风格化影像合成技术

摘要：风格化影像合成技术是一种将原始影像的颜色、纹理等内容变换到另一种表现模式的风格迁移技术。根据目标风格的不同，可分为基于物理的真实感风格化和应用艺术处理的非真实感风格化，卡通风格化是非真实感风格化中最灵活、定制性最强和难度最高的类型。**本文从风格控制方式的视角对卡通风格化相关技术进行归纳，将其划分为目标诱导式、语义学习式和参考匹配式**。目标诱导式又称白盒式风格控制，具有较强的可解释性，通过损失函数定制风格化需求，并通过手工调节各项损失权重来精细化控制目标风格。语义学习式通过定制单一风格数据集，由数据驱动语义特征学习来实现风格控制，此类方法可归纳为黑盒式风格控制，用户无法直接控制风格，而是通过筛选数据集来间接控制。参考匹配式更关注通用性，此类方法使用混合风格数据训练，在预测时通过用户提供参考图像的方式来定义目标风格，既具有比白盒式方法更高的灵活性，降低了手工定制风格的要求，又具有比黑盒式方法更好的可解释可控制性，并且无需重复训练模型，是大模型技术发展的主要方向。此外，模型预训练和交叉域迁移学习技术逐渐在针对时序影像的风格化任务中得到广泛应用，将时序风格化模型解耦为建模单帧风格和建模时序动作的两类可搭配子模型，建模单帧风格的子模型可由任意预训练的图像风格化模型构成，而建模时序动作的子模型可以使用带有物理约束的真实场景数据进行训练，二者结合能够实现具有定制风格和拟真动态的时序风格化影像，这种混合架构、即插即用式模型已逐渐成为风格迁移领域的研究热点之一。

关键词：卡通风格化，风格迁移，生成式对抗网络，影像合成

一、研究现状

卡通是一种流行的艺术形式，已被广泛应用于各种视觉创作任务中。现代卡通动画工作流允许艺术家使用各种原生艺术资源来创建内容，将自然影像等原生艺术资源转换为卡通动画资源的过程称为卡通风格化。卡通风格化是一项综合性任务，通常涉及风格迁移、图像转换和影像合成技术。

生成式对抗网络（Generative Adversarial Network，GAN）[1]是解决计算机视觉生成任务的有力工具，已经在图像转换[2][3]、超分辨率[4][5]、图像修补[6][7]、图像扩展[8]等领域取得了良好成效，卡通风格化任务正是图像转换技术的应用场景之一。基于 GAN 的图像转换技术研究开始于 Pix2Pix[2]，这是一个基于监督学习的条件控制 GAN 图像转换模型，它能够将图像从一个域（如自然图像）映射到另一个域（如卡通图像），但依赖于能够表征域特征的配对训练图像。CycleGAN[9]通过施加循环一致性约束，进一步解决了配对训练图像难以获取的问题。为生成差异化的目标域图像，多模态方法提取域不变特征并将其与定制风格融合，这使得目标域合成图像的自然度和丰富度得到极大提升。尽管图像转换技术已经取得了巨大的进步，但对于以卡通图像为目标域的图像转换任务而言，

生成模型的通用性较差且合成图像的视觉质量仍不够理想，这是因为卡通图像往往具有不同程度的夸张几何结构和错位的描边倾向，且风格统一和内容差异化需求在一定程度上存在内部矛盾。

针对特定语义对象的卡通风格化已经产生了一系列结果。在场景卡通风格化领域，CartoonGAN[10]、AnimeGAN[11]提出了强调“边缘清晰、着色平滑”的卡通风格化目标，通过构造语义内容损失和边缘损失来指导训练；White-Box[12]是一种多判别器 GAN 模型，它使用三种特征空间表示法来精细化控制卡通风格，巧妙结合导向滤波、超像素、纹理摘要等传统图像处理技术，设计了三种风格化损失函数，以一种“白盒”的方式对图像的表面特征、结构特征和纹理特征进行偏好设定，从而生成具有不同平滑度、结构概化度和细节精细度的卡通图像，提供了更加丰富的控制手段。这类卡通风格化技术能够将自然场景转换为高质量的卡通场景，但由于结构性约束较为严格，对夸张风格（大眼睛和简约的鼻子、嘴巴）的卡通人物肖像的合成无能为力。在肖像卡通风格化领域，APDrawingGAN[13]使用一种分层 GAN 架构，包含专用于面部特征区域（眼、鼻、口等）的六个局部网络以及用于捕获整体特征的全局网络，从而强化对肖像面部组件特征的刻画能力；MangaGAN[14]使用类似的分治思路，追加设计了一个几何变换网络用于重编排和夸张人物面部特征，从而生成极具漫画风格的卡通肖像。这类技术能够生成指定风格的卡通化肖像，但易受内容缺失干扰，导致出现不自然的纹理错位。Gated Mapping GAN[15]通过巧妙设计的门控映射单元实现了对场景和人物肖像卡通风格化的通用处理，并排除了场景和人物肖像卡通风格的相互干扰；该模型使用混合风格图像进行训练，从而支持基于参考的风格化，可由用户提供转换源图像和任意风格的参考图像并生成对应风格的合成图像；Gated Mapping GAN 还可迁移应用于视频卡通风格化合成任务中，但要求提供完整视频源。MoCoGAN-HD[16]是一个支持交叉图像域视频合成的高效模型，它将视频特征解耦为内容编码和动作编码，使用基于 LSTM[17]的动作生成器来建模时序信息，采用分层 GAN 架构，在帧级和视频级进行对抗训练；使用包含动作序列的真实视频训练动作生成器，搭载预训练的卡通图像合成网络，可将动作序列作用于卡通图像域以根据单张图像合成卡通风格视频。

二、典型算法

White-Box[12]是一种基于多判别器 GAN 的图像卡通风格化技术。此技术充分发挥了 GAN 模型的内容生成能力，并通过导向滤波、超像素、纹理摘要等传统图像处理技术增强了转换结果的可解释性和可控性。

