软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):2116−2133 [doi: 10.13328/j.cnki.jos.006769]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多域 VQGAN 的文本生成国画方法研究*

孙泽龙[1], 杨国兴[1], 温静远[1], 费楠益[2], 卢志武[1], 文继荣[1]

[1](中国人民大学 高瓴人工智能学院, 北京 100872)

[2](中国人民大学 信息学院, 北京 100872)

通信作者: 卢志武, E-mail: luzhiwu@ruc.edu.cn

摘　要: 随着生成式对抗网络的出现, 从文本描述合成图像最近成为一个活跃的研究领域. 然而, 目前文本描述往往使用英文, 生成的对象也大多是人脸和花鸟等, 专门针对中文和中国画的研究较少. 同时, 文本生成图像任务往往需要大量标注好的图像文本对, 制作数据集的代价昂贵. 随着多模态预训练的出现与推进, 使得能够以一种优化的方式来指导生成对抗网络的生成过程, 大大减少了对数据集和计算资源的需求. 提出一种多域 VQGAN 模型来同时生成多种域的中国画, 并利用多模态预训练模型 WenLan 来计算生成图像和文本描述之间的距离损失, 通过优化输入多域 VQGAN 的隐空间变量来达到图片与文本语义一致的效果. 对模型进行了消融实验, 详细比较了不同结构的多域 VQGAN 的 *FID* 及 *R*-precisoin 指标, 并进行了用户调查研究. 结果表示, 使用完整的多域 VQGAN 模型在图像质量和文本图像语义一致性上均超过原 VQGAN 模型的生成结果.

关键词: 文本生成图像; 多域生成; 中国画生成

中图法分类号: TP391

中文引用格式: 孙泽龙, 杨国兴, 温静远, 费楠益, 卢志武, 文继荣. 基于多域VQGAN的文本生成国画方法研究. 软件学报, 2023, 34(5): 2116–2133. http://www.jos.org.cn/1000-9825/6769.htm

英文引用格式: Sun ZL, Yang GX, Wen JY, Fei NY, Lu ZW, Wen JR. Text-to-Chinese-painting Method Based on Multi-domain VQGAN. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2116–2133 (in Chinese). http://www.jos.org.cn/1000-9825/6769.htm

Text-to-Chinese-painting Method Based on Multi-domain VQGAN

SUN Ze-Long[1], YANG Guo-Xing[1], WEN Jing-Yuan[1], FEI Nan-Yi[2], LU Zhi-Wu[1], WEN Ji-Rong[1]

[1](Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China)

[2](School of Information, Renmin University of China, Beijing 100872, China)

Abstract: With the development of generative adversarial networks (GANs), synthesizing images from textual descriptions has become an active research area. However, textual descriptions used for image generation are often in English, and the generated objects are mostly faces, flowers, birds, etc. Few studies have been conducted on the generation of Chinese paintings with Chinese descriptions. The text-to-image generation often requires an enormous number of labeled image-text pairs, and the cost of dataset production is high. With the advance in multimodal pre-training, the GAN generation process can be guided in an optimized way, which significantly reduces the demand for datasets and computational resources. In this study, a multi-domain vector quatization generative adversarial network (VQGAN) model is proposed to simultaneously generate Chinese paintings in multiple domains. Furthermore, a multimodal pre-trained model WenLan is used to calculate the distance loss between generated images and textual descriptions. The semantic consistency between images and texts is achieved by optimization of the hidden space variables input into multi-domain VQGAN. Finally, an ablation experiment is conducted to compare different variants of multi-domain VQGAN in terms of the *FID* and *R*-precision metrics, and a user investigation is carried out. The results demonstrate that the complete multi-domain VQGAN model outperforms the original VQGAN model in terms of image quality and text-image semantic consistency.

Key words: text-to-image generation; multi-domain generation; Chinese painting generation

---

* 基金项目: 国家自然科学基金 (61976220, 61832017); 北京高等学校卓越青年科学家计划 (BJJWZYJH012019100020098)
本文由“融合预训练技术的多模态学习研究”专题特约编辑宋雪萌副教授、聂礼强教授、申恒涛教授、田奇教授、黄华教授推荐.
收稿时间: 2022-04-16; 修改时间: 2022-05-29; 采用时间: 2022-08-24; jos 在线出版时间: 2022-09-20
CNKI 网络首发时间: 2023-03-23

当人们听到或者读到一段故事时, 会不由自主地在脑海中浮现出相应的画面. 将视觉世界和语言世界结合起来对人类来说是非常自然的一件事, 以至于我们时常意识不到这其实是个非常复杂的过程. 其实, 人类视觉心理意象能力在许多认知过程中都发挥着重要的作用, 如记忆、空间想象和推理等[1]. 受到人类如何可视化文字场景的启发, 建立一个能够理解视觉和语言之间关系的系统, 并能够生成出反映文字描述的高质量画面, 是人工智能发展的一个重要里程碑.

基于深度神经网络的文本生成图像模型已经非常多, 它们之间的区别主要体现在如何去表示文本和图像的语义空间, 以及使用什么样的模型或方法来连接这两种语义空间. 早期的方法, 如 StackGAN[2], DMGAN[3], AttnGAN[4]等, 尝试训练一个卷积生成器, 直接从给定文本的特征向量中预测图片的像素. 然而, 这些模型的泛化能力不强, 当应用于通用图像生成时, 在图像质量和文本图像匹配方面的效果较差.

最近, DALL-E[5]和 CogView[6]在文生成图领域的表现较好. 为了使得图像能够像自然语言一样进行表示, 这两个模型都采用 Transformer[7], 通过类似于 VQVAE (vector quantised-variational autoencoder)[8]和 VQGAN (vector quatization generative adversarial network)[9]的矢量量化模型, 将图像的隐空间特征向量进行离散化表示, 之后将文字和图像的隐空间特征表示拼接在一起输入到 Transformer 中, 即可在统一的框架内对跨模态文本图像数据进行训练. 然而, 虽然这种方式的生成效果较好, 但是训练这样的大规模模型需要数以亿计的成对文本图像数据, 构建这样的数据集代价十分昂贵.

目前的文生成图模型大多是针对某一特定的域 (如人脸、花、鸟等), 难以做到用同一个模型同时生成多域的图像. 且目前文本生成图像使用的文本主要是英文, 在中文和中国画的领域研究较少. 因此, 针对上述问题, 本文提出了基于中文多模态预训练模型 WenLan[10,11]和生成模型 VQGAN 的多域中国画生成方法, 能够使用少量的无文字标注的数据, 按域的不同, 从诗句中生成中国山水画. 我们提出模型的整体框架如图 1 所示, 展示了生成过程的两个阶段. 第 1 阶段, 我们设计了多域 VQGAN 的网络结构, 能够同时接收不同域的图片输入, 并能够生成不同域的图片. 第 2 阶段, 将训练好的多域 VQGAN 模型的编码器去除, 利用随机噪声生成图片, 作为 WenLan 的输入; 而 WenLan 会对文本和图片进行编码提取特征, 将得到的特征进行相似度计算作为损失函数, 不断更新随机噪声, 使得生成的图像能够越来越靠近文本的描述.

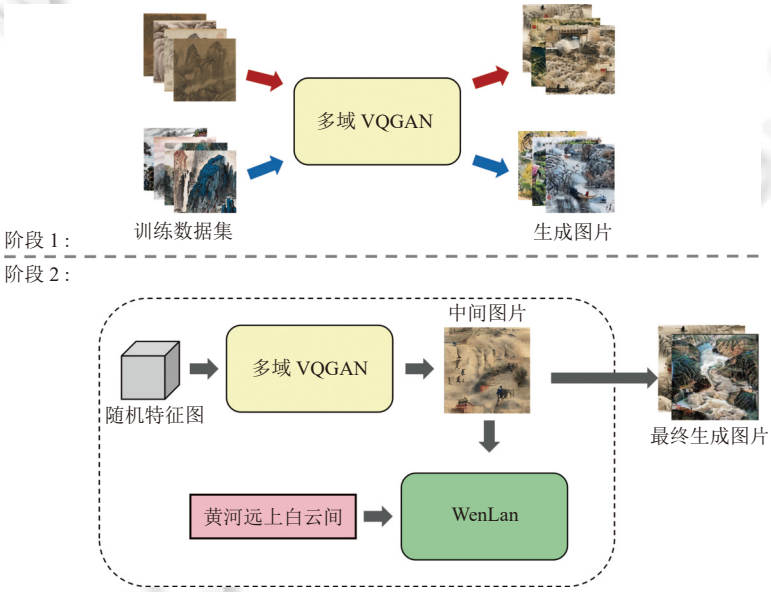


图 1 基于多域 VQGAN 的文本生成国画模型框架

本文的主要贡献有: (1) 提出了一种多域文生成图模型, 能够用一个模型生成多个域的图像. 且在训练过程中无需文字标注的数据集, 规避了目前模型难以获取大规模标注好的文本图像对的问题. (2) 提出的模型接收中文描

述来作为文本输入, 并能够使用中国古代诗句生成高清的中国画.

