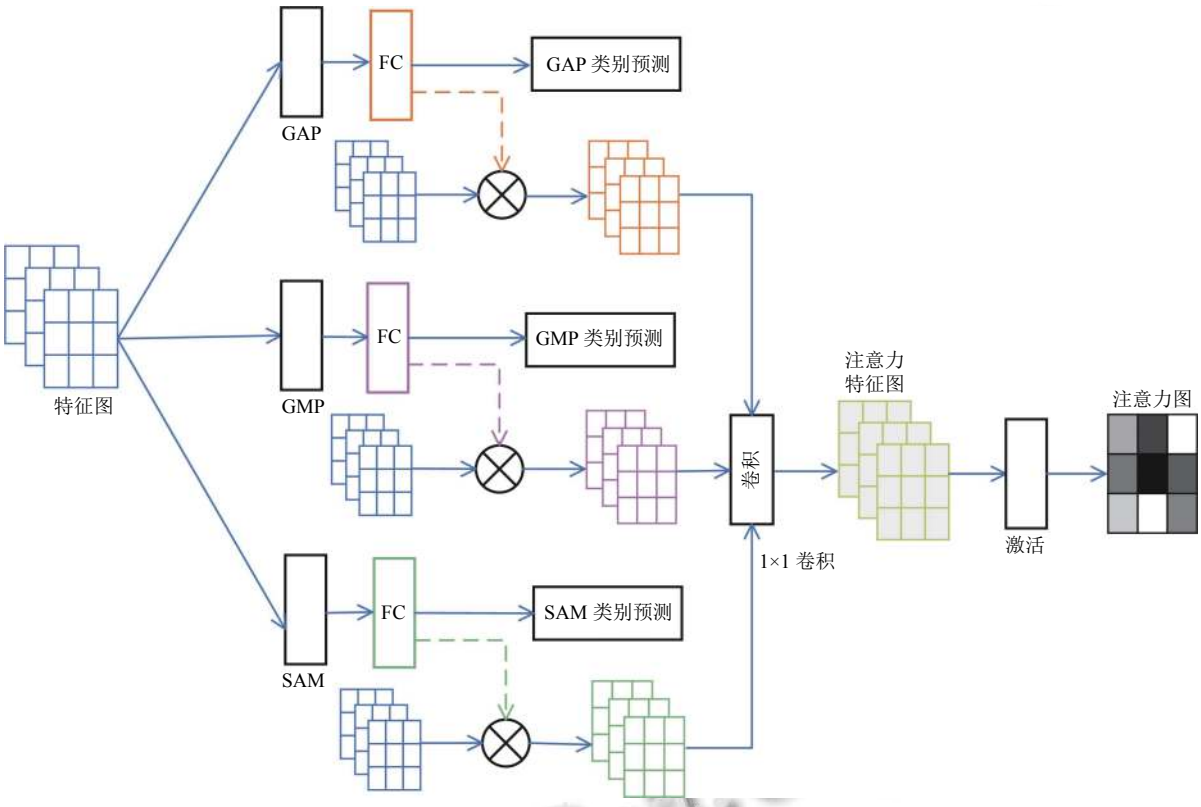


图 3 多角度注意力 (MA) 的详细处理流程

2.2 判别器

如图 2(c) 所示, 判别器由编码器 E、多角度注意力模块 (MA) Γ_D和分类器Φ组成. 编码器的结构在生成器和判别器中有所不同. 生成器的编码器包含下采样层和残差块, 而判别器的编码器只有下采样层. 嵌入在生成器和判别器中的多角度注意力采用相同的计算方法得到注意力特征图. 不同的是, 生成器中的注意力图引导网络关注两个域的判别性区域, 而判别器的注意力图则关注同一域中真实图像和生成图像的差异, 在关注局部区域时不能忽略全局信息, 因此将 p 设置为 1.