White-Box 卡通风格化系统如图 2-1 所示。左图呈现了 White-Box 所使用的三种特征空间，它们是用于反映图像低频特征与平滑度的表面特征空间，用于反映全局结构信息和颜色块状构型的结构特征空间，以及用于反映图像细节和边缘的纹理特征空间。这三种特征表示通过传统图像处理技术提取，用于构建风格化损失函数，以实现面向具体

需求的生成风格控制。右图呈现了 White-Box 所使用的多判别器 GAN 模型架构。编解码式生成器 G 用于实现图像从自然风格到卡通风格的转换，判别器 $D_{surface}(D_s)$旨在区分生成器输出和卡通图像表面特征的真实度，另一个判别器 $D_{texture}(D_t)$用于区分生成器输出和卡通图像纹理特征的真实度。预训练 VGG 网络[18]用于提取高级特征，结合损失函数设计对原始自然图像和生成图像的全局结构信息和内容信息施加空间约束。

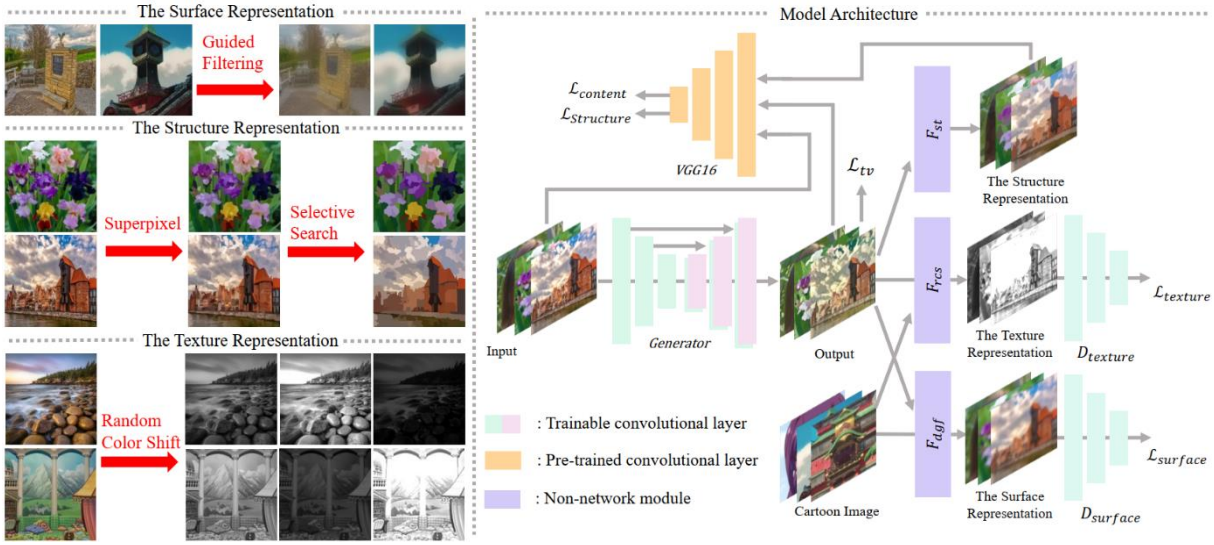

图2-1 White-Box 卡通风格化系统

F_{dgf}是可微导向滤波器模块，它接受图像I作为输入和导向图，输出表面特征图$F_{dgf}(I,I)$。在判别器D_s输出端构造表面特征损失函数：（式中I_p表示输入图像，I_c表示参考卡通图像）

$$L_{surface}(G,D_s) = \log D_s\left(F_{dfg}(I_c,I_c)\right) + \log\left(1 - D_s\left(F_{dfg}\left(G(I_p),G(I_p)\right)\right)\right) \tag{2-1}$$

F_{st}是结构摘要模块，它使用 Felzenszwalb 算法[19]结合选择搜索算法[20]生成语义分割图，并使用自适应着色算法[12]解决超像素技术带来的图像灰暗和低对比度问题，以增强结构特征提取质量。使用预训练 VGG16 网络[18]提取高维特征，对图像的全局结构信息施加空间约束，构造结构特征损失函数：（式中I_p表示输入图像，I_c表示参考卡通图像）

$$L_{structure} = \left\| VGG_n\left(G(I_p)\right) - VGG_n\left(F_{st}\left(G(I_p)\right)\right)\right\|_1 \tag{2-2}$$

F_{rcs}是纹理摘要模块，该模块接受三通道 RGB 彩色图像输入，执行图像去色和随机偏移处理，提取纹理相关的高频信息，生成单通道纹理特征图：

$$F_{rcs}(I_{rgb}) = (1-\alpha)\cdot\left(\beta_1 I_r + \beta_2 I_g + \beta_3 I_b\right) + \alpha \cdot Y \tag{2-3}$$

式中，I_{rgb}是三通道输入图像，I_r、I_g、I_b表示通道图像，Y表示输入图像的灰度化结果，分配权重$\alpha = 0.8$，$\beta_1,\beta_2,\beta_3 \sim U(-1,1)$。$F_{rcs}$的设计目标是在提取高频特征的同时，

尽量抑制颜色和亮度信息的干扰作用，使判别器D_t能够得到有效训练。在判别器D_t输出端构造纹理特征损失函数：（式中I_p表示输入图像，I_c表示参考卡通图像）

$$L_{texture}(G, D_t) = \log D_t\left(F_{rcs}(I_c)\right) + \log\left(1 - D_t\left(F_{rcs}\left(G\left(I_p\right)\right)\right)\right) \tag{2-4}$$

White-Box 模型训练。利用F_{dgf}、F_{st}、F_{rcs}所提取的三种特征对生成器和两个判别器进行同步优化，设置损失权重参数λ_1、λ_2、λ_3、λ_4、λ_5构造总损失函数：

$$L_{total} = \lambda_1 L_{surface} + \lambda_2 L_{texture} + \lambda_3 L_{structure} + \lambda_4 L_{content} + \lambda_5 L_{tv} \tag{2-5}$$

