基于混合生成对抗网络的多视角图像生成算法

卫 星 [1] 李 佳 [1] 孙 晓 [1] 刘邵凡 [1] 陆 阳 [1]

摘 要 多视角图像生成即基于某个视角图像生成其他多个视角图像, 是多视角展示和虚拟现实目标建模等领域的基本问题, 已引起研究人员的广泛关注. 近年来, 生成对抗网络 (Generative adversarial network, GAN) 在多视角图像生成任务上取得了不错的成绩, 但目前的主流方法局限于固定领域, 很难迁移至其他场景, 且生成的图像存在模糊、失真等弊病. 为此本文提出了一种基于混合对抗生成网络的多视角图像生成模型 ViewGAN, 它包括多个生成器和一个多类别判别器, 可灵活迁移至多视角生成的多个场景. 在 ViewGAN 中, 多个生成器被同时训练, 旨在生成不同视角的图像. 此外, 本文提出了一种基于蒙特卡洛搜索的惩罚机制来促使每个生成器生成高质量的图像, 使得每个生成器更专注于指定视角图像的生成. 在 DeepFashion, Dayton, ICG Lab6 数据集上的大量实验证明: 我们的模型在 Inception score 和 Top-k accuracy 上的性能优于目前的主流模型, 并且在结构相似性 (Structural similarity, SSIM) 上的分数提升了 32.29%, 峰值信噪比 (Peak signal-to-noise ratio, PSNR) 分数提升了 14.32%, SD (Sharpness difference) 分数提升了 10.18%.

关键词 深度学习, 计算机视觉, 图像翻译, 多视角图像生成

引用格式 卫星, 李佳, 孙晓, 刘邵凡, 陆阳. 基于混合生成对抗网络的多视角图像生成算法. 自动化学报, 2021, **47**(11): 2623−2636

DOI 10.16383/j.aas.c190743

Cross-view Image Generation via Mixture Generative Adversarial Network

WEI Xing[1] LI Jia[1] SUN Xiao[1] LIU Shao-Fan[1] LU Yang[1]

Abstract Cross-view image translation aims at synthesizing new image from one to another. It is a fundamental issue in areas such as multi-view presentations and object modeling in virtual reality, which has been gaining a lot interest from researchers around the world. Recently, generative adversarial network (GAN) has shown promising results in image generation. However, the state-of-the-arts are limited to flxed flelds, and it is di–cult to migrate to other scenes, and the generated images are ambiguous and distorted. In this paper, we propose a novel framework-ViewGAN that makes it possible to generate realistic-looking images with difierent views. The ViewGAN can be flexibly migrated to multiple scenarios of the multi-view image generation task, which has multiple generators and one multi-class discriminator. The multiple generators are trained simultaneously, aiming at generating images from difierent views. Moreover, we propose a penalty mechanism based on Monte Carlo search to make each generator focus on generating its own images of a speciflc view accurately. Extensive experiments on DeepFashion, Dayton and ICG Lab6 datasets demonstrate that our model performs better on inception score and top-k accuracy than several state-of-the-arts, and the SSIM (structural similarity) increased by 32.29%, the PSNR (peak signal-to-noise ratio) increased by 14.32%, and the SD (sharpness difference) increased by 10.18%.

Key words Deep learning, computer vision, image translation, multi-view images generation

Citation Wei Xing, Li Jia, Sun Xiao, Liu Shao-Fan, Lu Yang. Cross-view image generation via mixture generative adversarial network. *Acta Automatica Sinica*, 2021, **47**(11): 2623−2636

收稿日期 2019-10-25 录用日期 2020-02-23

Manuscript received October 25, 2019; accepted February 23, 2020

2020 年安徽省自然科学基金联合基金 (2008085UD08), 安徽省重点研发计划项目 (201904d08020008, 202004a05020004), 合肥工业大学智能制造技术研究院智能网联及新能源汽车技术成果转化及产业化项目 (IMIWL2019003, IMIDC2019002) 资助

Supported by Joint Fund of Natural Science Foundation of Anhui Province in 2020 (2008085UD08), Anhui Provincial Key Research and Development Program (201904d08020008, 202004a05020004), and Intelligent Networking and New Energy Vehicle Special Project of Intelligent Manufacturing Institute of Hefei University of Technology (IMIWL2019003, IMIDC2019002)

本文责任编委 吴建鑫

Recommended by Associate Editor WU Jian-Xin

1. 合肥工业大学计算机与信息学院 合肥 230601

1. School of Computer and Information, Hefei University of Technology, Hefei 230601

多视角图像生成指的是基于某个视角的图像生成其他视角的图像, 该问题在实际生活中具有很多应用, 例如: 电商网站上商品的多视角展示、虚拟现实中的目标建模和和数据集扩充等. 目前多视角图像生成已经吸引了来自计算机视觉、虚拟现实等众多领域研究人员的兴趣, 并取得了一定的进展[1−5].

早期工作中, 研究人员尝试使用变分自编码器 (Variational autoencoder, VAE)[6] 生成多视角图像. 变分自编码器采用变分推断和深度表示学习来

得到一个复杂的生成模型, 从而摆脱了传统耗时的采样过程. 但变分自编码器并不能很好地补充生成图像中的细节. 此外, 研究人员[7−8] 还尝试先建立目标的 3D 模型, 再生成目标视角的图像, 但这种方法的灵活性较弱, 只适合于合成椅子、杯子等简单物体的图像.

近年来, 有研究人员提出使用生成对抗网络 (Generative adversarial network, GAN)[9] 来生成多视角图像. 在文献 [4] 中, 研究人员将变分自编码器与生成对抗网络相结合, 提出了一种面向服装的多视角图像生成模型 VariGANs. VariGAN 模型将图像生成分为两步, 采用由粗到精的方式生成高分辨率的多视角图像, 本文模型也参考了这种由粗到精的生成方式. 但 VariGAN 模型局限于服装的多视角图像生成, 并不能有效迁移至其他领域.

在文献 [2] 中, 研究人员尝试在图像生成中引入语义指导, 提出了两种多视角图像生成模型 X-Fork 和 X-Seq. 这两个模型将已知视角的图像与目标视角的语义分割图共同输入模型, 填补了生成图像中的语义结构, 使得生成的图像更加真实. 受到文献 [2] 中工作的启发, 文献 [5] 中的研究人员提出了一种基于多通道注意力机制的 SelectionGAN 模型. SelectionGAN 模型将语义生成空间进一步扩大, 模型通过参考生成的中间结果, 进一步完善了图像中的语义细节, 在卫星图与地面图的翻译任务中取得了很好的成绩. 但以上的工作对于多视角生成任务中其他场景的兼容性较差, 因为并不是所有场景下都有充足的语义分割图来进行训练模型.

为解决上述问题, 本文提出了一种基于混合生成对抗网络的多视角图像生成模型 ViewGAN, 该模型可以灵活迁移至多视角生成任务中的各个场景. ViewGAN 包含多个生成器和一个多类别判别器, 每一个生成器负责生成某一视角的图像. 如图 1 所示, 模型分两步生成图像: 1) 模型运用粗粒度模块 (Coarse image module) 生成低分辨率 (Low resolution, LR) 下的目标图像; 2) 在低分辨率目标图像的基础上, 模型运用细粒度模块 (Fine image module) 完善图像的语义结构, 生成高分辨率 (high resolution, HR) 下的目标图像.

本文的 ViewGAN 模型与以往工作的不同之处在于: 1) ViewGAN 包含多个生成器和一个判别器, 每一个生成器负责生成某一视角的图像, 这保证了 ViewGAN 模型可以灵活迁移至各种多视角生成任务中, 甚至还可以运用到图像翻译的其他领域, 例如风格转换等; 2) 为了加强图像生成过程中的语义约束, 本文使用蒙塔卡罗搜索方法 (Monte Carlo search, MCS)[10] 对低分辨率目标图像进行多次采样, 并根据采样结果计算相应的惩罚值, 惩罚机制

图 1 本文模型 ViewGAN 在 DeepFashion、Dayton 和 ICG Lab6 数据集上的测试样例

Fig. 1 Examples of ViewGAN on three datasets, i.e., DeepFashion, Dayton and ICG Lab6

可以迫使每个生成器生成语义更加丰富的图像, 避免出现模式崩塌 (Mode collapse)[11]; 3) 模型中的多类别判别器使每个生成器更加专注于生成它们指定视角的图像, 避免生成与其他视角相似的图像, 从而进一步完善了图像的语义结构; 4) 本文将 ViewGAN 模型与目前主流的图像生成模型 (例如: Pix2Pix[12], VariGAN[4], X-Fork 和 X-Seq[2], SelectionGAN[5]) 进行了对比, 并在 3 个公开数据集上进行了大量的实验, 实验结果表明: 本文模型在 3 个数据集上都取得了最好成绩, 这表明了本文模型的灵活性和生成图像的高质量.