对图像$z\in\{X,G_X(y)\}$进行采样并将图像输入到判别器D_X中. D_X的编码器提取输入图像特征图θ, 即$\theta^n=E^n(x)$. 接下来将θ放入多角度注意力中, 得到类别预测概率$\Gamma_D(z)$和注意力特征图$a_D(z)$, 原理与生成器的多角度注意力相同. 最后, 将注意力特征图$a_D(z)$输入到分类器Φ:

$$\Phi(z)=\Phi(a_D(z))=\sigma(Conv(a_D(z))) \quad (3)$$

2.3 损失函数

为了促进网络生成更高质量的图像并使整个网络训练更加稳定, 总的损失函数包括跨域对抗损失、风格损失、内容重建损失和类别损失. 其中, 风格损失使生成的图像具有目标域图像的纹理, 内容重建损失使生成的图像保留源图像的内容. 为了减少图像中噪声的干扰, 风格损失和内容重建损失是使用图像的特征图而不是图像本身来计算的. 使用预训练的 VGG19 作

为感知网络来提取图像的高级语义特征.

2.3.1 跨域对抗损失

给定图像$x \in X$, 生成器的目标是合成具有 Y 域风格的图像$G_Y(x)$. 二值交叉熵损失公式如下[5]:

$$L_{\text{adv}}(X,Y) = E_{y\in Y}[\log(D_Y(Y))] + E_{x\in X}[\log(1 - D_Y(G_Y(x)))] \tag{4}$$

2.3.2 风格损失

本文引入风格损失函数L_{style}来惩罚生成图像和真实图像的风格特征差异. 风格损失函数使用 Gram 矩阵 M 来衡量每个维度的特征以及不同维度之间的关系. 对其求内积后得到的多尺度矩阵中, 对角元素提供了不同特征图的特征信息, 其余元素提供特征图之间的关联信息. 因此, M 不仅可以反映向量所包含的特征, 还可以反映不同特征之间的紧密程度. 计算两张图像的特征矩阵的 M, 通过比较 M 的差异来衡量两张图像的风格差异. L_{style}使翻译后的图像具有明显的目标域风格, 定义为[2]:

$$L_{\text{style}}(X) = E_{x\in X, y\in Y}[\|M_i^{\varphi}(x) - M_i^{\varphi}(G_X(y))\|_1] \tag{5}$$

其中, $M_i^{\varphi}(x)$是 VGG 网络φ中第 i 层激活的 Gram 矩阵. 在风格损失函数中, 第 i 层是 VGG19 网络的“conv4-4”.

2.3.3 内容重建损失

在图像翻译任务中, 翻译后的图像通常会改变输入图像的内容语义信息. 为了避免这个问题, 使用内容重建损失函数L_{rec}促进翻译图像保留源图像的内容信息. 具体地, 给定源图像$x \in X$, x 将被映射到目标域 Y. 进而将生成的目标域图像$G_Y(x)$作为G_X的输入图像, 将翻译后的图像还原输出源域图像$G_X(G_Y(x))$. 如果翻译后的图像可以很好地重建输入图像, 则认为生成器保留了源域图像的内容语义信息. 内容重建损失的公式如下[21]:

$$L_{\text{rec}}(X) = E_{x\in X}[\|\varphi_j(G_X(G_Y(x))) - \varphi_j(x)\|_1] \tag{6}$$

$\varphi_j(x)$为 VGG 网络φ中第 j 层下采样激活层, 在内容重建损失中, 第 j 层下采样激活层是 VGG19 网络的“conv5-4”.

2.3.4 类别损失

类别损失L_{cate}用来提示翻译网络准确定位当前两个域之间最具判别性的区域, 从而做出正确的类别预测. 使用Γ_{G_Y}表示域 Y 的生成器的类别预测器, 类别损失表示为[5]:

$$\begin{cases} L_{\text{cate}}^{G_Y} = \mathrm{E}_{x\in X, y\in Y}[\|\Gamma_{G_Y}(x) - \Gamma_{G_Y}(y)\|_1] \\ L_{\text{cate}}^{D_Y} = \mathrm{E}_{x\in X, y\in Y}[\|\Gamma_{D_Y}(y) - \Gamma_{D_Y}(G_Y(x))\|_2] \end{cases} \tag{7}$$

综上, MAGAN 完整的训练损失函数为:

$$\begin{cases} L_{G,\Gamma_G} = L_{\text{adv}} + L_{\text{style}} + L_{\text{rec}} + L_{\text{cate}}^{G} \\ L_{D,\Gamma_D} = -L_{\text{adv}} - L_{\text{cate}}^{D} \end{cases} \tag{8}$$

3 实验分析

3.1 基线模型

本文选取 CycleGAN[21]、DRIT[25]、UNIT[26]、MUNIT[17] 和 UGATIT[5] 作为基线模型. CycleGAN 首次利用未配对数据通过循环一致性损失来限制两个域之间的映射标准. DRIT 将特征信息解耦为内容空间和风格空间. UNIT 提出共享潜在空间的假设, 并结合可变分自编码来完成图像转换. 与 DRIT 类似, MUNIT 将图像信息分解为内容编码和样式编码.