式中，全方差损失L_{tv}用于增加生成图像的平滑度，抑制高频噪声和椒盐噪声；内容损失$L_{content}$用于确保卡通化图像和原始图像的语义内容一致性，该项损失在 VGG16 高维特征空间中计算：（式中，H、W、C是图像的空间维数，即高、宽、通道；∇_x、∇_y表示x、y方向的差分）

$$L_{tv} = \frac{1}{HWC}\left\|\nabla_x\left(G\left(I_p\right)\right) + \nabla_y\left(G\left(I_p\right)\right)\right\|_1 \tag{2-6}$$

$$L_{content} = \left\|VGG_n\left(G\left(I_p\right)\right) - VGG_n\left(I_p\right)\right\|_1 \tag{2-7}$$

White-Box 的突出优势在于“白盒”式的卡通风格精细化控制，而不再完全取决于训练数据分布。通过调节损失项权重λ_1、λ_2、λ_3、λ_4，可实现不同风格的卡通化，如图 2-2 所示。

图2-2 White-Box 卡通风格控制

Gated Mapping GAN[15]是一种通用型图像卡通风格化技术。其独特优势在于能够接收参考图像作为输入，并将源图像转换为参考图像风格，即转换风格不再完全由训练集决定，训练集也无需仅由单一风格卡通图像构成。Gated Mapping GAN 通过巧妙设计的门控映射单元实现了对场景和人物肖像卡通风格化的通用处理，只需在混合风格数据集上进行一次训练，该模型就能根据参考图像生成风格迥异的卡通场景、写实风格以及

夸张风格的人物肖像。

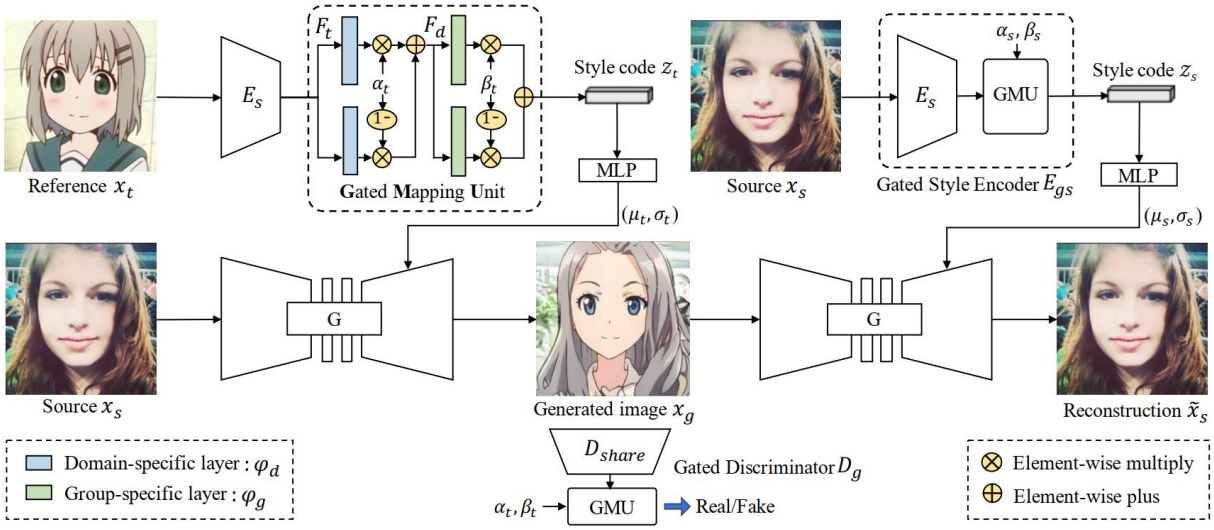

图2-3 Gated Mapping GAN 卡通风格化系统

Gated Mapping GAN 卡通风格化系统如图 2-3 所示。编解码式生成器 G 负责执行图像域转换，兼用于从源图像到转换域图像的生成以及从转换域图像到源图像的重建过程中。门控风格编码器E_{gs}用于生成图像风格编码，对参考图像x_t、源图像x_s分别生成风格编码z_t、z_s，再通过多层感知机（Multi-Layer Perceptron，MLP）将风格特征聚合到生成器 G 的解码端作为风格引导。门控映射单元（Gated Mapping Unit，GMU）是实现混合风格学习的关键模块，它由域调控层（自然/卡通图像域）和组调控层（人物肖像/场景）构成，用以排除图像域和组别之间的相互干扰，使不同风格类型都能得到独立且有效的学习。门控判别器D_g用于判别生成器 G 输出的转换域图像的真实度，它同样使用 GMU 来强化对特定风格图像真实度的判别能力

对于源图像x_s和参考图像x_t，首先通过E_{gs}生成风格编码z_t、z_s，然后通过 MLP 层并使用自适应实例标准化（Adaptive Instance Normalization，AdaIN）技术[21]将风格编码聚合到编解码式生成器 G 中，生成源x_s相对于参考x_t风格的生成图像$x_g = G(x_s, f(z_t))$；其中$f(z_t)$表示通过 MLP 生成的 AdaIN 参数（尺度μ和偏移σ）。针对源图像的重建$\tilde{x_s} = G(x_g, f(z_s))$用于内容一致性训练，所构成的循环结构使得模型训练能够在非配对数据集条件下进行。

GMU 用于图像域和组别引导信息嵌入。通过编码器E_s提取参考图像x_t的共同特征F_t，共同特征再通过域调控层$\varphi_{d_i}(i = 0,1)$生成域特征并通过调控门进行选择：（式中，$\alpha_t \in \{0,1\}$，表示图像属于自然图像域或卡通图像域）

$$F_d = \alpha_t \cdot \varphi_{d_0}(F_t) + (1 - \alpha_t) \cdot \varphi_{d_1}(F_t) \tag{2-8}$$