本文第 1 节介绍文生成图模型的相关方法和研究现状. 第 2 节介绍本文所需的基础知识, 包括编码字典、生成对抗网络和多模态预训练模型. 第 3 节和第 4 节介绍本文构建的基于多域 VQGAN 和 WenLan 的多域中国画生成模型. 第 5 节通过消融实验、用户调查实验, 将本文提出的模型与最新模型生成的图像进行对比, 证明了模型的有效性. 第 6 节是全文的总结.

1 相关工作

1.1 文本生成图像

通过文本描述生成图像的方法, 最早可以追溯到深度生成模型的早期, 当时 Mansimov 等人将文本信息添加到 DRAW[12]中. 之后, 由于生成对抗网络在图像合成技术上展现出的出色性能[13], 基于生成对抗网络的方法开始主导文本生成图像的任务.

2016 年, Reed 等人[14]首次尝试用 GAN 来进行文本生成图像的任务, 模型将文本信息作为条件输入生成器来进行约束的图像生成, 而判别器需要判断两种情况, 分别是真实数据、生成图像和文本描述不匹配的情况, 最终生成了分辨率为 64×64 的图像. 为了进一步提高生成图像的质量, Reed 等人在同年提出了 GAWWN 模型[15], 在之前模型的基础上, 通过指定关键点约束了图像中对象的不同部分, 使文本能够与图像细节相对应, 最终生成了分辨率为 128×128 的图像. 胡涛等人提出的基于单阶段 GANs 的文本生成图像网络[16], 在生成器中加入了通道-像素注意力模块, 并利用全局文本表示和局部词嵌入技术提供细粒度的判别信息, 能够使用单个生成器和判别器生成高质量的图像. 而堆叠式生成对抗网络 StackGAN[2], 则将生成过程看作为概略图细化的过程, 将传统方法中使用一对生成器判别器扩展成了使用两对生成器判别器. StackGAN 的生成过程分为两个阶段: 第 1 阶段在给定随机噪声向量和文本向量的情况下生成分辨率为 64×64 的粗略图像; 第 2 阶段将该初始图像和文本向量输入到第 2 个生成器, 最终输出分辨率为 256×256 的图像. 在这两个阶段, 每个判别器都被训练来区分匹配和不匹配的图像文本对. StackGAN++[17]通过端到端框架进一步改进了该体系结构, 其在 StackGAN 的基础上将堆叠结构改进为树状结构, 联合训练了 3 对生成器判别器; 模型还使用了一种色彩一致性正则化的方法, 旨在最小化不同尺度之间像素的均值和协方差之间的差异. FusedGAN[18]借鉴了同时训练有条件和无条件分布的想法, 由两个生成器组成, 一个用于有条件图像生成, 一个用于无条件图像生成, 两个生成器部分共享一个公共的潜在隐空间. 为了克服模型包含多个生成器的问题, HDGAN[19]提出了一种称为伴随层次嵌套对抗性目标, 通过规范在不同的中间层生成的低分辨率图像来使得生成器可以捕获复杂的图像信息, 在生成过程中的每个中间层都嵌套了一个对应的判别器, 用于区分生成图像的真假和与描述文本的语义相关性. 类似地, PPAN 模型[20]只使用 1 个生成器和 3 个不同的判别器, 其生成器采用金字塔框架[21], 通过自下而上的横向连接路径, 将低分辨率、语义强的特征与高分辨率、语义弱的特征相结合. AttnGAN[4]以 StackGAN++为基础, 将注意力机制融入到多阶段的生成过程中, 注意力机制允许网络根据相关单词和全局句子向量合成细粒度细节, 从而更好地对齐图像和文本. MirrorGAN[22]利用了“文本到图像的重新描述学习生成”的思想, 提出了 text-to-image-to-text 架构. 生成图像后, 反向生成与之匹配的文本, 并计算反向生成的文本与源文本之间的损失, 能够更好地学习生成图像和文本的语义一致性. DMGAN[3]使用动态记忆模型替换了 AttnGAN 中的注意力机制, 使得模型能够生成更加生动的图像. CAE-GAN[23]通过交叉注意力编码器, 将文本信息与视觉信息进行翻译和对齐, 以捕捉文本与图像信息之间的跨模态映射关系, 从而提升生成图像的逼真度和与输入文本描述的匹配度.

除了上述基于 GAN 的文生成图模型, 最新两个代表性工作 DALL-E[5]和 CogView[6]采用 Transformer 生成结构, 该 Transformer 结构具有至少几十亿参数, 并且需要利用海量的高质量文本-图像对来预训练该 Transformer. 在模型预训练时, 它们使用带有 VQVAE[8]的 tokenizer 将图像序列化, 并与序列化的文本拼接在一起输入 Transformer 模型中生成图像. 最后在处理文本图像生成类任务时, 模型会计算一个 Caption Score 对生成图像进行排序, 从而选择与文本最为匹配的图像作为结果. 虽然这种方式的生成效果较好, 但是训练这样的大规模模型需要数以亿计的成对文本图像数据, 构建这样的数据集代价非常昂贵, 同时训练模型占用的计算资源也是非常可观的, 不是一般

研究机构能够承受的.

1.2 基于优化的 GAN 反演方法

对于一个已经预训练好的无条件 GAN 模型, GAN 反演 (GAN inversion) 试图将原图 x 编码成隐空间表示 z^*, 使得由 z^* 生成的图片能够在视觉上与原图 x 相似. GAN 的反演主要有 3 种方式: 基于学习、基于优化或者基于混合方式. 其中基于优化的方法与本文使用方法较为相似, 因此在本节主要介绍该方式的相关内容. 图 2 展示了该方法的一般流程.

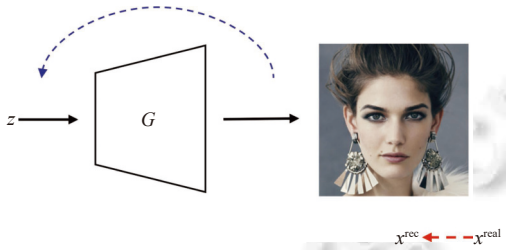


图 2 基于优化方法的 GAN 反演一般流程

现有的基于优化的 GAN 反演方法通常是通过优化隐向量来重建目标图像. 即:

$$z^* = \arg\min_z L(G(z), x) \tag{1}$$

其中, L 为某种距离度量函数, G 为已预训练好的生成器. 而基于优化的 GAN 反演方法主要有 3 个问题: 如何优化 z 向量, 如何解决局部最小问题及如何初始化 z 向量.

在以往的工作中, Creswell 等人[24]选择通过梯度下降的方法来优化隐向量 z, 并且在优化过程中, 一次性对一批图像样本进行反演, 这种方法不仅能抵消反演过程中批归一化带来的影响, 同时也能允许多个图片样本并行反演. Image2StyleGAN[25]中也使用了梯度下降的方法, 将给定图像映射到预训练好的 StyleGAN[26]的扩展潜空间 W+中, 并且通过线性插值, 交叉以及添加向量和比例差向量 3 种基本运算来对潜在空间的结构进行更深入的解释. Image2StyleGAN++[27]在其基础上创新性地加入了噪声, 以提高嵌入质量, 生成更高质量的嵌入图片, 并且通过先优化嵌入, 再优化噪声的方法得到了比同时优化两者更高的峰值信噪比. 而 Voynov 等人[28]则利用了雅可比矩阵分解 (Jacobian decomposition) 来分析预训练 GAN 模型的潜在空间, 并且还介绍了一种可以同时标识一批方向的轻量级方法来减轻计算大量雅可比矩阵带来的昂贵开销.