3.2 数据集

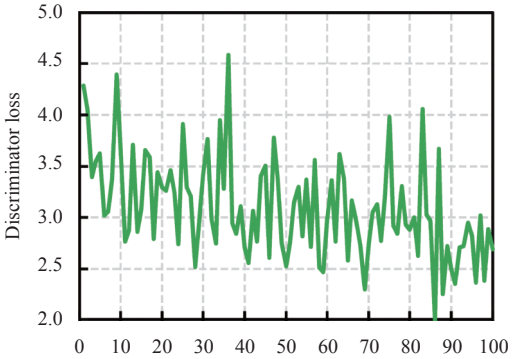
本文在 selfie2anime、cat2dog、horse2zebra 以及 vangogh2photo 这 4 个数据集上进行性能评估, 每个数据集的详细信息如表 1 所示, 所有图像大小均为 256×256. 特别地, selfie2anime 数据集只包含女性的照片, 人脸和动漫脸是差距较大的两个域. 两域间的图像翻译不仅需要颜色和纹理的变化, 更重要的是面部特征形状的转换, 具有更大的挑战性. 然而 horse2zebra 数据集的两个域仅在纹理上有所不同, 在形状上没有变化.

表 1 实验数据集详细信息

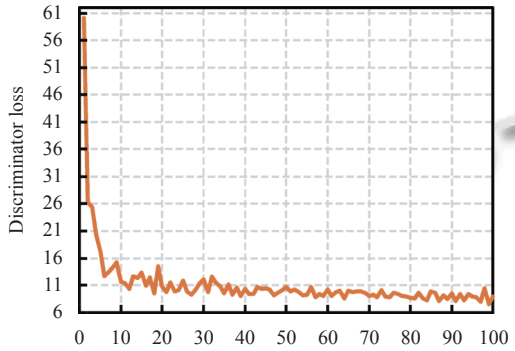
数据集	训练集	测试集
selfie2anime	6 800	200
cat2dog	2 035	200
horse2zebra	2 401	260
vangogh2photo	6 687	1 151

3.3 实现细节

第 2.1 节中介绍的网络结构作为 MAGAN 的基准骨干网络. 为了增强训练数据, 将图像以 0.5 的概率水平翻转. 对于训练参数, 批处理大小设置为 1, 模型训练总步长为 1 000 k. 使用 Adam 作为优化器, 其中β_1= 0.5 和β_2=0.999. 在前 500k 步的迭代训练中以 0.000 1 的学习率训练模型, 并将学习率设置为在后 500k 步的迭代训练中线性衰减. 图 4 展示了本文模型在训练过程中生成器损失和判别器损失的收敛曲线. 对于激活函数, 本文保留基线模型[5] 的设计, 在生成器中使用 ReLU, 在判别器中使用斜率为 0.2 的 LeakyReLU. 所有实验均在单个 NVIDIA RTX3090 GPU 上完成.