综上所述, 本文的主要贡献总结如下:

1) 提出了一种基于混合生成对抗网络的多视角图像生成模型 ViewGAN, 该模型包括多个生成器和一个判别器, 采用由粗到精的方式生成不同视角下的高质量图像.

2) 提出了一种基于蒙特卡洛搜索的惩罚机制来加强图像生成过程中的约束, 这使得每个生成器能够获得更充足的语义指导, 在对应视角的图像中增加更多的语义细节.

3) 在 3 个数据集上与目前的主流模型进行了大量的对比实验, 实验结果证明了 ViewGAN 在各种场景下的有效性与灵活性.

1 相关工作

1.1 图像生成

随着深度学习技术的发展, 图像生成已经变成了一个热门的话题. 变分自编码器 (VAE)[6]是一种基于概率图模型的生成模型. 在文献 [13] 中, 研究人员提出了一种可由视觉特征生成图像的 Attrib-

ute2Image 模型, 该模型通过合成前景和背景来建模图像. 之后研究人员[14] 尝试在 VAE 中引入注意力机制, 提出了一种 DRAW 模型, 该模型在一定程度上提升了图像的质量.

近年来, 研究人员在采用生成对抗网络[9] 在图像生成方向取得了不错的成绩, 生成器被训练生成图像来欺骗判别器, 判别器被训练区分真实图像和虚假图像. 之后大量基于 GAN 的变体相继提出, 例如条件生成对抗网络 (Conditional GANs)[15]、Bi-GANs[16]、InfoGANs[17] 等. GANs 还可以基于标签[15], 文本[18−19] 或者图[12, 20−22] 来生成图像.

但上述模型生成的图像普遍存在模糊、失真等问题, 模型并没有学会如何生成图像, 而是简单地重复训练集中图像的内容. 本文模型也是一种基于输入图像的条件生成对抗网络, 但本文模型凭借新颖的惩罚机制引入了更充足的语义指导, 进一步完善了图像的语义结构, 在一定程度上克服了图像失真的问题.

1.2　多视角图像生成

早期的研究人员通过对物体进行 3D 建模来合成不同视角的图像[7−8, 23]. 在文献 [8] 中, 研究人员提出一种 3D−2D 映射机制, 从而使模型可以从 2D 数据中学习到 3D 特征. 之后有研究人员[23]提出一种 3D-GAN 模型, 该模型可以依据复杂的概率空间生成 3D 物体.

在文献 [2] 中, 研究人员尝试使用图像翻译的方法进行多视角生成, 使用条件生成对抗网络在卫星图-街景图转换任务中取得了不错的成绩. 之后在文献 [4] 中, 研究人提出了一种面向服装的多视角图像生成模型 VariGANs. VariGAN 模型将图像生成分为两步, 采用由粗到精的方式生成高分辨率的多视角图像. 受到上述工作的启发, 文献 [5] 中的研究人员提出了一种基于多通道注意力机制的 SelectionGAN 模型. SelectionGAN 模型将语义生成空间进一步扩大, 模型通过充分参考生成的中间结果, 进一步完善了图像中的语义细节, 在卫星图与街景图的翻译任务中取得了很好的成绩.

但上述模型对于数据的要求极为严格, 模型需要大量的数据或者辅助的语义分割图进行训练, 这大大限制了模型的灵活性和兼容性. 为解决这个问题, 本文提出了一种基于混合生成对抗网络的多视角图像生成模型, 本文模型包括多个生成器和一个判别器, 这使得模型可以同时训练生成多个视角的图像. 大量实验结果证明, 在不需要大量的数据或者语义分割图辅助训练的前提下, 本文模型在 3 个数据集上都取得了不错的成绩.

2　GAN 的相关背景

生成对抗网络 (GAN)[9] 包括两个对抗学习的子网络: 一个生成器和一个判别器, 它们基于最大−最小博弈理论同时进行训练. 生成器 G 的目的在于将一个 d 维的噪声向量映射成一幅图像, 并尽可能地使生成的图像接近真实图像; 另一方面, 判别器 D 用来鉴别图像是来自于生成器的虚假图像还是来自真实数据的真实图像. 整个生成对抗网络的目标函数可表示为

$$\min_G \max_D L_{\mathrm{GAN}}(G,D) = \mathrm{E}_{x\sim p_{\mathrm{data}}(x)}[\ln D(x)] + \mathrm{E}_{z\sim p_z(z)}[\ln(1-D(G(z)))] \tag{1}$$

其中, x 表示由真实数据 p_{data} 采样得到的真实数据, z 表示从高斯分布 p_z 采样得到的 d 维噪声向量.

条件生成对抗网络 (Conditional GANs)[15] 通过引入辅助变量来控制模型的生成结果. 在条件生成对抗网络中, 生成器基于辅助信息生成图像, 判别器基于辅助信息和图像 (虚假图像或者真实图像) 做出判断. 整个网络的目标函数可表示为

$$\min_G \max_D L_{\mathrm{CGAN}}(G,D) = \mathrm{E}_{x,c\sim p_{\mathrm{data}}(x,c)}[\ln D(x,c)] + \mathrm{E}_{x',c\sim p_{\mathrm{data}}(x',c)}\left[\ln\left(1-D\left(x',c\right)\right)\right] \tag{2}$$

其中, c 表示引入的辅助变量, $x'=G(z,c)$ 表示生成器生成的图像.

除对抗损失外, 以往的工作[20−21] 还尝试最小化真实图像和虚假图像之间的 L1 或者 L2 距离, 这能够帮助生成器合成与真实图像更加相似的图像. 以往工作证明: 相比于最小化 L2 距离, 最小化 L1 距离更能够帮助模型减少图像中的模糊和失真, 因此我们在本文模型中也使用了 L1 距离. 最小化 L1 距离可表示为

$$\min_G L_{\mathrm{L1}}(G) = \mathrm{E}_{x,x'\sim p_{\mathrm{data}}(x,x')}\left[\|x-x'\|_1\right] \tag{3}$$

因此这类条件生成对抗网络的目标函数就是式 (2) 和式 (3) 之和.

3　ViewGAN 模型

3.1　任务定义

这里首先对多视角生成任务进行简短的定义. 假设有一个多视角集合 $V=\{v_1,\cdots,v_i,\cdots,v_n\}$, 其中 v_i 对应某一具体视角, 例如正面或者侧面. 一个物体在视角 v_i 下的图像定义为 I_{v_i}. 给定某个视角的图像 I_{v_i}, 多视角图像生成任务是指生成其他不同视角的图像 I_{v_j}, 其中 $v_j\in V, j\neq i$.

3.2 整体框架

本文提出的 ViewGAN 模型的整体框架如图 2 所示. 假设我们要生成 k 种不同视角的图像 (例如: 正面、侧面和背面三种视角), 我们使用 k 个生成器 $\{G_i\left(I_{g_i};\theta_g^i\right)\}_{i=1}^{i=k}$ 和一个判别器 $D\left(\left(I_{v_i},I_{v_j}\right);\theta_d\right)$, 其中 θ_g^i,θ_d 分别表示第 i 个生成器和判别器的网络参数. I_{v_i},I_{v_j} 表示从真实数据分布中采样得到的某一视角的图像, I_{g_i} 表示第 i 个生成器 G_i 生成的图像.

整个模型的训练可以分为两个对抗学习过程: 生成器的学习和判别器的学习. 第 i 个生成器 G_i 的目标是生成视角 v_i 下的图像, 并使得生成的图像能够欺骗判别器. 换句话说, 生成器的目标在于最小化合成图像与真实图像之间的距离. 与之相反, 判别器的目的在于尽可能的区分 k 种视角下的虚假图像和真实图像, 并针对虚假图像计算出准确的惩罚值.

3.3 多粒度生成器

生成器的整体框架如图 3 所示, 图中展示的是第 j 个生成器. 训练时中生成图像的过程分为 3 步: 1) 输入已知视角 v_i 下的图像 I_{v_i} 和目标视角 v_j 下的图像 I_{v_j}, 生成器首先使用粗粒度生成模块生成低分辨率的目标图像 I'_{g_j}; 2) 采用蒙特卡洛搜索策略对低分辨率目标图像进行 N 次采样, 从而得到 N 幅中间结果图像; 3) 引入注意力机制提取 N 幅中间结果图像的特征, 并将注意力机制的输出与已知视角 v_i 下的图像 I_{v_i} 输入到细粒度生成模块中, 细粒度生成模块生成最终结果, 即高分辨率的目标图像 I_{g_i}.