针对局部最小的问题, 一般通过使用两种类型的优化器来解决: 基于梯度的, 如 ADAM[29]和 L-BFGS[30], 以及无梯度的, 如协方差矩阵适应 (CMA)[31]. 如 Image2StyleGAN[25]使用了 ADAM 优化器, 而 Zhu 等人[32]的工作中则使用了 L-BFGS 方案. Huh 等人[33]使用了默认的优化超参数实验了各种无梯度优化的方法, 并发现将比较有挑战性的数据集中的图片反演到 StyleGAN 的隐空间时, 使用 CMA 及其变体 BasinCMA 的表现最好.

由于公式 (1) 通常是非凸的, 重建质量通常强烈依赖于 z 的初始化. Image2StyleGAN[25]中分析了两种初始化方式: 随机初始化和平均隐向量方式, 并且发现使用随机初始化方式得到的结果较好. 然而, 为了获得稳定的重建[32], 可能需要大量的随机初始化, 这使得实时处理无法实现. 因此, 一些工作[32,34]提出了使用编码器来获得更好的初始化向量.

基于优化的 GAN 反演方法与本文都是通过优化隐空间向量或特征图来影响最终的生成图片, 在优化方法上都是使用迭代的算法, 因此两者之间有一定的相似性. 然而, 需要注意的是, 传统的 GAN 反演方法没有引入文本信息, 而本文重点在于通过文本优化隐空间特征图, 以得到与文本语义相似的生成图片.

2 基础知识

2.1 编码字典

在没有监督的情况下学习图像的有效表示一直是机器学习的一个关键挑战. 自编码器是一个强大的生成模

型, 其原理是通过编码器将数据 x 进行编码, 映射到隐空间向量 z, 即 $z = encoder(x)$. 之后通过解码器将数据 x 从隐空间中的隐向量 z 重建成图像, 即 $x' = decoder(z)$. 若原图像和重建图像的重建损失较小, 即可认为该网络学习到了图像的有效隐空间表示.

变分自编码器[35]则在自编码器的基础上增加了一个限制, 即让 z 满足各向同性高斯分布, 称之为先验分布. 如此, 在变分自编码器训练结束后, 就可以在这个先验分布中随机采样, 得到一个随机的隐空间噪声, 再由解码器进行解码就能得到一张随机的图片. 可以看到, 变分自编码器的隐变量 z 的每一维都是一个连续的值.

而 Oord 等人[8]提出了一种称为矢量量化变分自动编码器 (VQVAE) 的方法, 用于学习图像的离散表示, 并使用卷积结构对其分布进行自回归建模. 而 VQGAN[9]引入了对抗性训练损失, 以加强对感知丰富的编码字典的学习. VQVAE 的隐变量 z 的每一维都为一个离散的整数, 这样可以有效地利用潜在空间. 而将 z 离散化的关键, 是学习一个编码字典. 假设编码字典由 K 个 D 维向量组成, 图片经过编码器后, 可以得到一个 $H\times W\times D$ 的特征图 z'. 则将这 $H\times W$ 个 D 维向量分别去编码字典里找到距离最近的向量, 用从编码字典得到的向量来代替自己, 最终得到量化后的 zq, 输入解码器得到图片.

2.2 生成对抗网络

Goodfellow 等人[36]于 2014 年首次提出生成对抗网络 (GAN). 学习 GAN 的初衷, 即生成不存在于真实世界的数据. 相比于传统的神经网络模型, GAN 是一种全新的非监督式架构. 其包括了两套独立的网络: 生成器和判别器, 两者作为互相对抗的目标. 如图 3 所示, 生成器 $G(z)$ 接收从先验噪声分布中随机采样得到的噪声 z 为输入, 输出为生成图像 x_g. 判别器 $D(x)$ 的输入为真实图片 x_r 或者 x_g, 输出是输入图片为真实图片的概率.

图 3 生成对抗网络结构

其训练过程可以看作两个网络在互相对抗的过程: 判别器被训练来区分真实的和生成的图像, 而生成器则被训练来捕捉真实的数据分布并生成尽可能真实的图像来骗过判别器. 具体来讲, 如 Goodfellow 等人生成对抗网络的训练过程可以定义为两个网络关于损失函数 $V(D,G)$ 的极大极小博弈过程. 判别器 $D(x)$ 的训练目标是最大化其分配给正确类别的可能性, 而生成器的训练目标是最小化其生成的图片被判别器判定为假的概率 $\log(1-D(G(z)))$. 其损失函数如下所示:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x\mid y)] + \mathbb{E}_{x\sim P_z}[\log(1-D(G(z\mid y)))] \tag{2}$$

2.3 多模态预训练模型

近年来, 人工智能在计算机视觉和自然语言处理方向均取得了很大的进展. 而融合二者的多模态深度学习也越来越受到关注. 多模态预训练的目标是将来自不同模态的数据进行对齐, 并能够将所学的知识迁移到各种下游任务. 多模态预训练模型根据其框架可分为两类: 单塔和双塔.

单塔网络模型将图片、文本等不同模态的输入一视同仁, 在同一个模型进行融合. 其代表有 UNITER[37]、Oscar[38]、M6[39]、VisualBERT[40]、Unicoder-VL[41]、VL-BERT[42]等模型. 它们利用一个特征融合模块 (例如 Transformer[7]) 来得到图像-文本对的嵌入. 其中, 一些单塔模型还使用对象检测器来检测图像区域, 并将这些区域与相应的单词进行匹配. 作为单塔模型的代表, UNITER 对 560 万图文对进行遮挡语言建模 (MLM)、遮挡区域建模 (MRM) 和图像文本匹配 (ITM) 的联合训练, 从而学习到通用的图像文本表示. Oscar[38]则使用 Fast R-CNN[43], 将检测到的对象标签与文本中的单词建立关联. 然而, 现有单塔结构通常通过假设文本和图像模态之间存在很强的语义相关性, 需要确定地模拟了图像-文本对之间的跨模态交互, 然而这一假设在现实场景中往往是无效的[10]. 此外, 单塔模型在推理阶段需要巨大的计算成本.

相比之下, 采用双塔结构的多模态预训练模型将不同模态的输入分别处理之后再进行交叉融合. 其使用单独的文本和图像编码器, 分别对文本和图像进行编码, 然后进行图文对匹配. 这种模式的检索效率更高, 但由于缺乏更深层次的图像-文本交互, 通常只能达到次优性能. 在最近的工作中, 如 LigntningDOT[44], 通过重新设计目标检测过程来应对这一挑战; 而 CLIP[45]、ALIGN[46]、WenLan 1.0[10]和 WenLan 2.0[11]则放弃了计算代价昂贵的对象检测器, 利用跨模态对比学习任务来进行模型训练.

3 基于多域 VQGAN 的国画生成

多模态预训练模型拥有强大的文字和图像语义理解能力, 可以很好地在文字和图像之间搭建起桥梁. 而 GAN 模型则能够根据随机的低维噪声生成效果逼真的图像. 因此, 本文将多模态预训练和 GAN 模型进行结合, 使用多模态预训练模型来指导 GAN 模型的图像生成过程. 具体地, 这种方法主要分为两个阶段: 使用多域 VQGAN 重构图像, 以及通过 WenLan 使用文字引导图像生成. 本文将分两节分别介绍这两个过程, 而在本节则主要介绍第 1 阶段多域 VQGAN 的设计及实现.

在以往的文生成图工作中, 许多模型都面临着如下的问题: (1) 需要收集大量成对的图像文本数据作为数据集, 代价昂贵. (2) 模型只能在特定的有限的域中生效, 面对多域生成时比较无力, 且难以生成高分辨率的图像. 因此, 在本文的工作中, 我们使用改进的多域 VQGAN 来进行图像的生成工作. 这样做的好处是, 我们在训练过程中, 无需配对的图文对, 只需要单独的图像即可, 大大减轻了数据集的构建压力. 经过改进的多域 VQGAN 能够接收多个域的训练数据集, 并生成多个域的图像, 且图像的质量及分辨率都会有提升.

生成对抗网络的目的是通过生成器和判别器的相互对抗训练, 使得生成样本的数据分布可以拟合真实样本的数据分布, 以此获得可以以假乱真的数据[47]. 与 VQVAE[8]类似, VQGAN 能够学习出一个有效的编码字典(codebook). 在生成图片时, 先随机生成特征图, 并在编码字典中进行量化, 再通过解码器解码就能得到效果比较真实的生成图像. 然而, 即使是同一类艺术形式, 也有很多不同的分类和风格, 比如国画又可分为水墨画, 设色画和白描画等, 本文将此定义为“域”. 如果将同属一种艺术形式, 但是不同域的图像混在一起训练生成模型, 那么最终生成的图像将无法区分其所属的域. 如果将不同域的图像分别作为不同模型的数据集, 那么不仅会有最终得到的模型数量太多的问题, 还可能会面临数据集较小导致模型过拟合的问题. 因此, 受到 StarGAN v2[48]的启发, 我们采用与其类似的分支方法, 将 VQGAN 的 3 个重要部分: 编码器, 编码字典和解码器进行了不同程度的共享和分支, 以便使用同一模型生成不同域的国画.