(a) 判断器损失的收敛曲线



(b) 生成器损失的收敛曲线

图 4 判别器损失和生成器损失的收敛曲线

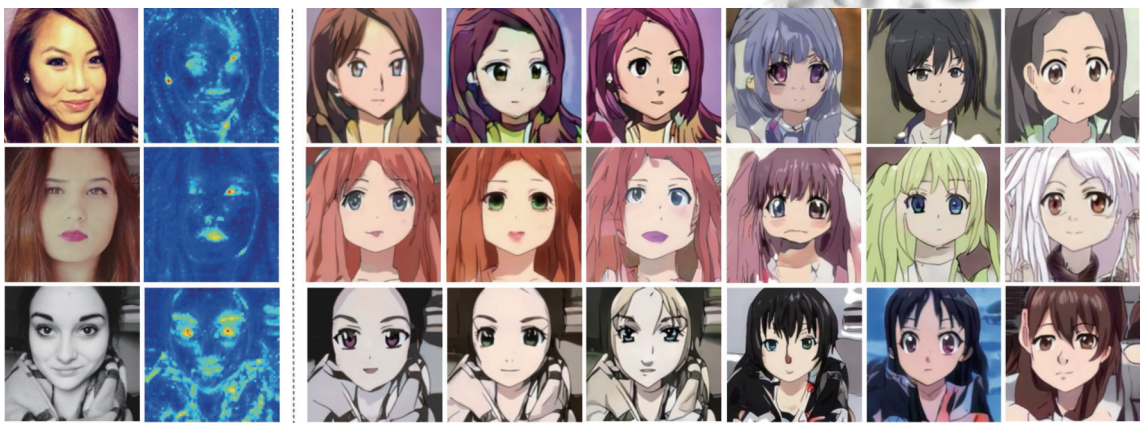
3.4 定性比较

为了验证本文模型 MAGAN 的有效性, 在所有翻译任务上进行了对比评估. 图 5 显示了 selfie2anime 数据集上对翻译结果的定性比较. 实验结果表明, CycleGAN 生成的动漫人在图 5(e) 中有明显失真. DRIT、UNIT 和 MUNIT 生成的图像与输入图像明显不一致. 它们没有充分保留源图像的全局信息. 如图 5(d) 所示, UGATIT 翻译的动漫脸面部轮廓不平滑, 并引入了冗余特征, 如第 3 行生成了两个鼻子. 本文的方法 (图 5(c)) 优于上述模型, 生成了高质量的动漫面孔. 本文对多角度注意力的精心设计使得网络能够精确地关注两个域最具有判别性的区域并有效地翻译局部结构, 从而去除翻译图像中的特征扭曲. 图 6 展示了其他数据集的定性比较结果.

3.5 定量比较

3.5.1 内核初始距离

内核初始距离 (kernal inception distance, KID)[27] 通过计算最大平均差的平方来衡量两组样本之间的差异. KID 值越小表明生成图像的分布更接近真实图像的分布. 在 KID 的计算中测试集被划分为 10 个大小为 10 的子集, 并将 KID 的度设置为 9. 在表 2 中, 使用 KID 来评估本文提出的模型和基线模型在所有数据集上的性能. 从表 2 可以看出, MAGAN 得到了最优的 KID 值. 这意味着本文提出的模型生成的图像具有最接近真实图像的分布, 因此它们在视觉上看起来很相似并且适用于多种图像翻译任务. 其他对比模型的 KID 分数来自 UGATIT[5].



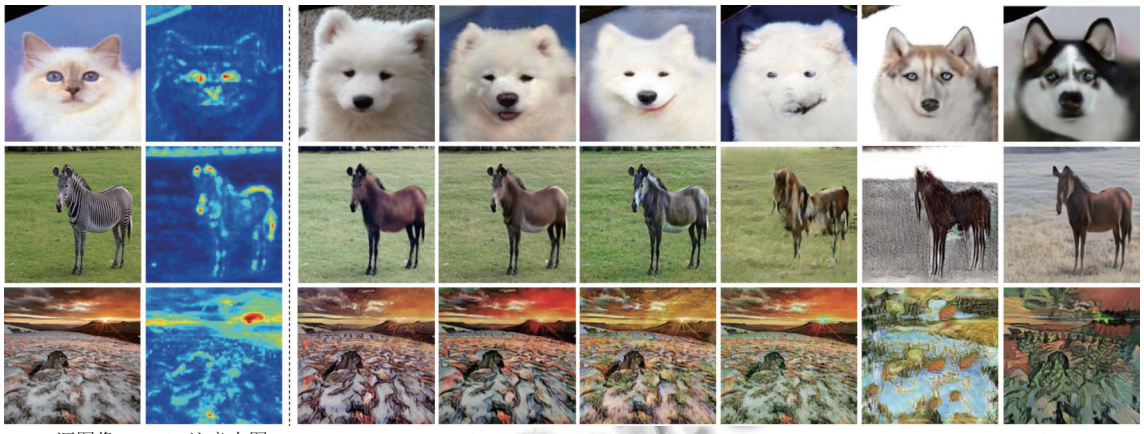
(a) 源图像 (b) 注意力图 (c) MAGAN (ours) (d) UGATIT (e) CycleGAN (f) UNIT (g) MUNIT (h) DRIT