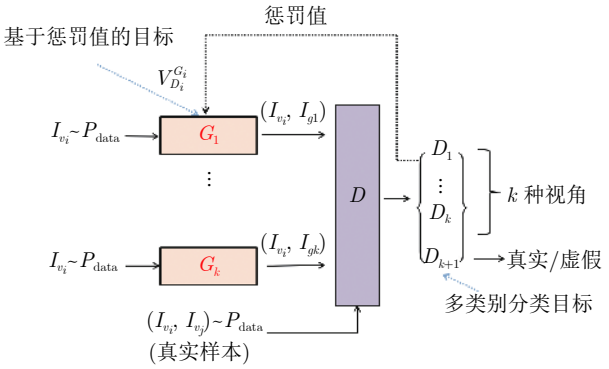

图 2 ViewGAN 模型的整体框架

Fig. 2 The framework of ViewGAN

3.3.1 惩罚机制

为加强图像生成过程中的语义约束, 进一步完善图像中的语义细节, 我们提出了一种基于蒙特卡洛搜索策略的惩罚值机制. 它使混合生成对抗网络中每个生成器更加专注于生成相应视角的图像.

我们对粗粒度生成模块输出的低分辨率目标图像 I'_{g_j} 进行 N 次蒙特卡洛搜索, 具体过程为

$$\left\{I_{g_j}^1,\cdots,I_{g_j}^N\right\}=MC^{G_\beta}\left(I'_{g_j};N\right) \tag{4}$$

$$I_{g_j}^i=G_\beta(I'_{g_j},z) \tag{5}$$

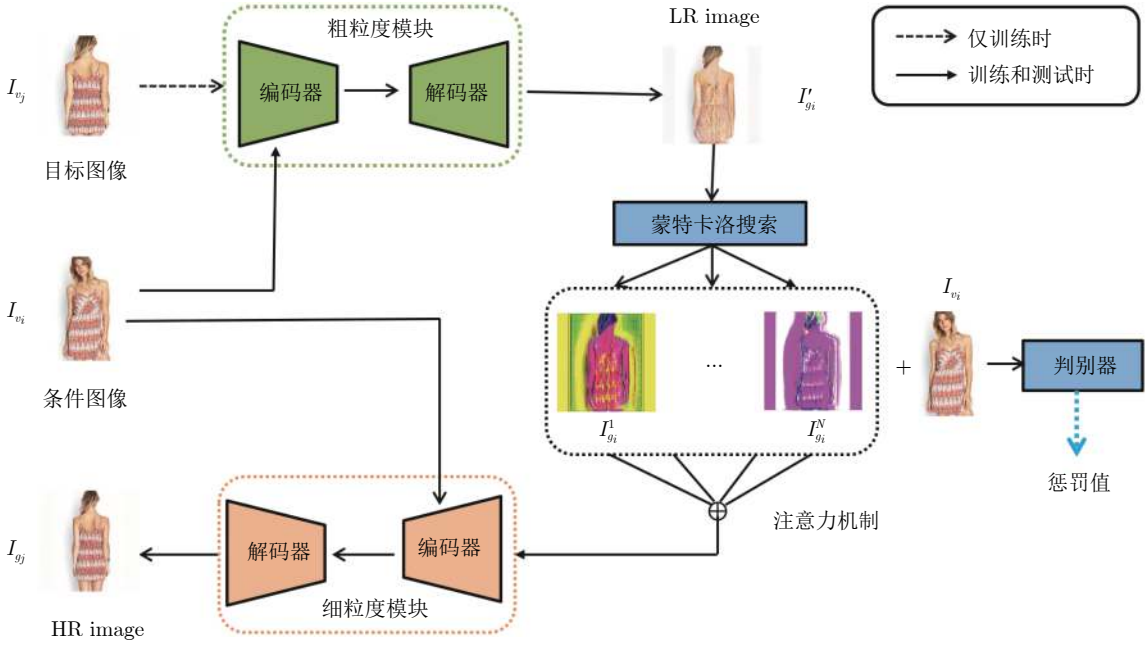

图 3 生成器 (G_j) 的整体框架

Fig. 3 The framework of the generator G_j

其中, $I_{g_j}^i$ $(i=1,\cdots,N)$表示采样得到的图像, MC^{G_β} 表示使用蒙特卡洛搜索模拟的状态. G_β 表示通过蒙特卡洛搜索技术虚拟出的另一个生成模块, 它与细粒度生成模块共享参数, 连续生成 N 次得到 N 张中间结果图像. 其中 $z\sim \mathrm{N}(0,1)$ 是为保证采样结果多样性而引入的噪声向量, 我们将编码器的输出与噪声向量之和作为解码器的输入, 噪声向量使得每次采样时模型对特征向量的着重点不同, 从而从低分辨率目标图像中挖掘出更细致、丰富的语义细节.

之后, 我们将 N 张中间结果图像和已知视角的图像送入判别器, 根据判别器的输出结果计算惩罚值. 计算过程可表示为

$$V_{D_j}^{G_j}=\frac{1}{N}\sum_{i=1}^{N}\left(1-D_j\left(I_{g_j}^i;\theta_d\right)\right)+\left\|I_{g_j}-I_{v_j}\right\|_1 \tag{6}$$

其中, I_{g_j} 表示第 j 个生成器生成的最终结果, I_{v_j} 表示对应的真实目标图像. $D_j(I_{g_j}^i;\theta_d)$ 表示多类别判别器输出的概率分布中图像 $I_{g_j}^i$ 是视角 v_j 下的真实图像的概率.

3.3.2　注意力机制

通过采样得到 N 幅中间结果图像后, 我们希望参考中间结果图像为下一步的生成提供充足的语义指导. 因此我们提出一种基于多通道的注意力机制, 区别于以往工作中合成图像仅从 RGB 三通道空间中生成的方法, 我们将采样得到的 N 幅中间结果作为特征集来构建更大的语义生成空间, 模型通过参考不同通道的信息来提取更加细粒度的信息. 并将计算结果输入到细粒度生成模块中, 从而得到高分辨率目标图像.

首先在 N 幅中间结果图像上进行一次卷积操作来得到相应的注意力权重矩阵 $\left\{I_A^i\right\}_{i=1}^N$, 计算结果可表示为

$$I_A^i=\mathrm{Softmax}\left(I_{g_j}^iW_A^i+b_A^i\right),\ i=1,\cdots,N \tag{7}$$

其中, $\left\{W_A^i,b_A^i\right\}_{i=1}^N$表示卷积操作的参数, Softmax$(\cdot)$ 表示用于归一化的 Softmax 函数. 最终, 将得到的注意力权重矩阵与相应的中间结果图像相乘, 从而得到最终输出为

$$I_{g_j}''=\left(I_A^1\otimes I_{g_j}^1\right)\oplus\cdots\oplus\left(I_A^N\otimes I_{g_j}^N\right) \tag{8}$$

其中, I_{g_j}'' 表示注意力机制的最终输出结果, 符号 $\oplus$ 和 $\otimes$ 分别表示矩阵的元素加法和元素乘法.

3.3.3　目标函数

综上所述, 生成器通过最小化以下目标函数来不断优化

$$J_{G_j}\left(\theta_g^j\right)=\mathrm{E}_{\left(I_{v_i},I_{v_j}\right)\sim p_{\mathrm{data}}}\left[\ln\left(1-D_{k+1}\left(I_{v_i},G_j\left(I_{v_i},I_{v_j}\right)\right)\right)\right]+V_{G_j}^{D_j} \tag{9}$$

其中, $D_{k+1}\left(I_{v_i},G_j\left(I_{v_i},I_{v_j}\right)\right)$表示生成器 G_j 生成的图像被判别器认为是虚假图像的概率, θ_g^j 表示第 j 个生成器 G_j 的参数. $V_{G_j}^{D_j}$ 表示使用式 (5) 得到的针对生成器 G_j 的惩罚值.

3.4　多类别判别器

参考文献 [24] 中有关半监督学习的工作, 我们使用了一种多类别判别器用来区分不同视角下的真实图像和虚假图像, 判别器的输入包括已知视角的图像和目标视角的图像.

3.4.1　目标函数

模型中包含 k 个生成器, 每个生成器负责生成某一视角下的图像, 所以判别器要输出 $k+1$ 种类的概率分布. 前 k 个类别的概率值 D_i $(i\in 1,\cdots k)$ 分别表示输入图像属于第 i 种视角的概率, 第 $k+1$ 个类别的概率值表示输入图像是虚假图像的概率. 判别器通过最小化以下目标来进行优化

$$J_D\left(\theta_d\right)=-\mathrm{E}_{I_v\sim P_{\mathrm{data}},I_g\sim P_g}\ln D_{k+1}\left(\left(I_v,I_g\right);\theta_d\right)-\sum_{i=1}^{k}\mathrm{E}_{\left(I_v,I_{v_i}\right)\sim P_{\mathrm{data}}}\ln D_i\left(\left(I_v,I_{v_i}\right);\theta_d\right)+\lambda L_c \tag{10}$$

其中, P_g 表示生成器生成的图像, P_{data} 表示真实数据分布. I_v 表示任一视角的真实图像, I_g 表示以 I_v 作为输入时生成器生成的图像, I_{v_i} 表示视角 v_i 下的真实图像, $D_i(\cdot)$ 表示判别器输出结果中第 i 个类别的概率值, L_c 表示类内损失. 下面详细阐述类内损失的含义, 参数 λ 用于控制两种损失的平衡.