具体地, VQGAN 的模型结构分为 4 部分: 编码器、编码字典、解码器及判别器. 本文提出的多域 VQGAN 由 k 域编码器、k 域编码字典, k 域解码器及判别器组成. 当 k=1 时, 多域 VQGAN 与原 VQGAN 模型相同. 当 k>1 时: 对于 k 域编码器, 下采样的网络会被多个域共享, 而其余部分会作为每个域独有的输出分支; 似于编码器, 解码器将会共享上采样网络, 其余部分作为每个域独有的输入分支; 编码字典则作为一个整体, 被每个域独立拥有. 我们发现, 在多域训练时适当共享一定量的网络能够互相提高图片的质量、多样性和与文字的相关性, 在第 5 节我们对 3 个组件的不同组合情况的生成效果进行了详细比较.

多域 VQGAN 的网络结构及训练过程, 如图 4 所示. 以域数为 2 为例, 在构建图片数据集时将图片通过简单的打标签的方式来进行区分其所属的不同的域, 之后将两个域的数据混合在一起. 在训练过程中, 同一批量数据可能含有两个域的数据, 它们通过共享的上采样网络后, 就按照域的编号选择进入的分支, 最后再统一经过上采样过程, 得到最终的高分辨率重建图片.

4 WenLan 引导的文本生成国画

虽然多域 VQGAN 模型在训练结束后, 能够生成不同风格的图像, 但是其接收的输入是随机的低维噪声, 生成的结果也是完全随机的图像. 因此, 本文引入了中文多模态预训练模型 WenLan 来接收中文文本输入, 以达到用文本信息指导图像生成的目的.

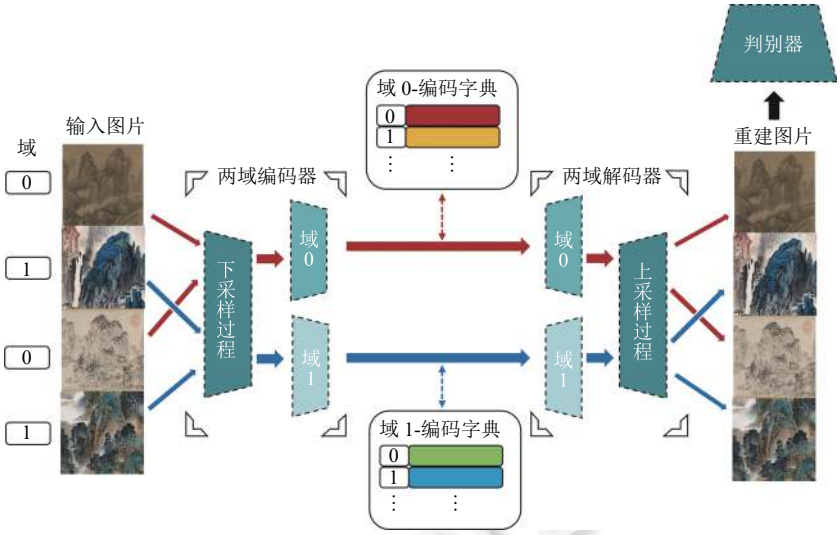


图 4 多域 VQGAN 的网络结构

WenLan 作为目前最大的中文多模态通用预训练模型, 在图文互检的任务上准确率和检索速度均取得出色的表现, 拥有很强的多模态语义理解能力, 能够很好地在中文文本和图像之间搭建起桥梁. 其提出的 BriVL 模型采用了双塔结构作为其模型架构, 将对比学习引入到了 BriVL 的双塔结构中. 其使用了独立的语言和视觉编码器来分别提取语言和视觉输入的特征向量, 在对比学习模块中将两种向量进行训练对齐. 这样的双塔结构允许我们将编码器模块替换为最新的单模态预训练模型, 从而加强模型的表达能力. 对于给定的某一对图文数据, BriVL 同时使用了视觉模态和语言模态去构建该图文数据的负样本, 并且基于 MoCo 的思想扩大了负样本数目, 从而进一步提高神经网络的表达能力. 而得益于 WenLan 的双塔结构, 预训练好的 WenLan 模型能够对图像和文本分别提取特征, 方便实际部署使用. 此外, WenLan 模型放松了对多模态间的强数据关联假设, 可以利用较为抽象、灵活的语义关联. 这种“弱相关”的假设在现实生活中更为普遍, 使得其模型的泛化能力变强, 也使得本文使用更加写意抽象的诗句来生成图片变得可能.

目前为止, WenLan 共发布了两个版本. 相比较于 WenLan 1.0[10]使用 3 000 万个图像文本对构成的数据集进行训练, WenLan 2.0[11]使用了 6.5 亿个来自互联网的弱相关图像文本对作为其数据集, 具有更加强大的泛化能力. 且 WenLan 2.0 移除了 WenLan 1.0 中的目标检测器, 是不依赖于物体检测结果的图文匹配模型, 减少了计算开销, 使得该工作更适合于实际工业界中的应用. 在实际应用中发现, WenLan 2.0 对于一些抽象概念, 具备着很符合人类直觉的理解能力, 比如“自然”, 其理解为大量的植被等; 对于“时间”, 其具象化理解为了一个钟表; 对于一些谚语和短语的理解也比较恰当. 因此, 本文使用 WenLan 2.0 来指导 VQGAN 的生成过程, 如图 5 所示. WenLan 2.0 由图像编码器和文本编码器组成, 其中图像编码器使用 EfficientNet[49]作为视觉主干网络, 文本编码器使用 RoBERTa-Large[50]作为文本主干网络. 在上述主干模型的输出后, BriVL 又堆叠了 4 个 Transformer[7]层, 以获得 2 560 维的视觉和文本特征, 并应用 InfoNCE 损失函数[51]将文本的特征与图像的特征对齐.

在进行图像生成时, 首先随机生成特征图, 并根据输入的域标签, 使用对应域的编码字典进行量化, 得到量化后的特征图, 并将其通过对应域的解码器生成随机的图像. 之后将生成的图像输入 WenLan 模型的图像编码器得到图像特征. 同时, 将输入的文本输入到 WenLan 模型的文字编码器得到文字特征. 最后计算图片特征和文字特征之间的相似性作为损失函数, 并通过梯度反传不断更新输入 VQGAN 的随机低维噪声来不断最小化该损失, 以达到生成图像的语义跟文本语义逐渐一致的效果. 如图 5 所示, 以某个域为例, 用 WenLan 2.0 指导 VQGAN 生成时的正向传播过程为: (1) 随机产生一个隐空间特征图, 将其使用对应域的编码字典进行量化, 得到量化后的隐空间特征图. (2) 将量化后的隐空间特征通过对应域的解码器生成图片. (3) 将生成的图片通过 WenLan 2.0 的图像编码

器得到图片的特征向量, 同时将输入的文本通过 WenLan 2.0 的文字编码器得到文本特征向量. (4) 计算图片特征向量和文本特征向量的余弦相似度, 衡量两者的距离作为损失函数. 而在反向传播过程 (虚线箭头) 中, 梯度经过 WenLan 2.0 和 VQGAN, 一直传播到随机隐空间特征图, 并通过梯度下降法更新随机隐空间特征图, 使得生成的图像语义越来越靠近输入文本描述的语义.