所选择出的域特征图再通过组调控层$\varphi_{g_i}(i = 0,1)$与调控门生成风格编码：（式中，$\beta_t \in \{0,1\}$，表示图像属于人物肖像或场景）

$$z_t = \beta_t \cdot \varphi_{g_0}(F_d) + (1-\beta_t) \cdot \varphi_{g_1}(F_d) \tag{2-9}$$

Gated Mapping GAN 模型训练。使用包含对抗损失L_{adv}、重建内容损失L_{rec}、风格损失L_{sty}、异化损失L_{ds}构造复合损失函数：

$$L_{total} = L_{adv} + \lambda_{rec}L_{rec} + \lambda_{sty}L_{sty} + \lambda_{ds}L_{ds} \tag{2-10}$$

对抗损失L_{adv}用于激励生成器 G 学习卡通图像分布特征：

$$L_{adv} = \mathbb{E}_{x_s,x_t}\left[\log\left(1 - D_g\left(G\left(x_s, f(z_t)\right)\right)\right)\right] + \mathbb{E}_{x_t}\left[\log\left(D_g(x_t)\right)\right] \tag{2-11}$$

重建内容损失L_{rec}用于保障生成结果的语义内容一致性，使得转换域图像可以被成功地重建为源图像：

$$L_{rec} = \left\| G\left(G\left(x_s, f(z_t)\right), f(z_s)\right) - x_s \right\|_1 \tag{2-12}$$

风格损失L_{sty}用于保障生成图像卡通风格与参考图像风格的一致性：

$$L_{sty} = \left\| E_{gs}\left(G\left(x_s, f(z_t)\right)\right) - z_t \right\|_1 \tag{2-13}$$

异化损失L_{ds}用于激励针对不同参考图像的差异化风格学习，对任意两个不同的参考风格编码$\left(z_{t_1}, z_{t_2}\right)$和同一个源图像$x_s$，合成图像应当具有显著差异：（注意此项损失带有负号）

$$L_{ds} = -\left\| G\left(x_s, f\left(z_{t_1}\right)\right) - G\left(x_s, f\left(z_{t_2}\right)\right) \right\|_1 \tag{2-14}$$

Gated Mapping GAN[15]的突出优势在于基于参考的风格控制，只需在混合数据集上进行一次训练，就可兼用于场景和人物肖像的卡通风格化。如图 2-4 所示，基于参考的风格控制使 Gated Mapping GAN 在实际应用中非常灵活且强大，它能实现夸张风格的人物肖像生成，且能保障夸张风格与写实风格、多种场景风格卡通图像生成之间的互不干扰性。

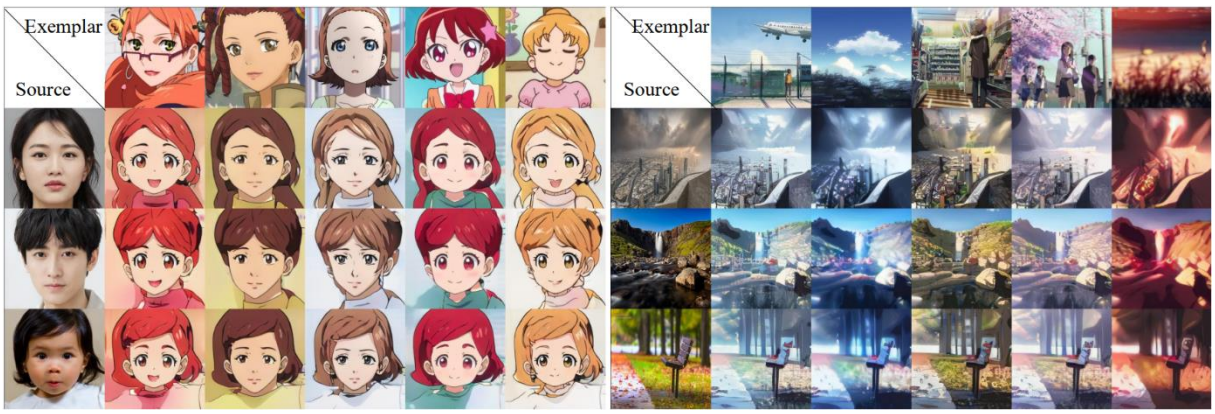

图2-4 Gated Mapping GAN 基于参考的卡通风格化

MoCoGAN-HD[16]是一种支持交叉图像域的高清视频合成技术。它将视频特征解耦为内容编码和动作编码，使用 LSTM[17]和差分隐空间编码来建模时序信息；MoCoGAN-HD 使用迁移学习技术来提高生成效率，在内容编码部分采用预训练的图像合成网络（例如 StyleGAN[22]、BigGAN[23]）。内容和动作的解耦使得 MoCoGAN-HD 可以从真实视频训练集中学习动作序列，但使用卡通图像合成网络来生成内容，从而生成模拟真实视频动作特征的卡通视频。