图 5 所有对比模型在 selfie2anime 任务上的翻译结果

3.5.2 用户研究

为了使实验结果的评价更符合人类的感知, 本文还通过用户研究对生成的图像效果进行定量比较. 感知研究实验涉及 20 名受试者. 对于每个参与者, 一次展示 1 张源域图像和 5 张由翻译模型生成的目标域图像, 参与者在不知情的情况下挑选出具有最佳

视觉效果的图像. 本文一共收到了来自 20 名受试者的 1 000 份结果. 表 3 中的结果表明我们的方法获得了更多的投票, 意味着用户更喜欢我们的方法翻译的图像.



(a) 源图像 (b) 注意力图 (c) MAGAN (ours) (d) UGATIT (e) CycleGAN (f) UNIT (g) MUNIT (h) DRIT

图 6 所有对比模型在其他数据集上的翻译结果

表 2 所有数据集上关于内核初始距离的定量比较 (KID×100±std×100)

模型	selfie2anime		cat2dog		horse2zebra		vangogh2photo	
	A2B	B2A	A2B	B2A	A2B	B2A	A2B	B2A
CycleGAN	13.08±0.49	11.84±0.74	8.92±0.69	9.94±0.36	8.05±0.72	8.00±0.66	5.46±0.33	4.68±0.36
UNIT	14.71±0.59	26.32±0.92	8.15±0.48	9.81±0.34	10.44±0.67	14.93±0.75	**4.26±0.20**	9.72±0.33
MUNIT	13.85±0.41	13.94±0.72	10.13±0.27	10.39±0.25	11.41±0.83	16.47±1.04	13.08±0.34	9.53±0.35
DRIT	15.08±0.62	14.85±0.60	10.92±0.33	10.86±0.24	9.79±0.62	10.98±0.55	12.65±0.35	7.72±0.34
UGATIT	11.61±0.57	11.52±0.57	7.07±0.65	8.15±0.66	7.06±0.80	7.47±0.71	4.28±0.33	5.61±0.32
MAGAN (ours)	**9.48±0.15**	**9.35±0.57**	**6.32±0.15**	**6.50±0.17**	**6.42±0.46**	**6.77±0.34**	4.28±0.21	**4.50±0.10**

表 3 用户研究评估结果 (%)

模型	selfie2anime	cat2dog	horse2zebra	vangogh2photo
CycleGAN	2.01	1.79	6.06	1.33
UNIT	1.14	1.66	1.12	42.65
MUNIT	0.56	1.03	0.64	0.53
UGATIT	20.25	21.40	29.33	20.69
MAGAN (ours)	**76.04**	**74.12**	**62.85**	**34.80**

3.6 消融实验

本文设计了消融实验来验证结合生成对抗网络及多角度注意力的图像翻译模型的性能. 所有的消融实验都在 selfie2anime 数据集上完成.