下面详细论述多类别判别器是如何协助每个生成器更专注于生成指定视角的图像, 避免出现图像模糊或者模式崩塌现象. 为了便于描述, 这里使用 X 表示采样得到的图像对, 例如虚假图像对 (I_v,I_g) 或者真实图像对 (I_v,I_{v_i}).

首先, 理想情况下第 i 个生成器可以学习到视角 v_i 下图像的真实分布 P_{r_i}. 判别器的目标函数如式 (9) 所示, 且 $\sum_{i=1}^{k+1}D_i=1, D_i\in[0,1],\forall i$. 由此可以得到判别器学习到的最优分布:

$$D_i\left(X;\theta_d\right)=\frac{P_{r_i(X)}}{\sum\limits_{i=1}^{k}P_{r_i}(X)+P_g(X)},\ i=1,\cdots,k$$

$$D_{k+1}\left(X;\theta_d\right)=\frac{P_g(X)}{\sum\limits_{i=1}^{k}P_{r_i}(X)+P_g(X)}$$

基于式 (5), 生成器的目标是最小化以下等式:

$$\begin{aligned}&\mathrm{E}_{X\sim P_g}\ln\left[\frac{P_g(X)}{P_{\text{avg}}(X)}\right]+\sum_{i=1}^{k}\mathrm{E}_{X\sim P_{r_i}}\ln\left[\frac{P_{r_i}(X)}{P_{\text{avg}}(X)}\right]-\\&(k+1)\ln(k+1)=\\&KL\left(\sum_{i=1}^{k}P_{g_i}(X)\|P_{\text{avg}}(X)\right)+\\&\sum_{i=1}^{k}KL\left(P_{r_i}(X)\|P_{\text{avg}}(X)\right)-\\&(k+1)\ln(k+1)\end{aligned}\tag{11}$$

其中, $P_{\text{avg}}(x)=\dfrac{\sum_{i=1}^{k}P_{r_i}(X)+P_g(X)}{k+1}$, KL 表示 KL 散度 (Kullback-Leibler divergence). 以上目标函数当 $P_{g_i}=P_{r_i}$ $(i=1,\cdots,k)$ 时取得全局最小值 $-(k+1)\times\ln(k+1)$. 当只有一个生成器时 $(k=1)$, 式 (6) 以 JS 散度 (Jensen-Shannon divergence) 取得最小值 $-\ln 4$.

其次, 当保持判别器参数 θ_d 不变时, 第 i 个生成器旨在最小化由判别器给出的惩罚值 $V_{D_i}^{G_i}$. 由于在判别器中 $\sum_{i=1}^{k+1}D_i=1$, 所以生成器等同于最小化 $\sum_{j=1,j\neq i}^{k+1}D_j(X;\theta_d)$. 换句话说, 为了降低判别器给出的惩罚值, 第 i 个生成器生成的视角 v_i 的图像必须与视角 v_i 的真实图像保持一致, 生成器将尽可能的完善图像的语义结构, 与其他视角的图像保持距离.

3.4.2 类内损失

判别器要对多个生成器生成的视角图像进行判断, 而每个视角的图像之间具有一定的重合部分, 仅采用 GAN 网络对抗损失进行监督学习容易导致图像的视角类别预测错误, 因此我们在对抗损失的基础上引入了类内损失用于减小类内特征差异, 并增加类间差异. 类内损失可表示为

$$L_c=\frac{1}{2}\sum_{i=1}^{m}\left\|I_g^{(i)}-c_{I_g}\right\|_2^2\tag{12}$$

理想情况下, c_{I_g} 应该在每次迭代更新特征时改变, 并使用整个训练集来计算特征中心, 但是由于训练集数量庞大, 这样计算的效率很低, 因此采用一个 Batch 的数据来代替整个训练集的数据. 每次特征中心位置的变化增量如式 (12) 所示, 其中, $I_g^{(i)}$ 表示一个 Batch 中第 i 张生成的图像, v_i 表示图像 $I_g^{(i)}$ 对应的视角, c_j 表示视角 v_j 下样本的特征中心

$$\Delta c_j=\frac{\sum_{i=1}^{m}\delta(v_i=v_j)\left(c_j-I_g^{(i)}\right)}{1+\sum_{i=1}^{m}\delta(v_i=v_j)}\tag{13}$$

其中, $\delta(\cdot)$ 是条件表达式, 当 $v_i=v_j$ 即当前图像 $I_g^{(i)}$ 属于视角 v_j 时, 条件 $\delta(v_i=v_j)$ 为 1, 其他情况下条件 $\delta(v_i=v_j)$ 为 0.

训练时将 k 个生成器与多类别判别器进行对抗学习, 并交替训练它们, 优化算法如算法 1 所示.

算法 1. ViewGAN 的对抗学习过程

输入. k 种视角下的真实图像数据集: $I_v=\{I_{v_1},\cdots,I_{v_k}\}$; 生成器: $\{G_i\}_{i=1}^{i=k}$; 判别器: D; g-step: 生成器的训练步长; d-step: 判别器的训练步长.

输出. 训练完成的生成器: $\{G_i\}_{i=1}^{i=k}$;

初始化生成器 $\{G_i\}_{i=1}^{i=k}$ 和判别器 D 随机权重.
repeat
  for g-steps **do**
    for i in $1\sim k$ **do**
      利用生成器 G_i 生成虚假图像;
      使用式 (5) 计算惩罚项 $V_{G_i}^{D_i}$;
      通过最小化式 (8) 更新生成器 G_i 的参数;
    end for
  end for
  for d-steps **do**
    使用生成器 $\{G_i\}_{i=1}^{i=k}$ 生成虚假图像 $I_g=\{I_{g_1},\cdots,I_{g_k}\}$;
    使用真实图像 $I_v=\{I_{v_1},\cdots,I_{v_k}\}$ 和虚假图像 $I_g=I_{g_1},\cdots,I_{g_k}$ 通过最小化式 (9) 更新判别器参数;
  end for
until ViewGAN 收敛;
return

算法 1 中, $I_g^{(i)}$ 表示一个 batch 中第 i 张生成的图像, $c_{I_g}\in\mathbf{R}^d$ 表示当前图像 $I_g^{(i)}$ 的特征中心, 其中, d 表示特征的维度, 共有 k 个特征中心分别对应 k 种视角图像.

3.5 网络结构

对于多视角生成问题, 输入和输出之间存在大量底层特征共享, 应该直接将特征在网络之间传递. 为了解决网络特征传递问题, 我们采用 U-Net[25] 作为生成器和判别器的基础结构并使用 Conv-BN-ReLu 模块作为中间结构, 网络结构如表 1 和表 2 所示. 其中 CONV BLOCK 卷积块由 3 个串联的卷积核大小为 3 的卷积层和滤波器大小为 2 的平均池化层组成, 其中卷积层的步长等于 1 并采用 1 个像素填充; DECONVBLOCK 由 2 个串联的卷积核大小为 3 的卷积层和滤波器大小为 2 的上采样层组成, 卷积层的设置与 CONV BLOCK 相同; HIDDEN LAYER 由 1 个大小为 3 的卷积核的卷积层组成.