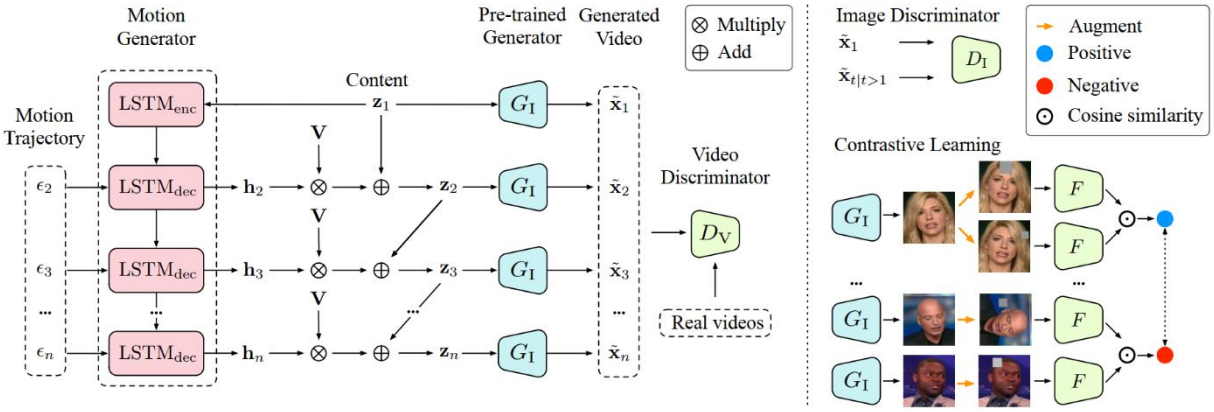

图2-5 MoCoGAN-HD 视频合成技术体系

MoCoGAN-HD 视频合成技术体系如图 2-5 所示。此架构包含一个动作生成器G_M和一个图像生成器G_I。G_M由两个 LSTM 网络构成，负责生成动作轨迹序列$Z=\{z_1,z_2,\cdots,z_n\}$（n是所需合成的帧数）；然后由G_I接收每一帧动作轨迹并合成视频$\tilde{v}=\{\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_n\}$，其中$\tilde{x}_t=G_I(z_t),t=1,2,\cdots,n$。采用双层级 GAN 结构，使用判别器$D_V$来判别合成视频序列的真实度，判别器$D_I$用于判别每一帧合成图像的真实度。

G_M包含一个 LSTM 编码器$\mathrm{LSTM_{enc}}$，它接收起始帧隐编码$z_1\in Z$生成初始隐状态，LSTM 解码器$\mathrm{LSTM_{dec}}$负责根据初始隐状态预测后续的$n-1$个隐状态：（式中，h 和 c 分别表示隐状态和累积状态，ϵ_t是服从标准正态分布的白噪声，用以生成差异化动作轨迹）

$$\begin{aligned} h_1,c_1&=\mathrm{LSTM_{enc}}(z_1)\\ h_t,c_t&=\mathrm{LSTM_{dec}}(\epsilon_t,(h_{t-1},c_{t-1})),t=2,3,\cdots,n\end{aligned} \tag{2-15}$$

为使G_M专注于动作序列生成，MoCoGAN-HD 采用基于主成分分析（Principal Component Analysis，PCA）的残差技术来生成后续帧隐编码：（式中，V是通过对Z进行m项随机采样，执行 PCA 生成的基，λ表示步长）

$$z_t=z_{t-1}+\lambda\cdot h_t\cdot V,t=2,3,\cdots,n \tag{2-16}$$

通过起始帧隐编码生成动作轨迹序列$G_M=\{z_1,z_2,\cdots,z_n\}$，再通过$G_I$合成视频$\tilde{v}=G_I\big(G_M(z_1)\big)$。

MoCoGAN-HD 模型训练。使用余弦相似度$\mathrm{sim}(u,v)=u^{\mathrm{T}}v/\|u\|\|v\|$作为距离衡量标准，构造动作异化损失：（式中，H 表示一个二层 MLP）

$$L_m = \frac{1}{n-1}\sum_{t=2}^{n}\mathrm{sim}(H(h_t),\epsilon_t) \tag{2-17}$$

以最大化L_m为优化目标，使生成的动作序列具有更多变化，解决$\mathrm{LSTM_{dec}}$的模式崩溃问题。

在判别器D_v和D_I的输出端构造交叉熵损失:（式中，p_v表示真实视频分布，p_z表示图像合成网络的输出分布）

$$L_{D_V} = \mathbb{E}_{v\sim p_v}\left[\log\left(D_V(v)\right)\right] + \mathbb{E}_{z_1\sim p_z}\left[\log\left(1-D_V\left(G_I\left(G_M(z_1)\right)\right)\right)\right] \tag{2-18}$$

$$\begin{aligned} L_{D_I} = {} & \mathbb{E}_{z_1\sim p_z}\left[\log\left(D_I\left(G_I(z_1)\right)\right)\right] \\ & + \mathbb{E}_{z_1\sim p_z, z_t\sim G_M(z_1)|t>1}\left[\log\left(1-D_I\left(G_I(z_t)\right)\right)\right] \end{aligned} \tag{2-19}$$

使用 InfoNCE 损失[24]增强生成视频内容的一致性。对于一批 N 个生成视频$\left\{\tilde{v}^{(1)},\tilde{v}^{(2)},\cdots,\tilde{v}^{(N)}\right\}$，在每个视频中随机抽取一帧得到图像集$\left\{\tilde{x}_t^{(1)},\tilde{x}_t^{(2)},\cdots,\tilde{x}_t^{(N)}\right\}$。然后如图 2-5 右图所示对每个抽取帧进行随机遮挡，生成两个变形$\left(\tilde{x}_t^{(ia)},\tilde{x}_t^{(ib)}\right)$，由此生成 2N 个样本；其中$\left(\tilde{x}_t^{(ia)},\tilde{x}_t^{(ib)}\right)$是正样本组，$\left(\tilde{x}_t^{(i\cdot)},\tilde{x}_t^{(j\cdot)}\right), i\neq j$是负样本组。图 2-5 中的 F 是自学习编码器，它由移除最后输出层的D_I网络加上两层 MLP 构成，其中D_I网络的权重参数与判别器D_I共享，构造对比损失:（式中，$\mathbb{I}_{[j\neq i]}\in\{0,1\}$等于 1 当且仅当$j\neq i$，$\tau$是温度参数，在该模型中被设置为 0.07）

$$\begin{aligned} & L_{contr} \\ & = -\sum_{i=1}^{N}\sum_{\alpha=a}^{b}\log\frac{\exp\left(\mathrm{sim}\left(F\left(\tilde{x}_t^{(ia)}\right),F\left(\tilde{x}_t^{(ib)}\right)\right)/\tau\right)}{\sum_{j=1}^{N}\sum_{\beta=a}^{b}\mathbb{I}_{[j\neq i]}\cdot\exp\left(\mathrm{sim}\left(F\left(\tilde{x}_t^{(ia)}\right),F\left(\tilde{x}_t^{(j\beta)}\right)\right)/\tau\right)} \end{aligned} \tag{2-20}$$