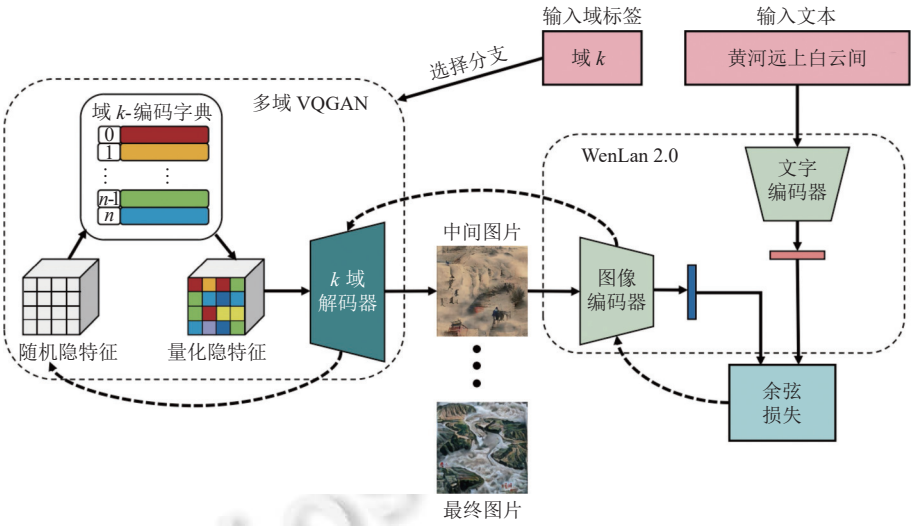


图 5 WenLan 指导的多域 VQGAN 生成过程

5 实验分析

5.1 实验数据

我们在本文中使用的图片数据集主要分为两部分: 水墨画和设色画.

● 设色画. 设色画数据集为我们自己构建的数据集. 我们通过各种渠道, 包括互联网、微信订阅号收集了 7000 余张原始数据. 我们花费了大量的时间来收集和整理这些数据: 第一, 人工过滤掉非风景以及低质量的画作. 第二, 人工裁剪出画心, 去除题跋. 第三, 针对画心又进行了裁剪, 去除提款等文字信息. 第四, 将每幅画作根据其最短边裁剪成多张正方形的画作, 并将其分辨率调整为 512×512. 总之, 最终得到的设色画数据共 2925 张. 图 6 展示了本文构建的设色画数据集.

图 6 设色画数据集采样示例

● 水墨画. 水墨画数据集包含 2 811 张 512×512 的图片. 其来源主要分为两部分: 一部分来自 Xue[52]公开的数据集, 其数据来源于开放访问的博物馆画廊: 史密森·弗里尔画廊、大都会艺术博物馆、普林斯顿大学艺术博物馆和哈佛大学艺术博物馆. 该数据集已经过该作者的手动过滤及裁剪, 我们在使用时又经过一次过滤, 最终共得到 1 979 张分辨率为 512×512 的图片. 水墨画数据集的另一部分为本文搜集的来源于互联网上的高分辨画作, 画作的年代主要分布于唐、明、清等, 经过人工过滤、去除文字信息以及裁剪等预处理工作后, 得到共计 832 张分辨率为 512×512 的图片. 图 7 采样展示了本文使用的水墨画数据集, 其中第 1 行为公开数据集, 第 2 行为本文构建的数据集.

图 7　水墨画数据集采样示例

5.2　评价指标

在本文中, 我们对于图像质量和图像文本对齐度进行了定量分析. 此外, 我们对用户主观评价进行了调研, 得到用户对于生成结果的评价结果作为定性分析.

● 图像质量分析. 我们采用 *FID* 指标[53]来评价图像质量. *FID* 是一种基于 Inception 网络的得分计算方法. Inception 网络是基于 ImageNet 数据集进行训练的图片分类器, 在计算 *FID* 指标时, 去掉了预训练 Inception-v3[54]的最后一个池化层, 来提取图片的特征, 之后测量真实图像分布和生成图像分布之间的距离. 为了计算 *FID*, 假设得到的图片特征遵循多维高斯分布, 那么就可以用均值和协方差矩阵来计算两个分布之间的距离. 公式 (3) 给出了真实数据和生成数据之间的 *FID*, 其中 $\left(\mu_r, \sum r\right)$ 为对真实图像提取的特征分布的均值和协方差, $\left(\mu_g, \sum g\right)$ 为生成图片特征分布的均值和协方差.

$$FID = \left\|\mu_r - \mu_g\right\|_2^2 + \mathrm{Tr}\left(\sum r + \sum g - 2\left(\sum r \sum g\right)^{1/2}\right) \tag{3}$$

FID 表示的是生成图像的特征向量与真实图像的特征向量之间的距离, 该距离越近, *FID* 越小, 说明生成模型的效果越好, 即图像的清晰度高, 且多样性丰富.

● 图像文本对齐度. 能否生成真实的图像只是评价文生成图模型的一个方面, 另一个重要方面是评估文本描述与生成的图像之间的语义对齐度. 因此本文采用了 *R*-precision 指标[4], 其原理是通过对生成图像的特征和文本特征之间的检索结果进行排序, 来衡量文本描述和生成的图像之间的视觉语义相似性. 对于每张生成图片特征. 计算其与文本集合中每个文本特征的余弦相似度, 并按照相似性递减的顺序对文本进行排序. 如果生成该图片使用的真实文本出现在前 *R* 名中, 则视为成功.

● 用户主观评价. *FID* 指标能够对图像质量和图像多样性做出定量评价, *R*-precision 指标能够很好地评价图像和文本相似性. 然而, *FID* 和 *R*-precision 在评价维度上还不够全面. 为此, 我们引入了用户主观评价, 能够在对象保真度、图像可解释性、人类常识等多维度方面对实验结果进行定性评价[55]. 在进行用户主观评价时, 首先使用生成模型从一定数量的随机抽样文本中生成一定数量的图片, 之后向用户提供文本和对应的生成图像, 由用户来选

出最优图像或进行排序.

5.3 消融实验

为了充分探讨 k 域编码器、k 域编码字典及 k 域解码器对于最终生成效果的影响, 我们设计了多种组合方式, 并在两个数据集上使用相同的参数进行训练, 并通过测试得到 *FID* 和 *R*-precision 指标.

● 实验设计. 本节共设计了 8 种网络结构: 原 VQGAN 结构, 即编码器、编码字典和解码器均为 1 域; 只将编码器、编码字典和解码器其中一个替换为两域结构, 其余均为 1 域结构; 将编码器、编码字典和解码器其中两个替换为两域结构, 其余为保持原 VQGAN 的结构; 编码器、编码字典和解码器均替换为两域结构. 模型具体部署情况见表 1.

我们以水墨画和设色画两个域作为训练模型的数据集, 以下所有的指标都是在两个域上各自进行计算得到的. 我们搜集了 200 句诗句, 从中随机挑选出了 100 句用于生成图片, 每句诗对应生成了 5 张图片, 共得到 500 张图片, 构成了我们生成图片的测试集. 其中, 诗句集的抽样示例如表 2 所示.

● *FID* 的计算. 在计算 *FID* 指标时, 真实图像集使用了训练集, 生成图像集使用了生成得到的 500 张图片. 将两个数据集合分别使用 Inception-v3 网络提取特征, 每张图片得到一个 2 048 维的向量, 之后利用公式 (3) 计算 *FID*.

● 检索召回实验. 在计算 *R*-precision 指标时, 我们使用生成的图像来召回他们对应的文字描述. 具体来说, 我们使用了生成的 500 张图像来对 200 句诗句做检索. 首先我们分别利用 WenLan 1.0 和 WenLan 2.0 提取出所有生成图像和所有文本的特征. 然后我们计算每一张生成图片的特征向量与全部文本特征的余弦相似度. 最后, 我们对每幅图片的候选文本描述按照相似性递减进行排序, 并使用最相似的文本来计算 *R*-precision 指标. 即, 只有图像对应的真实文本在相似度排名中排名第一时, 才会认为是文本召回成功.

本文的模型均在相同的超参数设置下运行了 300 轮. 模型每个部分的参数量如表 3 所示. 可以看出, 在使用两域的完整模型时, 较使用两个原 VQGAN 模型时, 节省了约 24M 的模型参数量. 而当域数增多时, 模型参数量的节省效果将会更加明显, 大大降低了对硬件的需求, 提高了训练效率.

表 1 模型的 8 种组合方式

模型	编码器是否为两域结构	编码字典是否为两域结构	解码器是否为两域结构
原VQGAN模型	×	×	×
两域编码器	√	×	×
两域编码字典	×	√	×
两域解码器	×	×	√
两域 (编码器+解码器)	√	×	√
两域 (编码器+编码字典)	√	√	×
两域 (编码字典+解码器)	×	√	√
完整模型	√	√	√

表 2 生成图像所用诗句示例

编号	诗句
文字0	白日依山尽, 黄河入海流.
文字1	不知香积寺, 数里入云峰.
文字2	轮台东门送君去, 去时雪满天山路.
文字3	恰似一江春水向东流.