表 1　生成器网络结构
Table 1　Generator network architecture

网络	层次	输入	输出
Down-Sample	CONV(*N*64, *K*4×4, *S*1, *P*3)-BN-Leaky Relu	(256, 256, 3)	(256, 256, 64)
	CONV(*N*128, *K*4×4, *S*2, *P*1)-BN-Leaky Relu	(256, 256, 64)	(128, 128, 128)
	CONV(*N*256, *K*4×4, *S*2, *P*1)-BN-Leaky Relu	(128, 128, 128)	(64, 64, 256)
	CONV(*N*512, *K*4×4, *S*2, *P*1)-BN-Leaky Relu	(64, 64, 256)	(32, 32, 512)
Residual Block	CONV(*N*512, *K*4×4, *S*1, *P*1)-BN-Leaky Relu	(32, 32, 512)	(32, 32, 512)
	CONV(*N*512, *K*4×4, *S*1, *P*1)-BN-Leaky Relu	(32, 32, 512)	(32, 32, 512)
	CONV(*N*512, *K*4×4, *S*1, *P*1)-BN-Leaky Relu	(32, 32, 512)	(32, 32, 512)
	CONV(*N*512, *K*4×4, *S*1, *P*1)-BN-Leaky Relu	(32, 32, 512)	(32, 32, 512)
	CONV(*N*512, *K*4×4, *S*1, *P*1)-BN-Leaky Relu	(32, 32, 512)	(32, 32, 512)
Up-Sample	DECONV(*N*256, *K*4×4, *S*2, *P*1)-BN-Leaky Relu	(32, 32, 512)	(64, 64, 256)
	DECONV(*N*128, *K*4×4, *S*2, *P*1)-BN-Leaky Relu	(64, 64, 256)	(128, 128, 128)
	DECONV(*N*64, *K*4×4, *S*1, *P*3)-BN-Leaky Relu	(128, 128, 128)	(256, 256, 64)
	CONV(*N*3, *K*4×4, *S*1, *P*3)-BN-Leaky Relu	(256, 256, 64)	(256, 256,3)

表 2　判别器网络结构
Table 2　Discriminator network architecture

网络	层次	输入	输出
Input Layer	CONV(*N*64, *K*3×3, *S*1, *P*1)-Leaky Relu	(256, 256, 3)	(256, 256, 64)
	CONV BLOCK	(256, 256, 64)	(256, 256, 64)
	CONV BLOCK	(256, 256, 64)	(128, 128 128)
	CONV BLOCK	(128, 128 128)	(64, 64 256)
Inner Layer	HIDDEN LAYER	(64, 64 256)	(32, 32 512)
	HIDDEN LAYER	(32, 32 512)	(32, 32 64)
	DECONV BLOCK	(32, 32 64)	(64, 64 64)
	DECONV BLOCK	(64, 64 64)	(128, 128, 64)
	DECONV BLOCK	(128, 128, 64)	(256, 256, 64)
Output layer	CONV(*N*64, *K*3×3, *S*1, *P*1)-Leaky Relu CONV(*N*64, *K*3×3, *S*1, *P*1)-Leaky Relu CONV(*N*3, *K*3×3, *S*1, *P*1)-Tanh	(256, 256, 64)	(256, 256, 3)

U-Net 是一种在编码器和解码器之间具有跳跃连接的网络, 这种网络结构允许更多的原图像纹理的信息在高层特征层中进行传播. 对于每个生成器 G_i , 编码器第一层卷积层滤波器大小为 64, 解码器最后一层卷积层用于将通道数映射为输出图片通道数 (彩色图片为 3 通道, 黑白图片为 2 通道). 除了编码器的第一层卷积层, 其余所有卷积层后都连接 BatchNorm 层进行特征归一化.

4　实验与分析

4.1　数据集

1) DeepFashion[26]. 该数据集包含 8 697 幅服饰的多视角图像, 每件服饰具有三个视角: 正面、侧面和背面. 从中挑选出 6 000 幅图像作为训练集, 2 000 幅图像作为测试集, 图像尺寸为 256×256 像素.

2) Dayton[27]. 该数据集包含超过 13 万幅街道视角−鸟瞰视角的图像, 从中挑选出 55 000 幅图像作为训练集, 5 000 幅图像作为测试集. 图像的原始尺寸为 354×354 像素, 我们将图像尺寸调整为 256×256 像素.

3) ICG Lab6[28]. 该数据集包含 6 名人员的室内场景活动图, 共使用 4 个不同方位的静态摄像头进行拍摄. 从中挑选 6 000 幅图像作为训练集, 1 500 幅图像作为测试集. 图像的原始尺寸为 1 024×768 像素, 我们将图像尺寸调整为 256×256 像素.

4.2 基线模型

我们将本文模型与目前主流的多视角图像生成、图像翻译模型进行对比.

1) Pix2Pix[12]. 采用对抗损失学习从 $x \in X$ 到 $y \in Y$ 的映射, 其中, x 和 y 分别表示不同域 X 和 Y 中的图像, 在图像翻译任务上取得了较好成绩.

2) X-Fork[2]. 与 Pix2Pix 结构类似, 生成器通过学习映射: $G:\{I_a\} \rightarrow \{I_b, S_b\}$ 来生成多视角图像. 其中 I_a, I_b 分别表示视角 a 和视角 b 下的图像, S_a 表示视角 b 下的语义分割图.

3) X-Seq[2]. 两个 CGAN($G1$,$G2$) 的组合模型, 其中 $G1$ 合成目标视角的图像, $G2$ 基于 $G1$ 的输出图像合成目标视角的语义分割图. 两个生成器之间的输入–输出依赖约束了生成的图像和语义分割图, 有效地提升了图像的质量.

4) VariGAN[4]. 变分自编码器和 GAN 网络的组合模型, 采用由粗到精的方法生成高分辨率的多视角图像, 在 DeepFashion 等服饰数据集上取得了较好的结果.

5) SelectionGAN[5]. 在 X-Seq 模型的基础上引入了一种多通道注意力机制来选择性地学习模型的中间结果, 从而实现了一种由粗到精的级联式语义指导, 使合成图像具有更丰富的语义细节.

4.3 定量评估

在定量实验中, 我们采用 Inception score[24], Top-k 预测准确率指标从高层特征空间的角度来评估合成图像. 此外, 我们还采用一些像素级别的相似度指标来衡量生成的图像, 包括: 结构相似性 (Structural similarity, SSIM)、峰值信噪比 (Peak signal-to-noise ratio, PSNR) 和 SD (Sharpness difference).

4.3.1 Inception score 和 Top-k 预测准确率

1) Inception score 指标. Inception score (IS) 是一种面向生成模型的常见定量指标, 它可以衡量模型生成的图像是否清晰、生成的图像是否多样. 其计算式为

$$IS(G) = \exp\left(\mathrm{E}_{x\sim p_g} D_{\mathrm{KL}}(p(y|x)\|p(y))\right) \tag{14}$$

其中, G 表示生成器, x 表示生成的图像, y 表示合成图像的预测标签.

因为 Dayton 数据集和 DeepFashion 数据集中包含了 ImageNet 数据集[29] 中未标注的户外物体类别和服装类别, 所以不能直接使用预训练的 Inception 模型. 对于 Dayton 数据集, 使用在 Places 数据集[30] 上训练的 AlexNet 模型[31] 进行评分; 对于 DeepFashion 数据集, 使用预训练的 PaperDoll[32] 模型进行评分; 对于 ICG Lab6 数据集, 采用在 ImageNet 数据集上训练的 Inception 模型进行评分.

同时我们注意到: 这些预训练模型针对合成图像输出的置信度分数较为分散, 合成图像并没有包含所有类别的目标. 因此我们只在 Top-1 和 Top-5 类别上计算 Inception score, 其中 "Top-1"表示每幅图像的预测标签中概率最大的 k 个标签不变, 其余标签的概率进行平滑处理.

2) Inception score 评估结果. 基于 Inception score 的实验结果如表 3 所示. 从实验结果可以看出: 本文模型 ViewGAN在 DeepFashion 数据集和 ICG Lab6 数据集上均优于基线模型. 其中 DeepFashion 数据集的图像风格、服装样式等变化较大, 以往模型很难生成这种多样性较强的图像, 而本文模型通过采用分布式生成的方法, 使模型有足够的内存来学习如何生成各种样式的服装以及同一服装不同视角下的变化. ICG Lab6 数据集的图像取自复杂的室内环境, 对图像分辨率的要求较高. 以往模型缺乏对图像细节的补充, 导致生成模糊、失真的图像, 而本文模型采用惩罚机制加强了对图像语义细节的约束, 能够生成更加清晰的高质量图像.

同时我们注意到 ViewGAN 在 Dayton 数据集

表 3 各模型 Inception score 统计表, 该指标越高表明模型性能越好

Table 3 Inception score of different models (For this metric, higher is better)

模型	DeepFashion			Dayton			ICG Lab6		
	all classes	Top-1 class	Top-5 class	all classes	Top-1 class	Top-5 class	all classes	Top-1 class	Top-5 class
Pix2Pix	3.37	2.23	3.44	2.85	1.93	2.91	2.54	1.69	2.49
X-Fork	3.45	2.57	3.56	3.07	2.24	3.09	4.65	2.14	3.85
X-Seq	3.83	2.68	4.02	2.74	2.13	2.77	4.51	2.05	3.66
VariGAN	3.79	2.71	3.94	2.77	2.19	2.79	4.66	2.15	3.72
SelectionGAN	3.81	2.72	3.91	3.06	**2.27**	3.13	5.64	2.52	4.77
ViewGAN	**4.10**	**2.95**	**4.32**	**3.18**	2.27	**3.36**	**5.92**	**2.71**	**4.91**
Real Data	4.88	3.31	5.00	3.83	2.58	3.92	6.46	2.86	5.47

下 Top-1 类别的得分仅次于 SelectionGAN. 这主要是因为 Dayton 数据集中的多视角图像是区别较大的户外图像, 这种多视角图像生成任务具有较大的难度, SelectionGAN 中引入了目标图像的语义分割图来辅助生成, 本文模型却没有引入外部知识. 因此本文模型生成的图像具有一定的不确定性, 从而导致图像中存在模糊的区域, 但 ViewGAN 的得分与 SelectionGAN 的得分很接近, 这也表明了本文模型的潜力.