3.6.1 多角度注意力的实验分析

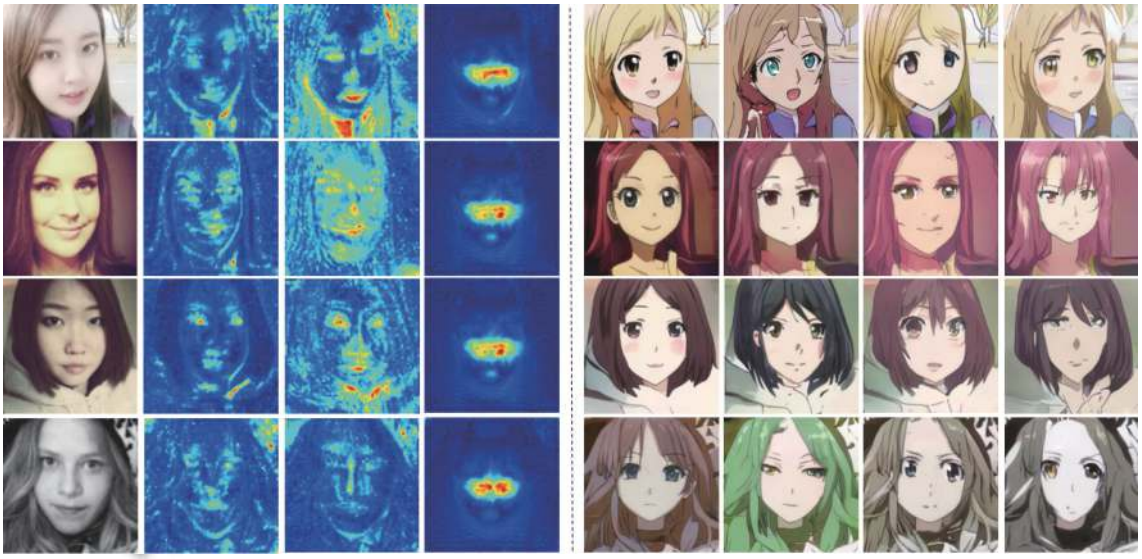
在图 7(b) 中, 多角度注意力图促使网络关注两个域最具辨别性的图像区域. 设计了 3 个基线: “w/ CAM”(图 7(f))、“w/ SAM”(图 7(g)) 和“w/o attention” (图 7(h)). 其中, “w/o attention”没有使用注意力机制, 翻译效果不尽人意. “w/ CAM”通过类激活映射技术计算注意力图. 由于 CAM 忽略了不同通道之间的特征关联, 导致翻译的图像中存在严重的特征扭曲. “w/ SAM” 只能捕获局部特征信息, 边缘定位效果较差. 相比之下, 多角度注意力不仅可以定位到完整的目标, 还可以充分利用目标的空间相关性来生成高质量的动漫图像.

由表 4 可知, “w/o attention”在 selfie2anime 数据集上的 KID 分数最高. 验证了注意力机制在提高网络性能方面发挥了重要作用. 本文提出的模型在人脸和动漫脸相互转换的任务上的 KID 值最低, 分别为 9.48 和 9.35, 表明本文提出的模型生成的图像更加真实.

3.6.2 特征激活程度的实验分析

本文研究了特征激活度对网络性能的影响, 并可视化了不同激活度下的注意力图. p 的取值有 1 (“w/ low”)、2 (“w/ middle”) 或 3 (“w/ high”). 同时, 为了验证相同激活度在不同网络架构的效果差异, MAGAN 分别为生成器网络和判别器网络设置了每个激活度的特征激活. 如图 8 所示, 每列表示生成器中使用相同的激活程度, 每行表示判别器中使用相同的激活程度. 其中, 第 1–3 行是注意力特征图的可视化, 第 4–6 行是生成器输出相应的图像. 当 p=1 时, 网络倾向于关注全局信息, 而对重要敏感区域 (例如眼睛和嘴巴) 的信息关注度不够; 随着 p 值的增加, 更多的注意力

被放在最具辨别性的局部区域; 当 p=3 时, 网络失去了捕捉边界定位细节的能力 (例如, 面部轮廓扭曲). 生成器网络需要对辨别性最大的区域赋予更多的权重, 因此将 p 设置为 2 效果最好. 与生成器网络不同, 判别器网络必须从整体上判断图像的真实性, 因此将 p 设置为 1 最合适. 在表 5 中, 生成器“w/ middle”和判别器“w/ low”的模型在 selfie2anime 数据集上的 KID 得分最低.