表 3 模型各部分参数量

域	编码器 (M)	编码字典 (k)	解码器 (M)
一域	29.3	262	42.4
两域	41.1	524	54.3

两域的编码器、编码字典和解码器按照表 1 所示的 8 种组合情况各自训练, 并计算 *FID* 和 *R*-precision 指标. 具体实验结果见表 4. 可以看到, 在使用两个域的数据集分别去训练原 VQGAN 模型时, *FID* 值较低, 说明生成图像的质量较好. 但是无论用 WenLan 1.0 或者 WenLan 2.0, 其 *R*-precision 指标几乎是所有模型中最低的; 而当加入两域编码器或两域解码器后 (表 4 第 2、4 行), *R*-precision 指标有了明显提升. 说明将两个域的数据集联合起来训练时, 得益于使用两域编码器或两域解码器时的共享网络部分, 能够使得两个域的数据集互相帮助, 丰富各自的图片对象. 比如, 当输入文本是黄河时, 若设色画数据集中没有河流, 若使用原 VQGAN 模型是无法生成黄河的. 但是

使用部分共享网络, 将两个数据集联合起来训练时, 便可以借鉴到水墨画中的黄河或者河流信息, 以此可以提升两个域的生成图像与输入文本的语义一致性. 然而, 单独使用两域编码器或两域解码器时, 因为共享网络部分过多, (如单独使用两域编码器时, 编码器的下采样网络、编码字典和解码器会被共享), 两个数据集的分布也会互相影响, 导致 *FID* 指标较大; 而在单独使用两域编码字典时 (表 4 第 3 行), *FID* 值在 8 个模型最大, *R*-precision 指标在 7 个使用了两域组件的模型中也最低. 说明单独使用两域编码字典难以使得两个数据集互相帮助, 反而会每个域的分布混乱; 当只缺少两域解码器时 (表 4 第 6 行), *FID* 值较大, 说明生成图片时, 解码器在区分不同域的分布上起着较为重要的作用. 在表 4 第 2–4 行中, 只使用两域解码器时, *FID* 值也较只使用两域编码器或两域解码器要低也说明这个问题; 而表 4 第 5 行中, 即只缺少两域编码字典时, *R*-precision 指标较低, 说明在生成图片时, 每个域拥有自己独立的编码字典对于提升文本图像的语义一致性有较大帮助; 而当我们使用完整模型, 即图 4 所示的网络结构时, 无论是 *FID* 值还是 *R*-precision 指标都取得了最好的效果, 说明两个域的数据集能够很好地联合起来训练, 提高每个域的生成质量和效果.

表 4　模型在不同组合情况下的实验结果比较

模型	水墨画			设色画		
	FID	*R*-precision W2	*R*-precision W1	*FID*	*R*-precision W2	*R*-precision W1
原VQGAN模型	138.845	0.656	0.162	88.226	0.450	0.104
两域编码器	160.277	0.812	0.202	93.120	0.788	0.202
两域编码字典	164.789	0.660	0.168	105.573	0.684	0.158
两域解码器	154.890	0.780	0.168	89.793	0.768	0.194
两域 (编码器+解码器)	139.007	0.668	0.158	90.368	0.754	0.178
两域 (编码器+编码字典)	160.540	0.758	0.230	96.323	0.796	0.220
两域 (编码字典+解码器)	145.027	0.778	0.210	90.946	0.778	0.178
完整模型	**136.032**	**0.872**	**0.236**	**87.780**	**0.818**	**0.238**

为了更直观的比较两域编码器、编码字典和解码器的不同的组合方式对生成结果的影响, 以设色画为例, 图 8 给出了在不同模型下, 使用表 2 中的诗句生成的结果. 可以看出, 使用原 VQGAN 模型生成的图片 (图 8 第 1 列) 质量已经较好了, 但是使用完整的多域 VQGAN 模型生成的图片, 在颜色和对象丰富度上更胜一筹. 如使用文字 0: “白日依山尽, 黄河入海流”和文字 3: “恰似一江春水向东流”时, 完整 VQGAN 模型生成的图像中“春水”“黄河”的意象更加明确, 更能感受到黄河的澎湃汹涌和春水东流时的气势磅礴, 对于波浪、水纹的刻画也更加明显细致. 在使用文字 0 和文字 2 时, 完整 VQGAN 模型对于“依山尽”和“天山”的刻画也更细致, 而在原模型生成的图片中 (图 8 第 1 列的第 1、3 行), 基本看不出山的形状. 而在只使用部分两域组件的模型 (图 8 第 2–7 列) 生成图像中, 只使用两域编码字典时 (图 8 第 3 列) 的生成效果最差, 而使用两域编码器和两域解码器时 (图 8 第 5 列) 生成效果有了较大改善, 比较符合上述由表 4 得到的结论.

图 8　不同分支情况下多域 VQGAN 的生成结果

图 8 不同分支情况下多域 VQGAN 的生成结果 (续)

5.4 用户主观评价

为了更好地对我们模型使用诗句生成多域国画的效果进行评价, 在本节引入了用户主观评价. 我们针对水墨画和设色画两个域, 每个域随机选取了 60 句诗句. 针对每句诗句分别使用表 1 中的 8 种模型各生成一张图片. 为了确保实验的公平性, 生成图片的序号均被打乱. 我们召集了 7 位志愿者, 每位志愿者将从 8 张生成图片中选择出: (1) 最接近真实画作的图像. (2) 与文本描述最相关的图像. 用户评价的实验结果如表 5 所示. 可以看出, 使用完整模型时的生成效果要优于其他模型; 使用原 VQGAN 模型和只使用两域解码器时的效果较好; 而只使用两域编码字典 (表 5 第 3 行) 和使用两域编码器+两域编码字典 (表 5 第 6 行) 时的效果最差. 总之, 这些结论基本符合第 5.3 节中得到的结论.

表 5 用户主观评价研究结果 (%)

模型	水墨画			设色画		
	真实性	语义相关性	平均	真实性	语义相关性	平均
原VQGAN模型	20.34	17.97	19.15	19.66	19.32	19.49
两域编码器	7.12	10.84	9.98	4.40	7.45	8.12
两域编码字典	4.06	5.08	4.57	3.72	4.73	4.22
两域解码器	13.22	13.90	13.56	14.57	14.23	14.40
两域 (编码器+解码器)	12.20	10.84	11.52	14.23	11.18	12.71
两域 (编码器+编码字典)	4.41	8.14	6.27	3.38	7.45	5.41
两域 (编码字典+解码器)	12.20	10.17	11.19	11.52	10.16	10.84
完整模型	**26.44**	**23.05**	**24.75**	**28.47**	**25.42**	**26.95**

5.5 生成实例

为了更加直观地展示我们模型的生成效果, 我们给出了完整多域 VQGAN 模型结合 WenLan 2.0 生成的水墨画及设色画 (迭代 1 000 次), 同时给出了 CogView (在线 demo) 生成的水墨画图像, 对比结果如后文图 9–图 11 所示. 可以看出, CogView 生成的结果较为简单, 在色彩, 对象丰富程度, 刻画细节及语义相关性上均不如本文所提模型的生成结果. 总之, 我们的生成结果具有更强的视觉冲击度和更好的文本图像语义一致性.

本文所提方法, 在使用一张 Tesla V100-PCIE-32GB 的 GPU, 迭代次数设置为 1 000 次, 生成图片分辨率为 256×256 时, 生成每张图片的平均耗时约为 213.206 s, 在相同的硬件条件下, 我们复现了 DALL-E[5]中所述的方法, 模型参数的设置也与论文中的描述相同, 其平均生成一张 256×256 的图片的耗时为 569.482 s. 由此可见本文提出的模型在生成效率方面也较优.

6 总 结

利用文本描述生成图像是一项非常具有挑战性的任务. 本文提出了一种基于多域 VQGAN 的文本生成国画方法. 针对现有文本生成图像任务大多基于英文、生成图像的域比较有限以及训练需要大量标注的图像文本对等问题, 本文提出基于多域 VQGAN 来生成多域图像, 并结合 WenLan 使用文本指导多域 VQGAN 的生成过程, 使得能够使用中文诗句生成多域的中国画. 通过消融实验、用户主观评价以及将使用本文方法生成的图像与使用 CogView 中国水墨风格生成的图像进行对比等, 充分验证了本文所提出的方法能够较好地按照文本语义生成高分辨率的中国画.