为保障帧间高级语义特征的一致性，引入 CGAN[25]中采用的特征匹配损失函数L_f，但将其中的 L_1 距离替换为余弦相似度。MoCoGAN-HD 的最终优化目标为:（式中，λ_m、λ_f、λ_{contr}是权重参数）

$$\min_{G_M}\left(\max_{D_V}L_{D_V}+\max_{D_I}L_{D_I}\right)+\max_{G_M}\left(\lambda_m L_m+\lambda_f L_f\right)+\min_{D_I}\left(\lambda_{contr}L_{contr}\right) \tag{2-21}$$

由于使用了预训练的图像合成网络以及改进的残差隐编码技术，MoCoGAN-HD 在生成效率、分辨率和质量上都达到了高水平，损失函数改进使得此模型能够生成动作变化更加丰富长视频，并且支持交叉图像域视频合成。如图 2-6 所示，前两行视频帧是利用人类面部表情生成的狗面部表情动态（G_I使用 AFHQ-Dog[26]训练，G_M使用 VoxCeleb[27]训练），后两行视频帧是利用人类面部表情生成的卡通角色面部表情动态（G_I使用 AnimeFaces[28]训练，G_M使用 VoxCeleb[27]训练）。

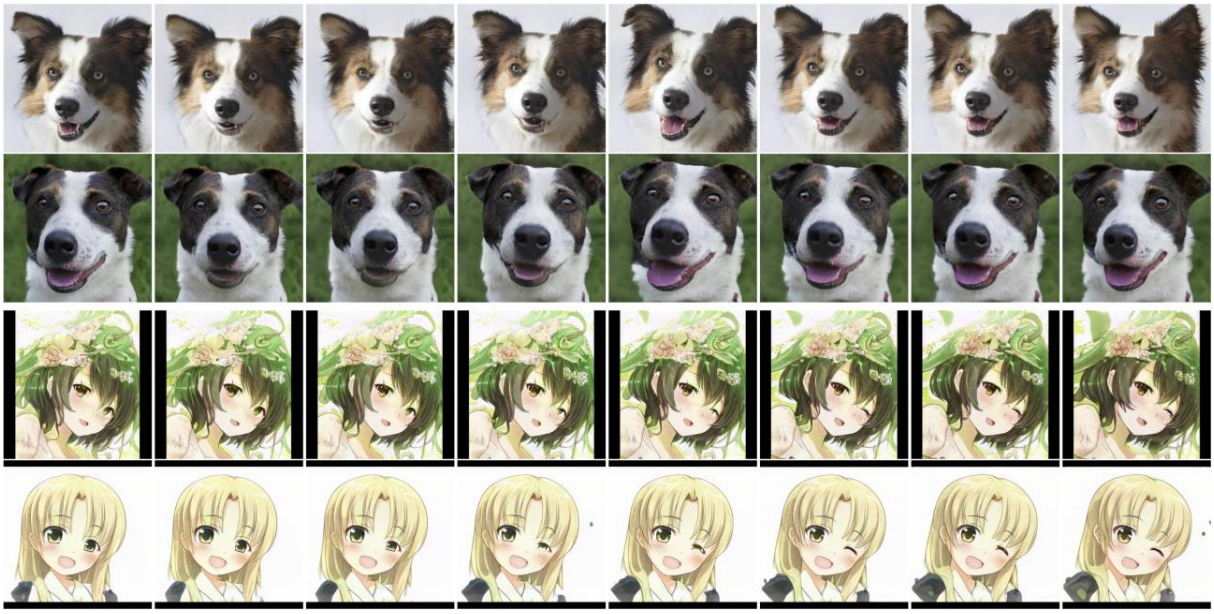

图2-6 MoCoGAN-HD 交叉图像域视频合成

表2-1 视觉合成技术对比

合成技术	输入	输出	风格控制
White-Box	源图像	卡通风格化图像	训练时控制，调节损失函数
Gated Mapping GAN	源图像+参考图像	卡通风格化图像	应用中控制，使用参考图像
MoCoGAN-HD	源图像	视频序列	训练图像合成网络时控制,改变训练集

合成技术	训练集需求	视频合成支持
White-Box	统一风格图像集	逐帧转换，需要源视频
Gated Mapping GAN	混合风格图像集	逐帧转换，需要源视频
MoCoGAN-HD	统一风格图像集,统一风格视频集	仅需起始帧或关键帧，自动合成视频

White-Box、Gated Mapping GAN、MocoGAN-HD 三种视觉合成技术的对比结果如表 2-1 所示。

- White-Box[12]的独特优势在于可通过调节三种特征风格损失的权重以及插值技术实现卡通风格的“白盒”式精细化控制。该模型需要使用统一风格的图像集进行训练，并且生成器的转换风格在训练时确定，如需改变转换风格则必须重新训练生成器。该模型的转换风格忠实于原始内容，对场景图像的转换效果较好，但无法实现夸张风格人物肖像的卡通化。
- Gated Mapping GAN[15]的突出优势在于基于参考的风格控制，只需在混合风格图像集上进行一次训练，就可兼用于场景和人物肖像的卡通风格化。基于参考的风格控制使该模型在实际应用中非常灵活且强大，它能实现夸张风格的人物肖像生成，且能保障夸张风格与写实风格、多种场景风格卡通图像生成之间的互不干扰性，在生成不同风格卡通图像时也无需重新训练。但也正是因为该模型使用混合风格图像集进行训练，缺乏精细化控制手段，导致生成图像风格与参考图像风格对应关系的

可解释性较差，是一种“黑盒”模型。

- MoCoGAN-HD[16]实现了视频合成与交叉图像域转换的一体化，其显著优势在于无需完整源视频即可创作新视频。前述两项技术均需在源视频的基础上执行逐帧转换，合成视频的动态内容完全取决于源视频，而该模型具有更为强大的创作能力，仅需起始帧或关键帧即可合成视频，是三个模型中唯一具备动态创作能力的模型。该模型还是图像合成网络的黏合剂，预训练的 StyleGAN[22]、BigGAN[23]乃至 White-Box、Gated Mapping GAN 均可集成到 MoCoGAN-HD 中。但在易用性上，该模型具有与 White-Box 类似的缺陷，对训练集的需求较高，且改变风格时需要重新训练。