(a) 源图像 (b) 多角度注意力结构的注意力图 (c) “w/ CAM”的注意力图 (d) “w/ SAM”的注意力图 (e) MAGAN (ours) (f) “w/ CAM” (g) “w/ SAM” (h) “w/o attention”

图 7 对多角度注意力消融研究的定性比较

表 4 selfie2anime 数据集上对多角度注意力消融研究的定量比较 (KID×100±std×100)

Method	CAM	SAM	selfie2anime	anime2selfie
w/o attention	—	—	13.30±0.98	15.49±0.54
w/ CAM	√	—	11.61±0.57	11.52±0.57
w/ SAM	—	√	12.84±0.98	14.51±0.49
w/ MA	√	√	**9.48±0.15**	**9.35±0.57**

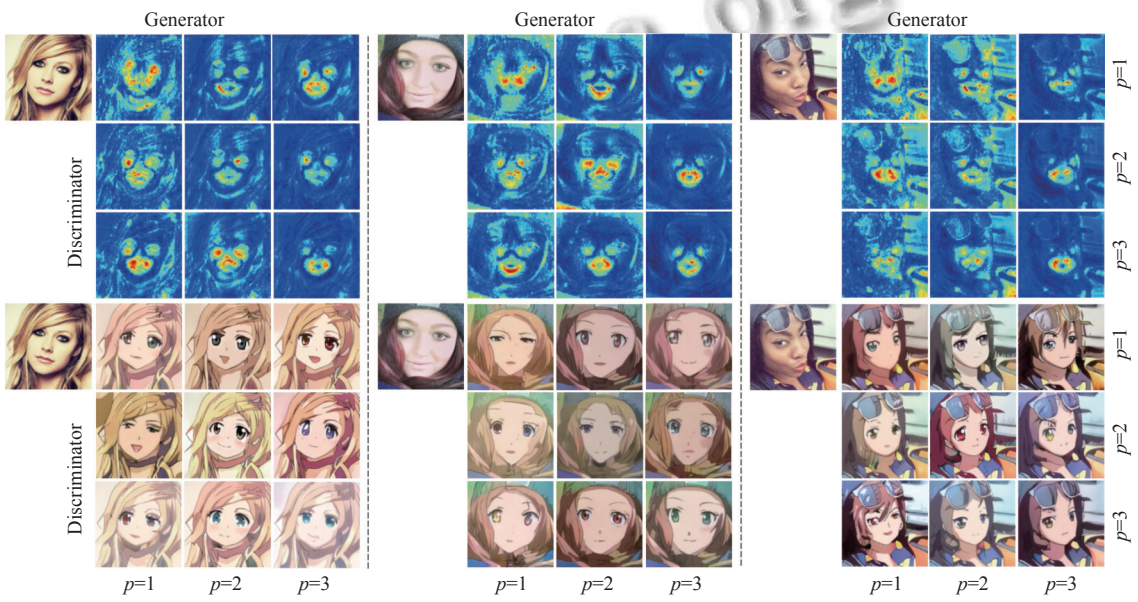


图 8 对特征激活程度消融研究的定性比较

表 5 在 selfie2anime 数据集上对特征激活程度消融研究的定量比较 (KID×100±std×100)

D	“w/ low”		“w/ middle”		“w/ high”	
	selfie2anime	anime2selfie	selfie2anime	anime2selfie	selfie2anime	anime2selfie
“w/ low”	11.68±0.41	12.49±0.81	**9.48±0.15**	**9.35±0.57**	11.95±0.45	10.09±0.21
“w/ middle”	12.79±0.83	11.95±1.49	12.05±0.29	11.21±1.67	12.13±0.55	11.27±0.67
“w/ high”	12.63±0.04	11.93±0.40	9.86±0.33	11.57±1.20	13.27±0.41	11.51±0.58

4 结论与展望

本文提出了一种结合生成对抗网络及多角度注意力的无监督图像翻译模型. 本文所提出的注意力模型能够有效地引导生成器和判别器关注判别性区域. 其不但可以保留源图像的全局信息, 而且能够将局部风格转换为目标域的风格. 实验结果表明, 与其他模型相比, 本文提出的 MAGAN 生成了高质量的图像. 此外, 多角度注意力可以增强模型的抗干扰能力, 适用于多种类型的图像翻译任务. 然而, 由于我们的注意力模块不能实现内容和风格的完全解耦, 所以生成的图像中包含预期外的内容. 因此, 在未来的工作中, 将研究一种新的模型以完全解耦图像的内容和风格, 从而实现更精准的图像翻译.

参考文献

1 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

2 Choi Y, Uh Y, Yoo J, *et al*. StarGAN v2: Diverse image synthesis for multiple domains. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8185–8194. [doi: 10.1109/CVPR42600.2020.00821]

3 Liang XD, Zhang H, Lin L, *et al*. Generative semantic manipulation with mask-contrasting GAN. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 574–590.