3) Top-k 指标. 此外, 我们还计算了真实图像和合成图像的 Top-k 预测准确率. 我们使用与 Inception score 同样的与训练模型来获得真实图像的标注和合成图像的预测标签. 实验计算了 Top-1 预测准确率和 Top-5 预测准确率. 每种准确率的计算方法有两种方法: a) 考虑所有的测试图像; b) 只考虑那些预测标签概率值超过 0.5 的测试图像.

4) Top-k 评估结果. 基于 Top-k 的实验结果如表 4 所示. 由实验结果可知: 本文模型在 3 个数据集上的性能均优于基线模型, 显著提升了预测准确率. 这说明本文模型生成图像具有较高的清晰度、丰富的语义细节, 在复杂多变的 DeepFashion 数据集和 Dayton数据集上表现出了较强的鲁棒性.

值得注意的是, 本文模型在 ICG Lab6 数据集上的准确率要略低于在其他两个数据集的准确率. 这主要是因为 ICG Lab6 数据集中的图像包含较多小目标物体, 这对合成清晰图像来说本身就是一个很大的挑战, 因此以往模型最高只达到了 76.44% 的准确率. 而本文模型引入基于蒙特卡洛搜索的惩罚机制, 充分利用了模型的中间结果, 保证了图像具有更细致的语义细节, 最高达到了 93.25% 的准确率.

4.3.2　结构相似性、峰值信噪比和 Sharpness difference

1) 指标. 参考文献 [33−34] 中的工作, 我们利用结构相似性、峰值信噪比和 SD (Sharpness difference) 指标来衡量合成图像与真实图像之间的像素级相似度.

2) 结构相似性 (SSIM). 基于图像的亮度、对比度等属性评估图像之间的相似度, 其取值范围为 [−1, 1], 值越大则图像之间的相似度越高. 结构相似性的计算式为

$$SSIM\left(I_g, I_v\right)=\frac{\left(2\mu_{I_g}\mu_{I_v}+c_1\right)\left(2\sigma_{I_gI_v}+c_2\right)}{\left(\mu_{I_g}^2+\mu_{I_v}^2+c_1\right)\left(\sigma_{I_g}^2+\sigma_{I_v}^2+c_2\right)} \tag{15}$$

其中, μ_{Ig}, μ_{I_v} 分别表示合成图像 I_g 和真实图像 I_v 的均值, σ_{I_g}, σ_{I_v} 分别表示图像 I_g 和 I_{I_v} 的标准差. c_1, c_2 是为了避免分母为 0 而引入的常数.

3) 峰值信噪比 (PSNR). 通过测量到达噪音比率的顶点信号来评估合成图像相对于真实图像的质量. 峰值信号比越大, 合成图像的质量越高. 峰值信噪比的计算式为

$$PSNR\left(I_g, I_v\right)=10\lg\left(\frac{\max_{I_g}^2}{mse(I_g, I_v)}\right) \tag{16}$$

其中, $mse(I_g, I_v)=\dfrac{1}{n}\sum_{i=1}^{n}(I_v[i]-I_g[i])^2$, $\max_{I_g}=255$.

4) SD (Sharpness difference) 测量图像生成过程中清晰度的损失, 为了计算合成图像和真实图像之间的清晰度差异, 我们参考文献 [35] 中的思路, 计算图像之间的梯度变化

$$SharpDiff\left(I_g, I_v\right)=10\lg\left(\frac{\max_{I_g}^2}{grads}\right) \tag{17}$$

$$grads=\frac{1}{N}\sum_{i}\sum_{j}\left(\left|\left(\nabla_i I_v+\nabla_j I_v\right)-\left(\nabla_i I_g+\nabla_j I_g\right)\right|\right) \tag{18}$$

其中, $\nabla_i I=\left|I_{i,j}-I_{i-1,j}\right|$, $\nabla_j I=\left|I_{i,j}-I_{i,j-1}\right|$. 式 (17) 中的 *SharpDiff* 可看作是梯度的倒数, 我们希望图像之间的梯度尽量小, 所以 *SharpDiff* 就应该尽量大.

5) 结果. 基于 SSIM, PSNR, SD 的实验结果如表 5 所示. 由实验结果可以看出: 本文模型 ViewGAN 在 3 个数据集上的得分均高于基线模型的得分. 相

表 4　各模型 Top-k 预测准确率统计表, 该指标越高表明模型性能越好

Table 4　Accuracies of different models (For this metric, higher is better)

模型	DeepFashion				Dayton				ICG Lab6			
	Top-1 class		Top-5 class		Top-1 class		Top-5 class		Top-1 class		Top-5 class	
Pix2Pix	7.34	9.28	25.79	32.68	6.80	9.15	23.55	27.00	1.33	1.62	5.43	6.79
X-Fork	20.68	31.35	50.45	64.12	30.00	48.68	61.57	78.84	5.94	10.36	20.83	30.45
X-Seq	16.03	24.31	42.97	54.52	30.16	49.85	62.59	80.70	4.87	8.94	17.13	24.47
VariGAN	25.67	31.43	55.52	63.70	32.21	52.69	67.95	84.31	10.44	20.49	33.45	41.62
SelectionGAN	41.57	64.55	72.30	88.65	42.11	68.12	77.74	92.89	28.35	54.67	62.91	76.44
ViewGAN	**65.73**	**95.77**	**91.65**	**98.21**	**69.39**	**89.88**	**93.47**	**98.78**	**58.97**	**83.20**	**88.74**	**93.25**

表 5 各模型 SSIM, PSNR, SD 和速度统计表, 其中 FPS 表示测试时每秒处理的图像数量, 所有指标得分越高表明模型性能越好

Table 5 SSIM, PSNR, SD of different models. FPS is the number of images processed per second during testing (For all metrics, higher is better)

模型	DeepFashion			Dayton			ICG Lab6			速度 (帧/s)
	SSIM	PSNR	SD	SSIM	PSNR	SD	SSIM	PSNR	SD	
Pix2Pix	0.39	17.67	18.55	0.42	17.63	19.28	0.23	15.71	16.59	166±5
X-Fork	0.45	19.07	18.67	0.50	19.89	19.45	0.27	16.38	17.35	87±7
X-Seq	0.42	18.82	18.44	0.50	20.28	19.53	0.28	16.38	17.27	75±3
VariGAN	0.57	20.14	18.79	0.52	21.57	19.77	0.45	17.58	17.89	70±5
SelectionGAN	0.53	23.15	19.64	0.59	23.89	20.02	0.61	26.67	19.76	66±6
ViewGAN	**0.70**	**26.47**	**21.63**	**0.74**	**25.97**	**21.37**	**0.80**	**28.61**	**21.77**	62±2

比于目前主流的 SelectionGAN 模型, 本文模型 ViewGAN 在 ICG Lab6 数据集上的 SSIM 分数提升了 32.29%, SD 分数提升了 10.18%, 在 DeepFashion 数据集上的 PSNR 得分提升了 14.32%.

上述定量实验结果证明: 面向室内、室外等复杂环境, 本文模型能够学会如何生成高质量多视角图像, 而不是简单地重复训练集中的图像. 这种分布式的生成方式使每个生成器专注于学习固定视角的图像, 提出的惩罚机制能够进一步完善图像的语义细节, 使得生成的图像更加逼真, 显著地化解了图像中的人工痕迹.

6) 速度对比实验. 为验证各个模型在速度上的差异, 我们在 Dayton 数据集上对各个模型测试时的速度进行了对比实验. 实验结果如表 5 所示, 从实验结果看出: ViewGAN 的测试速度低于所有基线模型, 这要是因为采用由粗到精的两阶段生成方法势必会造成计算量的增加, 此外蒙特卡洛搜索耗时较多. 但事实上, ViewGAN 和 SelectionGAN 的速度差距不大, 且 62 帧/s 可以满足实际应用中的需要.