三、存在问题

White-Box[12]、Gated Mapping GAN[15]、MoCoGAN-HD[16]分别代表了可解释、通用化和创作性赛道的时新技术体系，但在易用性、细节质量以及训练需求方面仍存在一些缺陷。

风格切换。White-Box 的风格控制主要通过设定损失权重的方式在训练时完成，虽然可以通过插值后处理的方式进一步微调细节平滑程度，但总体而言风格控制手段较为有限，风格切换的代价较大（需要重新训练）。MoCoGAN-HD 将动作和内容生成分离，一定程度上降低了风格切换的难度，但在更换动作训练集或图像合成器后仍需通过优化训练来生成恰当的结果。Gated Mapping GAN 对风格切换的支持较好，可直接在应用中控制，并可通过增量训练扩充参考风格转换能力。

图3-1 Gated Mapping GAN 合成异常

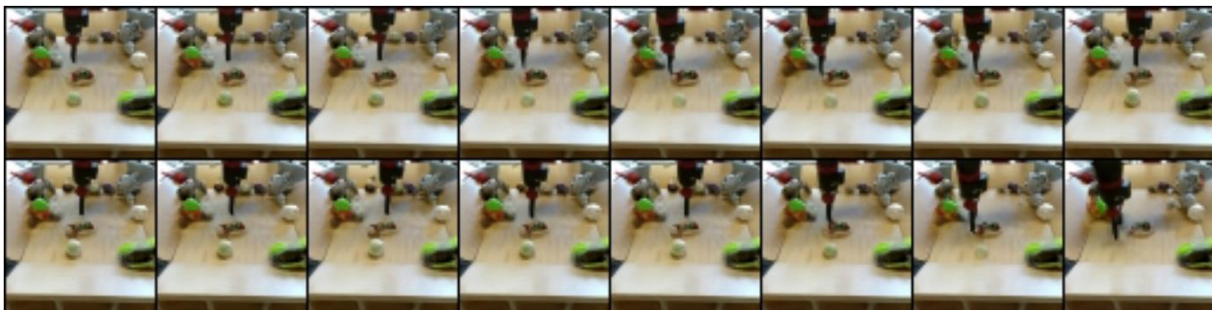

图3-2 MoCoGAN-HD 合成异常（比较首帧和尾帧）

细节合成质量。White-Box 是忠实于源图像的合成技术，细节合成质量高，但无法实现线条扭曲、轮廓变形等夸张艺术效果。Gated Mapping GAN 对人物肖像中的配饰物（眼镜、耳饰等）的转换效果不理想，这类配饰物在合成图像中可能被丢弃或产生干扰

性结果（如图 3-1a）。MoCoGAN-HD 作为一种视频合成模型，更易受到训练数据噪声和多样性干扰，容易引入面积伸缩、边缘变形等异常。

图像域间干扰。White-Box 和 MoCoGAN-HD 对训练集风格统一的需求较高，当训练集风格差异过大时会导致合成质量下降。由于绝大多数人物肖像带有环境背景和非人物的装饰物，Gated Mapping GAN 在转换这类人物肖像时不能很好地将人物肖像和环境分开，这使得环境背景与其它装饰物也被施加人物肖像转换的夸张风格，导致诸多不合理的色块变化和纹理错位（如图 2-3 中示例、图 2-4）。

内容一致性。White-Box 对源图像的内容忠实度较高，并且较好地解决了色彩偏差问题。相较之下，Gated Mapping GAN 为了将源图像迁移到参考图像风格，非常容易产生色差与亮度偏差（如图 2-4）；对于人物肖像转换，在使用自适应风格损失后可增强对源图像内容的忠实度，但色差问题依然存在；对于场景图像转换，如果源图像和参考图像的语义差异过大，Gated Mapping GAN 会生成不真实且不自然的结果（如图 3-1b）。MoCoGAN-HD 生成差异化帧动态的异化需求与内容一致性的保持存在天然矛盾，只能通过调节损失权重来折衷；对于特征稀疏的图像，MoCoGAN-HD 将更加难以分辨图像中的物体对象，导致合成动态中的小物体发生合体、变形或丢失等异常（如图 3-2）。

四、未来的研究热点

模型高效性和易用性改良。卡通风格化方法主要基于卷积生成式对抗网络来实现，此模型具有学习能力强大、支持无监督学习等优点，但也存在训练稳定性差、易出现模式崩溃等问题，这通常需要通过人工调整学习率规划、学习步骤甚至改变损失函数来解决。规模过大的网络模型导致较大的时间开销，尤其是对 White-Box 等一类通过重新训练来实现风格切换的合成技术而言，其实用性大打折扣；这类技术需要通过模型轻量化或迁移学习方法来降低训练开销，或者像 Gated Mapping GAN 那样实现通用式的“一次训练”。Gated Mapping GAN 为通用合成模型的研究提供了思路，但其目前对图像域的划分依然是粗粒度的（仅包含场景和人物肖像），图像域的细粒度划分以及图像域间干扰排除问题仍有待进一步研究。

实用性研究。现代卡通艺术已经从 2D 迈入 3D 时代，在 2D 动画仍保有一定市场占有率的同时，大量动画制作者也开始使用 3D 模型制作动画。多视角采集技术能够以特定旋转角度为步长快速采集一个真实物体或人体的多视角图像，这些多视角图像存在非常显著的空间约束性，空间约束保持的多视角图像卡通风格化（如双目图像卡通风格化）是一个极富挑战性的课题，优质的具有约束保持特性的合成结果（特别是满足三视图约束的结果）对 3D 建模工作具有较强的指导作用。纯 3D 模型卡通风格化旨在对真实物体扫描模型或高真实感雕刻模型进行修改，直接生成 3D 卡通模型；另一条赛道与图形学渲染技术相关，卡通风格化对非真实感渲染技术而言具有非常高的应用价值。在平面卡通领域，现有模型对重度风格化的支持仍不足，这一类风格化需求涉及强烈的纹理扭