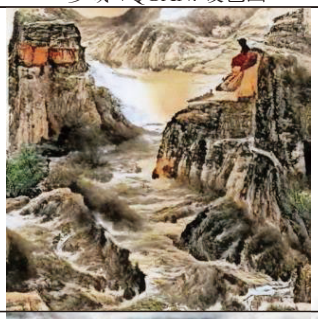
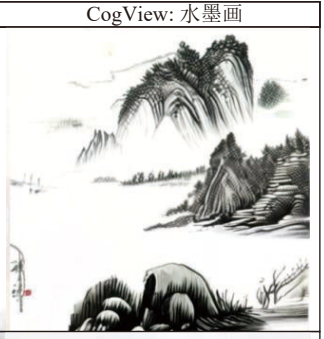
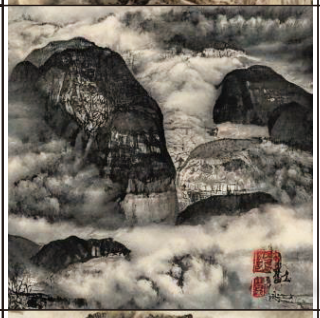
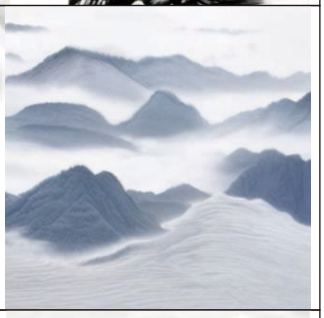
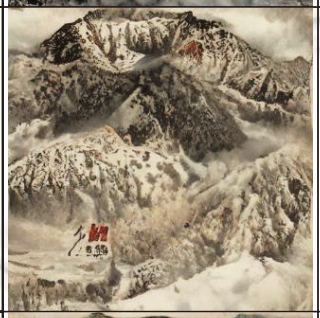
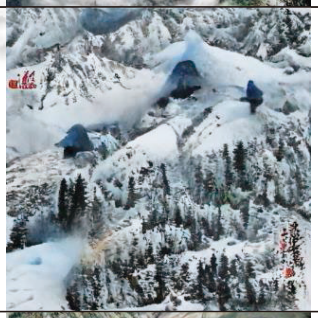
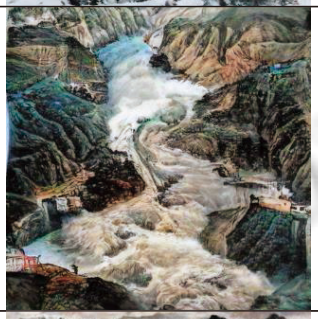
文本	多域 VQGAN: 水墨画	多域 VQGAN: 设色画	CogView: 水墨画
白日依山尽, 黄河入海流.			
除却巫山不是云			
雪景寒岭			
黄河远上白云间			
数峰无语立斜阳			

图 9　多域 VQGAN 与 CogView 生成结果对比 1

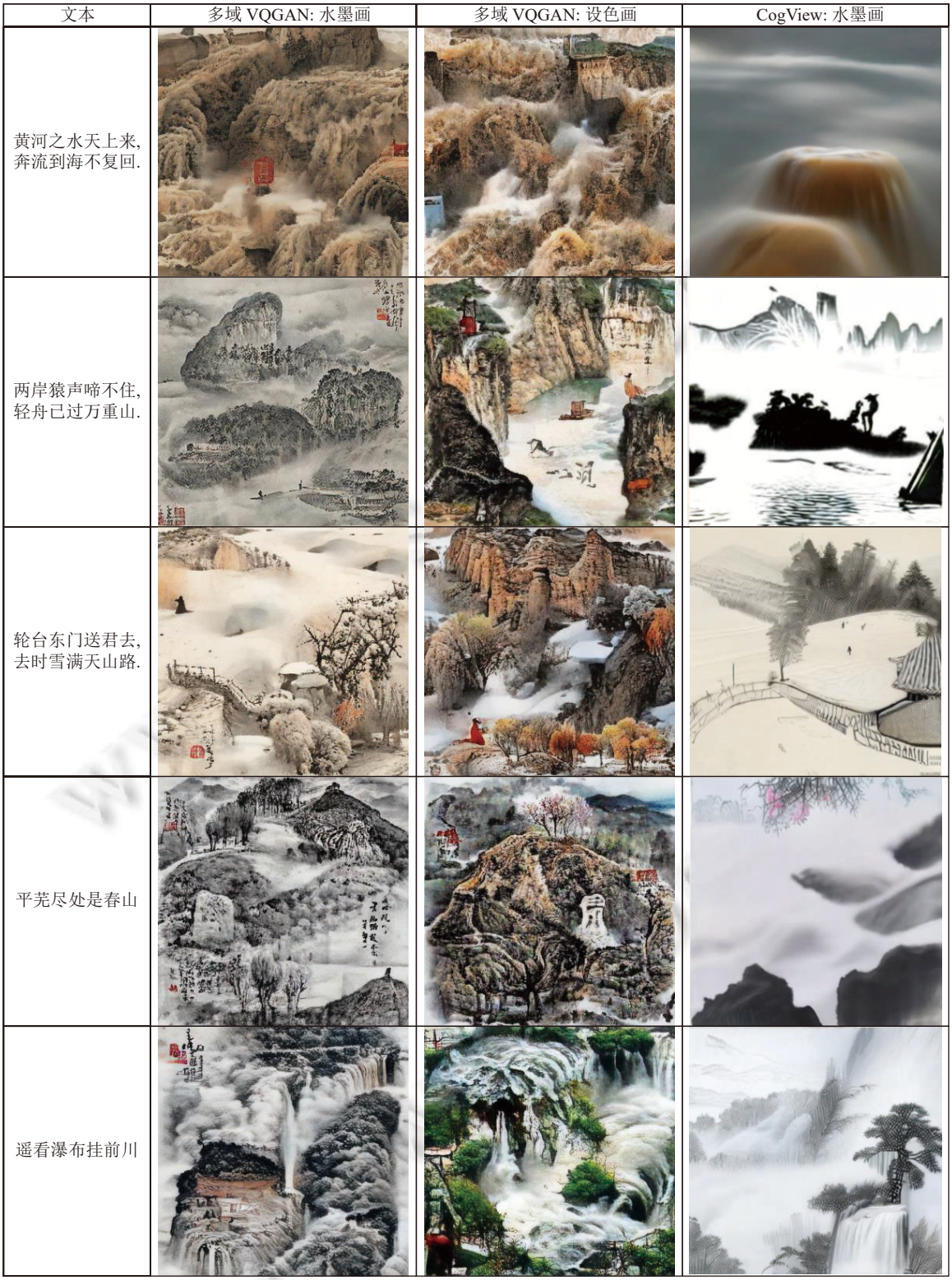


图 10　多域 VQGAN 与 CogView 生成结果对比 2

图 11　多域 VQGAN 与 CogView 生成结果对比 3

References:

[1] Kosslyn SM, Ganis G, Thompson WL. Neural foundations of imagery. Nature Reviews Neuroscience, 2001, 2(9): 635–642. [doi: 10.1038/35090055]

[2] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas D. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5907–5915. [doi: 10.1109/ICCV.2017.629]

[3] Chen ZL, Wang C, Wu HM, Shang K, Wang J. DMGAN: Discriminative metric-based generative adversarial networks. Knowledge-

based Systems, 2020, 192: 105370. [doi: 10.1016/j.knosys.2019.105370]

[4] Xu T, Zhang PC, Huang QY, Zhang H, Gan Z, Huang XL, He XD. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1316–1324. [doi: 10.1109/CVPR.2018.00143]

[5] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I. Zero-shot text-to-image generation. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8821–8831.

[6] Ding M, Yang ZY, Hong WY, Zheng WD, Zhou C, Yin D, Lin JY, Zou X, Shao Z, Yang HX, Tang J. CogView: Mastering text-to-image generation via transformers. In: Proc. of the 34th Advances in Neural Information Processing Systems. Curran Associates Inc., 2021. 19822–19835.

[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[8] van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6309–6318.

[9] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12873–12883. [doi: 10.1109/CVPR46437.2021.01268]

[10] Huo YQ, Zhang ML, Liu GZ, *et al*. WenLan: Bridging vision and language by large-scale multi-modal pre-training. arXiv:2103.06561, 2021.

[11] Fei NY, Lu ZW, Gao YZ, Yang GX, Huo YQ, Wen JY, Lu HY, Song RH, Gao X, Xiang T, Sun H, Wen JR. WenLan 2.0: Make AI imagine via a multimodal foundation model. arXiv:2110.14378, 2021.