7) 最小数据量实验. 为了验证训练集规模对于模型性能的影响, 我们在 DeepFashion 数据集上对 ViewGAN 的最小训练样本量进行了探究实验. 实验结果如表 6 所示, 从实验结果可以看出: 随着训练集规模的缩小, ViewGAN 的性能下降较为缓慢. 直至训练集规模缩小至 60% 时, ViewGAN 在各指标上的得分才低于基线模型 SelectionGAN 在完整数据集上的得分, 这表明 ViewGAN 具有较强的鲁棒性, 即使在小规模数据集上仍能学习到关键的特征信息, 在一定程度上克服了以往模型泛化能力不强的缺点.

4.4 定性评估

在 3 个数据集上的定性评估结果如图 4 ~ 6 所示, 测试图像的分辨率均为 256×256 像素. 从实验结果可以看出: 本文模型 ViewGAN 生成的图像更加清晰, 有关物体或场景的细节更丰富. 在 DeepFashion 数据集中, 以往模型易生成模糊失真的图像, ViewGAN 学会了如何生成多样性强的服饰图像, 在服饰的图案、人物的姿态上具有更多的语义细节. 在 Dayton 数据集中, ViewGAN 能够生成更加自然的图像, 图像中的房屋、草木和汽车都更符合实际, 减轻了图像中的人工痕迹. 在 ICG Lab6 数据集中, ViewGAN 在面对复杂环境时仍表现出较好的性能, 图像中的桌椅、电脑等小目标都十分逼真, 在清晰度方面非常接近真实图像.

表 6 最小数据量实验结果

Table 6 Minimum training data experimental results

数据量 (幅)	SSIM	PSNR	SD
6 000 (100%)	0.70	26.47	21.63
5 400 (90%)	0.68	26.08	20.95
4 800 (80%)	0.66	24.97	20.31
4 200 (70%)	0.59	23.68	20.00
3 600 (60%)	0.51	21.90	18.89

4.5 消融分析

为了分析本文模型中不同组件的功能, 我们在 DeepFashion 数据集上进行了消融分析实验. 实验结果如表 7 所示. 由实验结果可知: 相比于模型 A, 模型 B 的性能更好, 这表明由粗到精的两阶段生成方法能够更好地提升图像的清晰度. 模型 C 的性能得到进一步的提升, 这是因为本文模型采用的混合生成对抗网络有效地扩充模型的内存容量, 使得每个生成器更擅长生成指定视角的图像. 模型 D 通过引入类内损失促进了每个生成器的学习, 在提升系统稳定性的同时提升了图像的质量. 而模型 E 的得分表明: 而本文提出的惩罚机制显著提升了模型的

图 4　各模型在 DeepFashion 数据集上的测试样例

Fig. 4　Results generated by different models on DeepFashion dataset

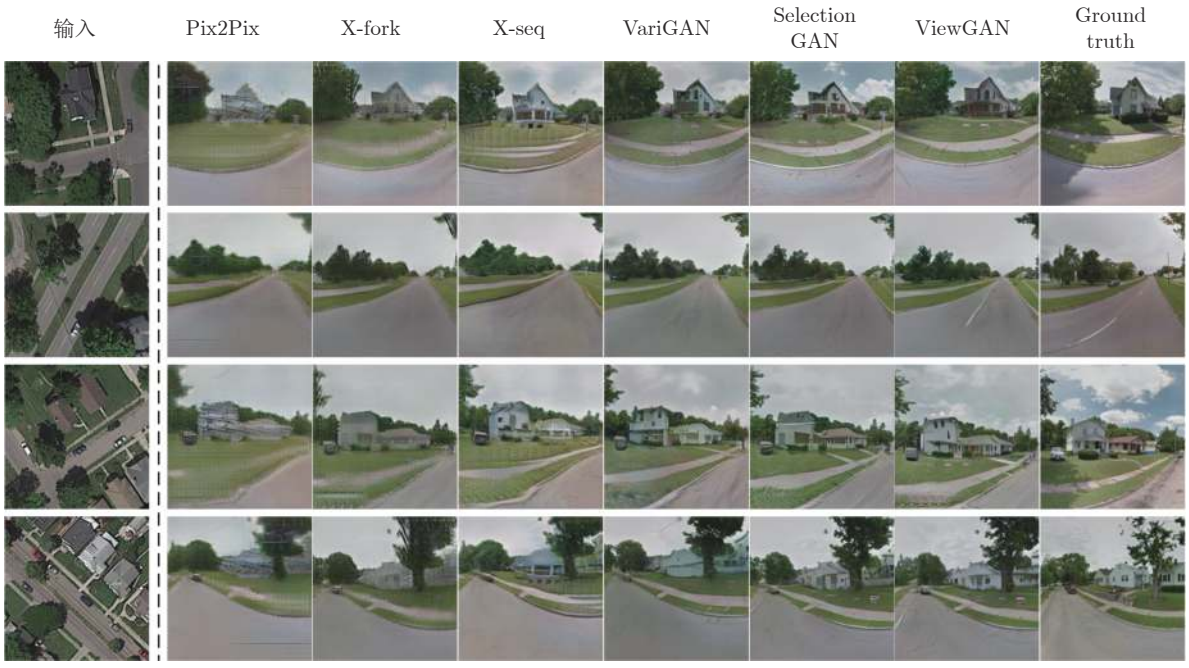

图 5　各模型在 Dayton 数据集上的测试样例

Fig. 5　Results generated by different models on Dayton dataset

性能, 使模型在生成过程中得到了充足的语义约束, 这大大增强了合成图像的清晰度和真实感.

为进一步探索本文提出的基于蒙特卡洛搜索的惩罚机制, 我们将 ViewGAN 生成图像的过程进行了可视化, 如图 7 所示. 从图中可以看出, 低分辨率目标图像仅仅画出了人物的基本轮廓, 缺乏服饰的细节. 而利用蒙特卡洛搜索进行多次挖掘后, 提取出了不同层次的语义信息, 如服饰的蕾丝边、手臂的轮廓等, 之后调用细粒度模块将这些语义信息填补到目标图像中, 从而得到最终逼真的高分辨率目

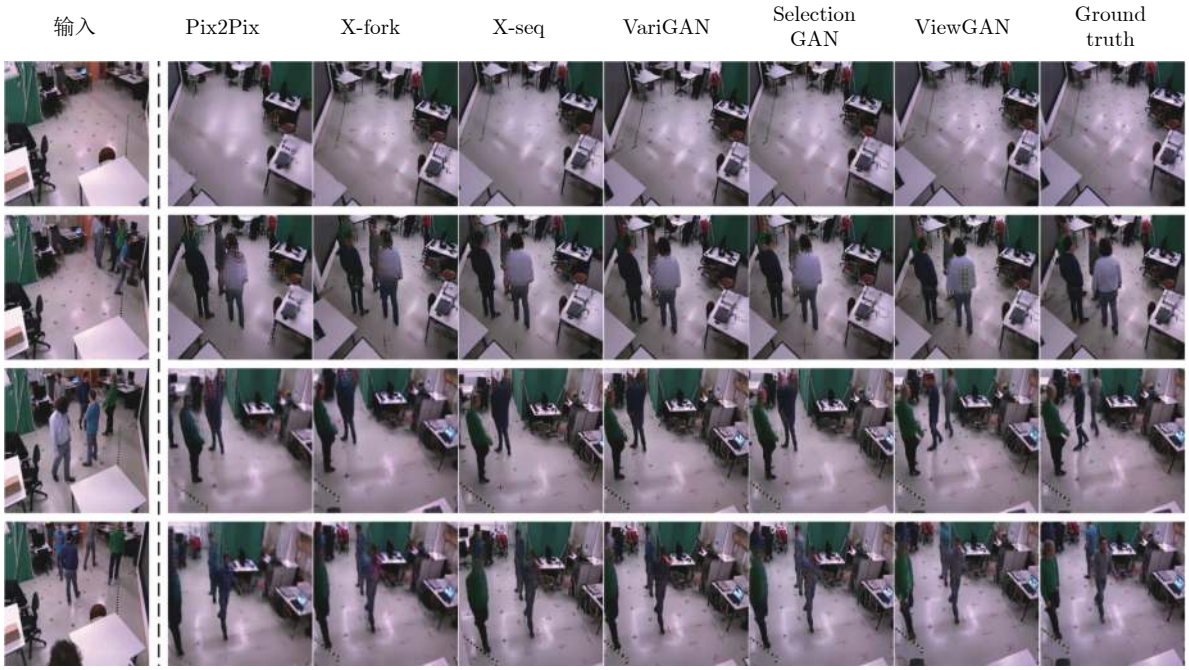

图 6 各模型在 ICG Lab6 数据集上的测试样例

Fig. 6 Results generated by different models on ICG Lab6 dataset

表 7 消融分析实验结果

Table 7 Ablations study of the proposed ViewGAN

模型	结构	SSIM	PSNR	SD
A	Pix2Pix	0.46	19.66	18.89
B	A+由粗到精生成方法	0.53	22.90	19.31
C	B+混合生成对抗网络	0.60	23.77	20.03
D	C+类内损失	0.60	23.80	20.11
E	D+惩罚机制	0.70	26.47	21.63

标图像.