4 Liu MY, Huang X, Mallya A, *et al*. Few-shot unsupervised image-to-image translation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 10550–10559. [doi: 10.1109/ICCV.2019.01065]

5 Kim J, Kim M, Kang H, *et al*. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. Proceedings of the 8th International Conference on Learning Representations (ICLR). Addis Ababa: OpenReview.net, 2020. 1–19.

6 Zhou BL, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2921–2929. [doi: 10.1109/CVPR.2016.319]

7 Li YJ, Fang C, Yang JM, *et al*. Diversified texture synthesis with feed-forward networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 266–274. [doi: 10.1109/CVPR.2017.36]

8 Chen SSC, Cui H, Du MH, *et al*. Cantonese porcelain classification and image synthesis by ensemble learning and generative adversarial network. Frontiers of Information Technology & Electronic Engineering, 2019, 20(12): 1632–1643.

9 Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4396–4405. [doi: 10.1109/CVPR.2019.00453]

10 吴福祥, 程俊. 基于自编码器生成对抗网络的可配置文本图像编辑. 软件学报, 2022, 33(9): 3139–3151. [doi: 10.13328/j.cnki.jos.006622]

11 秦魁, 侯新国, 周锋, 等. fire-GAN: 基于生成对抗网络的火焰图像生成算法. 激光与光电子学进展, 2022, 1–13. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=JGDJ202207130F9. (2022-07-17).

12 Chen L, Wu L, Hu ZZ, *et al*. Quality-aware unpaired image-to-image translation. IEEE Transactions on Multimedia, 2019, 21(10): 2664–2674. [doi: 10.1109/TMM.2019.2907052]

13 Liu Y, Chen W, Liu L, *et al*. SwapGAN: A multistage generative approach for person-to-person fashion style transfer. IEEE Transactions on Multimedia, 2019, 21(9): 2209–2222. [doi: 10.1109/TMM.2019.2897897]

14 Song YH, Yang C, Lin Z, *et al*. Contextual-based image inpainting: Infer, match, and translate. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 3–18.

15 王清和, 曹兵, 朱鹏飞, 等. 基于自判别循环生成对抗网络的人脸图像翻译. 中国科学: 信息科学, 2022, 52(8): 1447–1462.

16 Yang C, Kim T, Wang RZ, *et al*. Show, attend, and translate: Unsupervised image translation with self-regularization and attention. IEEE Transactions on Image Processing, 2019, 28(10): 4845–4856. [doi: 10.1109/TIP.2019.2914583]

17 Huang X, Liu MY, Belongie S, *et al*. Multimodal unsupervised image-to-image translation. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 179–196.

18 Choi Y, Choi M, Kim M, *et al*. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8789–8797. [doi: 10.1109/CVPR.2018.00916]

19 Isola P, Zhu JY, Zhou TH, *et al*. Image-to-image translation with conditional adversarial networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5967–5976. [doi: 10.1109/CVPR.2017.632]

20 Wang TC, Liu MY, Zhu JY, *et al*. High-resolution image synthesis and semantic manipulation with conditional GANs. Proceedings of the 2018 IEEE CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8798–8807. [doi: 10.1109/CVPR.2018.00917]

21 Zhu JY, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2242–2251. [doi: 10.1109/ICCV.2017.244]

22 Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1510–1519. [doi: 10.1109/ICCV.2017.167]

23 Zhang H, Goodfellow I, Metaxas DN, *et al*. Self-attention generative adversarial networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 7354–7363.

24 Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon: OpenReview.net, 2017. 1–13.

25 Lee HY, Tseng HY, Huang JB, *et al*. Diverse image-to-image translation via disentangled representations. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 36–52.

26 Liu MY, Breuel T, Kautz J. Unsupervised image-to-image translation networks. Proceedings of the 31st International Conference on Neural Information Processing System. Long Beach: ACM, 2017. 700–708.

27 Bińkowski M, Sutherland DJ, Arbel M, *et al*. Demystifying MMD GANs. Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver: OpenReview.net, 2018. 1–36.

(校对责编: 牛欣悦)