曲和结构的器质性变化，例如动物的人貌化（常见于动物拟人动画中）、将拼接成的自然生物图像合成为异想生物图像（常见于科幻、奇幻动画中）等。

创作驱动与风格迁移的泛化。视频创作驱动的风格迁移是另一个极具吸引力的研究领域，其风格迁移不再局限于画面风格，而被拓展到动作风格、运镜风格乃至剪辑风格中，“如何在合成视频帧的同时创作符合卡通动画的画面、动作、运镜风格，使其适配动画表现手法”是一个非常新颖有趣的研究方向。

参考文献

[1] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets[C]// GHAHRAMANI Z, WELLING M, CORTES C, et al. Advances in Neural Information Processing Systems 27. La Jolla, California: Neural Information Processing Systems Foundation, 2014: 2672-2680.

[2] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-To-Image Translation With Conditional Adversarial Networks[C]// IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2017: 1125-1134.

[3] 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, New Jersey: Curran Associates Incorporation, 2017: 700-708.

[4] LEDIG C, THEIS L, HUSZAR F, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network[C]// IEEE. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2017: 105-114.

[5] WANG X T, YU K, WU S X, et al. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks[C]// LEAL-TAIXE L, ROTH S. Computer Vision – ECCV 2018 Workshops: Part V. Cham, Switzerland: Springer International Publishing, 2019: 63-79.

[6] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context Encoders: Feature Learning by Inpainting[C]// IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2016: 2536-2544.

[7] LUGMAYR A, DANELLJAN M, ROMERO A, et al. RePaint: Inpainting using Denoising Diffusion Probabilistic Models[C]// IEEE. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2022: 11451-11461.

[8] YAO K, GAO P L, YANG X, et al. Outpainting by Queries[C]// AVIDAN S, BROSTOW G, CISSE M, et al. Computer Vision – ECCV 2022. Cham, Switzerland: Springer Nature Switzerland, 2022: 153-169.

[9] YI R, LIU Y J, LAI Y K, et al. Unpaired Portrait Drawing Generation via Asymmetric Cycle Mapping [C]// IEEE. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2020: 8217-8225.

[10] CHEN Y, LAI Y K, LIU Y J. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization[C]// IEEE. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2018: 9465-9474.

[11] CHEN J, LIU G, CHEN X. AnimeGAN: A Novel Lightweight GAN for Photo Animation[C]// LI K, LI W, WANG H, et al. Artificial Intelligence Algorithms and Applications. Singapore: Springer Singapore, 2020: 242-256.

[12] WANG T C, LIU M Y, ZHU J Y, et al. Learning to Cartoonize Using White-Box Cartoon Representations[C]// IEEE. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2020: 8090-8099.

[13] YI R, LIU Y J, LAI Y K, et al. APDrawingGAN: Generating Artistic Portrait Drawings From Face Photos With Hierarchical GANs[C]// IEEE. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2019: 10735-10744.

[14] SU H, NIU J W, LIU X F, et al. MangaGAN: Unpaired Photo-to-Manga Translation Based on The Methodology of Manga Drawing[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California: AAAI Press, 2021: 2611-2619.

[15] MEN Y F, YAO Y, CUI M M, et al. Unpaired Cartoon Image Synthesis via Gated Cycle Mapping[C]// IEEE. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2022: 3491-3500.

[16] TIAN Y, REN J, CHAI M L, et al. A Good Image Generator Is What You Need for High-Resolution Video Synthesis[C/OL]// 9th International Conference on Learning Representations. [S.l.]: OpenReview, 2021: 1-23. https://openreview.net/forum?id=6puCSjH3hwA.

[17] GREFF K, SRIVASTAVA R K, KOUTNIK J, et al. LSTM: A Search Space Odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2222-2232.

[18] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]// 3rd International Conference on Learning Representations: ICLR 2015 Conference Track Proceedings. [S.l.]: International Conference on Learning Representations, 2015: 1-14.

[19] FELZENSZWALB P F, HUTTENLOCHER D P. Efficient Graph-Based Image Segmentation[J]. International journal of computer vision, 2004, 59(2): 167-181.

[20] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[21] HUANG X, BELONGIE S. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization[C]// IEEE. 2017 IEEE International Conference on Computer Vision. Los Alamitos, California: IEEE Computer Society, 2017: 1510-1519.

[22] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and Improving the Image Quality of StyleGAN[C]// IEEE. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2020: 8107-8116.

[23] BROCK A, DONAHUE J, SIMONYAN K. Large Scale GAN Training for High Fidelity Natural Image Synthesis [C/OL]// 7th International Conference on Learning Representations. [S.l.]: OpenReview, 2019: 1-35. https://openreview.net/forum?id=B1xsqj09Fm.

[24] VAN DE OORD A, LI Y Z, VINYALS O. Representation Learning with Contrastive Predictive Coding[J/OL]. ArXiv, 2018: 1-13. https://arxiv.org/abs/1807.03748.

[25] WANG T C, LIU M Y, ZHU J Y, et al. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs[C]// IEEE. 2018 IEEE/CVF Conference on

Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2018: 8798-8807.

[26] CHOI Y J, UH Y J, YOO J J, et al. StarGAN v2: Diverse Image Synthesis for Multiple Domains[C]// IEEE. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE Computer Society, 2020: 8185-8194.

[27] NAGRANI A, SHUNG J S, XIE W D, et al. Voxceleb: Large-scale speaker verification in the wild[J]. Computer Speech & Language: Article 101027, 2020, 60: 1-15.

[28] DANBOORU COMMUNITY. Danbooru2019 Portraits: A Large-Scale Anime Head Illustration Dataset[DB/OL]. (2019-03-12)[2023-01-11]. https://www.gwern.net/Crops#danbooru2019-portraits.