[12] Gregor K, Danihelka I, Graves A, Rezende DJ, Wierstra D. DRAW: A recurrent neural network for image generation. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: PMLR, 2015. 1462–1471.

[13] Wu H, Xu D. Survey of digital image compositing. Journal of Image and Graphics, 2012, 17(11): 1333–1346 (in Chinese with English abstract). [doi: 10.11834/jig.20121101]

[14] Reed SE, Akata Z, Yan XC, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York City: PMLR, 2016. 1060–1069.

[15] Reed S, Akata Z, Mohan S, Tenka S, Schiele B, Lee H. Learning what and where to draw. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 217–225.

[16] Hu T, Li JL. Text to image generation based on single-stage GANs. Information Technology and Network Security, 2021, 40(6): 50–55 (in Chinese with English abstract). [doi: 10.19358/j.issn.2096-5133.2021.06.009]

[17] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas DN. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947–1962. [doi: 10.1109/TPAMI.2018.2856256]

[18] Bodla N, Hua G, Chellappa R. Semi-supervised FusedGAN for conditional image generation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 689–704. [doi: 10.1007/978-3-030-01228-1_41]

[19] Zhang ZZ, Xie YP, Yang L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6199–6208. [doi: 10.1109/CVPR.2018.00649]

[20] Gao LL, Chen DY, Song JK, Xu X, Zhang DX, Shen HT. Perceptual pyramid adversarial networks for text-to-image synthesis. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conf. and the 9th AAAI Symp. on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019. 1019. [doi: 10.1609/aaai.v33i01.33018312]

[21] Lai WS, Huang JB, Ahuja N, Yang MH. Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 624–632. [doi: 10.1109/CVPR.2017.618]

[22] Qiao TT, Zhang J, Xu DQ, Tao DC. MirrorGAN: Learning text-to-image generation by redescription. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1505–1514. [doi: 10.1109/CVPR.2019.00160]

[23] Tan XY, He XH, Wang ZY, Luo XD, Qing LB. Text-to-image generation technology based on Transformer cross attention. Computer Science, 2022, 49(2): 107–115 (in Chinese with English abstract). [doi: 10.11896/jsjkx.210600085]

[24] Creswell A, Bharath AA. Inverting the generator of a generative adversarial network. IEEE Trans. on Neural Networks and Learning Systems, 2019, 30(7): 1967–1974. [doi: 10.1109/TNNLS.2018.2875194]

[25] Abdal R, Qin YP, Wonka P. Image2StyleGAN: How to embed images into the StyleGAN latent space? In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4432–4441. [doi: 10.1109/ICCV.2019.00453]

[26] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4401–4410. [doi: 10.1109/CVPR.2019.00453]

[27] Abdal R, Qin YP, Wonka P. Image2StyleGAN++: How to edit the embedded images? In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8296–8305. [doi: 10.1109/CVPR42600.2020.00832]

[28] Voynov A, Babenko A. Unsupervised discovery of interpretable directions in the GAN latent space. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 9786–9796.

[29] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

[30] Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical Programming, 1989, 45(1): 503–528. [doi: 10.1007/BF01589116]

[31] Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation, 2001, 9(2): 159–195. [doi: 10.1162/106365601750190398]

[32] Zhu JY, Krähenbühl P, Shechtman E, Efros AA. Generative visual manipulation on the natural image manifold. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 597–613. [doi: 10.1007/978-3-319-46454-1_36]

[33] Huh M, Zhang R, Zhu JY, Paris S, Hertzmann A. Transforming and projecting images into class-conditional generative networks. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 17–34. [doi: 10.1007/978-3-030-58536-5_2]

[34] Guan SY, Tai Y, Ni BB, Zhu FD, Huang FY, Yang XK. Collaborative learning for faster StyleGAN embedding. arXiv:2007.01758, 2020.

[35] Kingma DP, Welling M. Auto-encoding variational bayes. arXiv:1312.6114, 2013.

[36] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

[37] Chen YC, Li LJ, Yu LC, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu JJ. UNITER: UNiversal image-text representation learning. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 104–120. [doi: 10.1007/978-3-030-58577-8_7]

[38] Li XJ, Yin X, Li CY, Zhang PC, Hu XW, Zhang L, Wang LJ, Hu HD, Dong L, Wei FR, Choi Y, Gao JF. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 121–137. [doi: 10.1007/978-3-030-58577-8_8]

[39] Lin JY, Men R, Yang A, Zhou C, Ding M, Zhang YC, Wang P, Wang A, Jiang L, Jia XY, Zhang J, Zhang JW, Zou X, Li ZK, Deng XD, Liu J, Xue JB, Zhou HL, Ma JX, Yu J, Li Y, Lin W, Zhou JR, Tang J, Yang HX. M6: A Chinese multimodal pretrainer. arXiv:2103.00823, 2021.

[40] Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557, 2019.

[41] Li G, Duan N, Fang YJ, Gong M, Jiang DX. Unicoder-Vl: A universal encoder for vision and language by cross-modal pre-training. Proc. of the 2020 AAAI Conf. on Artificial Intelligence, 2020, 34(7): 11336–11344. [doi: 10.1609/aaai.v34i07.6795]

[42] Su WJ, Zhu XZ, Cao Y, Li B, Lu LW, Wei FR, Dai JF. VL-BERT: Pre-training of generic visual-linguistic representations. arXiv:1908.08530, 2019.

[43] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

[44] Sun SQ, Chen YC, Li LJ, Wang SH, Fang YW, Liu JJ. LightningDOT: Pre-training visual-semantic embeddings for real-time image-text retrieval. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 982–997.

[45] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[46] Jia C, Yang YF, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung YH, Li Z, Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 4904–4916.

[47] Fu FF. Neural network methods for multi-style Chinese art paintings generation [MS. Thesis]. Chengdu: Sichuan University, 2021 (in Chinese with English abstract). [doi: 10.27342/d.cnki.gscdu.2021.000359]

[48] Choi Y, Uh Y, Yoo J, Ha JW. StarGAN v2: Diverse image synthesis for multiple domains. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8188–8197. [doi: 10.1109/CVPR42600.2020.00821]

[49] Tan MX, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 6105–6114.

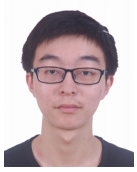
[50] Liu YH, Ott M, Goyal N, Du JF, Joshi M, Chen DQ, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized

BERT pretraining approach. arXiv:1907.11692, 2019.
[51] van den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.
[52] Xue A. End-to-end Chinese landscape painting creation using generative adversarial networks. In: Proc. of the 2021 IEEE Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3863–3871. [doi: 10.1109/WACV48630.2021.00391]
[53] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6629–6640.
[54] Szegedy C, Vanhoucke V, Ioffe S, Shles J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826. [doi: 10.1109/CVPR.2016.308]
[55] Frolov S, Hinz T, Raue F, Hees J, Dengel A. Adversarial text-to-image synthesis: A review. Neural Networks, 2021, 144: 187–209. [doi: 10.1016/j.neunet.2021.07.019]

附中文参考文献:

[13] 吴昊, 徐丹. 数字图像合成技术综述. 中国图象图形学报, 2012, 17(11): 1333–1346. [doi: 10.11834/jig.20121101]
[16] 胡涛, 李金龙. 基于单阶段GANs的文本生成图像模型. 信息技术与网络安全, 2021, 40(6): 50–55. [doi: 10.19358/j.issn.2096-5133.2021.06.009]
[23] 谈馨悦, 何小海, 王正勇, 罗晓东, 卿粼波. 基于Transformer交叉注意力的文本生成图像技术. 计算机科学, 2022, 49(2): 107–115. [doi: 10.11896/jsjkx.210600085]
[47] 付菲菲. 多风格国画生成的神经网络方法 [硕士学位论文]. 成都: 四川大学, 2021. [doi: 10.27342/d.cnki.gscdu.2021.000359]

孙泽龙(1999－), 男, 硕士生, 主要研究领域为机器学习, 文本生成图像.



费楠益(1997－), 男, 博士生, 主要研究领域为计算机视觉, 视觉语言多模态.

杨国兴(1998－), 男, 博士生, 主要研究领域为机器学习, 图像生成.

卢志武(1978－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 计算机视觉.

温静远(1998－), 男, 硕士生, 主要研究领域为多模态学习, 计算机视觉.

文继荣(1972－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为信息检索, 数据挖掘, 机器学习, 数据库.