5 总结与未来方向

本文提出了一种多视角图像生成模型 ViewGAN, 它可基于不同视角的图像合成新视角的图像. 模型首先利用粗粒度模块生成低分辨率目标图像, 之后利用蒙特卡洛搜索挖掘中间结果的语义信息, 细粒度模块基于搜索结果合成高分辨率目标图像. 在 3 个公开数据集 DeepFashion, Dayton 和 ICG Lab6 上的定量实验与定性实验证明: 相比于目前的主流模型, 本文模型能够生成更加清晰的、多样性的目标图像. 其中消融分析实验证明了本文提出的基于蒙塔卡罗搜索的惩罚机制显著提升了图像的质量. 此外, 本文模型可灵活迁移至多视角生成的各个场景, 未来我们将进一步探索该方法在图像风格迁移、图像翻译等领域的应用.

图 7 ViewGAN 生成图像的可视化过程 ((a) 输入图像; (b) 粗粒度模块合成的低分辨率目标图像; (c) 蒙特卡洛搜索的结果; (d) 细粒度模块合成的高分辨率目标图像)

Fig. 7 Visualization of the process of ViewGAN generating images ((a) The input image; (b) The LR image generated by coarse image module; (c) Intermediate results generated by Monte Carlo search module; (d) The HR image generated by fine image module)

References

1 Zhu X E, Yin Z C, Shi J P, Li H S, Lin D H. Generative adversarial frontal view to bird view synthesis. In: Proceedings of the 2018 International Conference on 3D Vision (3DV). Verona, Italy: IEEE, 2018. 454–463

2 Regmi K, Borji A. Cross-view image synthesis using conditional GANs. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 3501–3510

3 Zhai M H, Bessinger Z, Workman S, Jacobs N. Predicting ground-level scene layout from aerial imagery. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 4132–4140

4 Zhao B, Wu X, Cheng Z Q, Liu H, Jie Z Q, Feng J S. Multi-view image generation from a single-view. In: Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Republic of Korea: ACM, 2018. 383–391

5 Tang H, Xu D, Sebe N, Wang Y Z, Corso J J, Yan Y. Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 2412–2421

6 Kingma D P, Welling M. Auto-encoding variational bayes. In: Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014. Banff, Canada, 2014.

7 Park E, Yang J, Yumer E, Ceylan D, Berg A C. Transformation-grounded image generation network for novel 3D view synthesis. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 702–711

8 Choy C B, Xu D F, Gwak J Y, Chen K, Savarese S. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 628–644

9 Wang Kun-Feng, Gou Chao, Duan Yan-Jie, Lin Yi-Lun, Zheng Xin-Hu, Wang Fei-Yue. Generative adversarial networks: The state of the art and beyond. *Acta Automatica Sinica*, 2017, **43**(3): 321–332
(王坤峰, 苟超, 段艳杰, 林懿伦, 郑心湖, 王飞跃. 生成式对抗网络GAN的研究进展与展望. 自动化学报, 2017, **43**(3): 321–332)

10 Browne C B, Powley E, Whitehouse D, Lucas S M, Cowling P I, Rohlfshagen P, et al. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 2012, **4**(1): 1–43

11 Srivastava A, Valkov L, Russell C, Gutmann M U, Sutton C. VEEGAN: Reducing mode collapse in GANs using implicit variational learning. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, USA: Curran Associates, Inc., 2017. 3308–3318

12 Isola P, Zhu J Y, Zhou T H, Efros A A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1125–1134

13 Yan X C, Yang J M, Sohn K, Lee H. Attribute2Image: Conditional image generation from visual attributes. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 776–791

14 Gregor K, Danihelka I, Graves A, Rezende D, Wierstra D. DRAW: A recurrent neural network for image generation. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR, 2015.

15 Tang Xian-Lun, Du Yi-Ming, Liu Yu-Wei, Li Jia-Xin, Ma Yi-Wei. Image recognition with conditional deep convolutional generative adversarial networks. *Acta Automatica Sinica*, 2018, **44**(5): 855–864
(唐贤伦, 杜一铭, 刘雨微, 李佳歆, 马艺玮. 基于条件深度卷积生成对抗网络的图像识别方法. 自动化学报, 2018, **44**(5): 855–864)

16 Dumoulin V, Belghazi I, Poole B, Mastropietro O, Lamb A, Arjovsky A, et al. Adversarially learned inference [Online], available: https://arxiv.gg363.site/abs/1606.00704, February 21, 2017

17 Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates, Inc., 2016. 2180–2188

18 Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 5907–5915

19 Johnson J, Gupta A, Fei-Fei L. Image generation from scene graphs. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1219–1228

20 Zhu J Y, Park T, Isola P, Efros A A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2242–2251

21 Choi Y, Choi M, Kim M, Ha J W, Kim S, Choo J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 8789–8797

22 Dosovitskiy A, Springenberg J T, Tatarchenko M, Brox T. Learning to generate chairs, tables and cars with convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(4): 692–705

23 Wu J J, Xue T F, Lim J J, Tian Y D, Tenenbaum J B, Torralba A, et al. Single image 3D interpreter network. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 365–382

24 Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates, Inc., 2016. 2234–2242

25 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer, 2015. 234–241

26 Liu Z W, Luo P, Qiu S, Wang X G, Tang X O. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 1096–1104

27 Vo N N, Hays J. Localizing and orienting street views using overhead imagery. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 494–509

28 Possegger H, Sternig S, Mauthner T, Roth P M, Bischof H. Robust real-time tracking of multiple objects by volumetric mass densities. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE, 2013. 2395–2402

29 Deng J, Dong W, Socher R, Li L J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009. 248–255

30 Zhou B L, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(6): 1452–1464

31 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates, Inc., 2012. 1097–1105

32 Yamaguchi K, Kiapour M H, Berg T L. Paper doll parsing: Retrieving similar styles to parse clothing items. In: Proceedings of

the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 3519−3526

33 Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 105−114

34 Tang Yi-Ling, Jiang Shun-Liang, Xu Shao-Ping, Liu Ting-Yun, Li Chong-Xi. Asymmetrically distorted stereoscopic image quality assessment based on ocular dominance. *Acta Automatica Sinica*, 2019, **45**(11): 2092−2106
(唐祎玲, 江顺亮, 徐少平, 刘婷云, 李崇禧. 基于眼优势的非对称失真立体图像质量评价. 自动化学报, 2019, **45**(11): 2092−2106)

35 Mathieu M, Couprie C, LeCun Y. Deep multi-scale video prediction beyond mean square error [Online], available: https://arxiv.gg363.site/abs/1511.05440, February 26, 2016

卫 星 合肥工业大学副教授. 2009 年于中国科技大学获得博士学位. 主要研究方向为深度学习与物联网工程, 无人驾驶解决方案.
E-mail: weixing@hfut.edu.cn
(**WEI Xing** Associate professor at Hefei University of Technology. He received his Ph.D. degree from University of Science and Technology of China in 2009. His research interest covers deep learning and internet of things engineering, and driverless solutions.)

李 佳 合肥工业大学计算机与信息学院硕士研究生. 主要研究方向为自然语言处理, 情感对话生成.
E-mail: lijiajia@mail.hfut.edu.cn
(**LI Jia** Master student at the School of Computer and Information, Hefei University of Technology. His research interest covers natural language processing and emotional conversation generation.)

孙 晓 博士, 合肥工业大学计算机与信息学院情感计算研究所副教授. 主要研究方向为情感计算, 自然语言处理, 机器学习与人机交互, 本文通信作者. E-mail: sunx@hfut.edu.cn
(**SUN Xiao** Ph.D., associate professor at the Institute of Affective Computing, Hefei University of Technology. His research interest covers affective computing, natural language processing, machine learning and human-machine interaction. Corresponding author of this paper.)

刘邵凡 合肥工业大学硕士研究生. 2018 年于合肥工业大学获得学士学位. 主要研究方向为目标检测和领域自适应.
E-mail: frank-uzi@hotmail.com
(**LIU Shao-Fan** Master student at Hefei University of Technology. He received his bachelor degree from Hefei University of Technology in 2018. His research interest covers object detection and domain adaptation.)

陆 阳 合肥工业大学教授. 2002 年于合肥工业大学获得博士学位. 主要研究方向为物联网工程和分布式控制系统. E-mail: luyang.hf@126.com
(**LU Yang** Professor at Hefei University of Technology. He received his Ph.D. degree from Hefei University of Technology in 2002. His research interest covers IOT (internet of things) engineering and distributed control system.)