

文章编号: 1671-7449(2024)02-0154-07

# 基于CLIP模型和文本重建的人脸图像生成方法研究

李源凡, 张丽红\*

(山西大学 物理电子工程学院, 山西 太原 030006)

**摘要:** 针对文本生成人脸方法中生成图像与文本描述不一致、图像分辨率较低等问题, 提出一种跨模态文本生成人脸图像网络框架。首先, 采用CLIP预训练模型对文本进行特征提取, 通过条件增强模块增强文本语义特征并生成隐藏向量; 然后, 将隐藏向量通过映射网络投影到预训练模型StyleGAN的隐式空间中获得解纠缠隐藏向量, 将该向量输入到StyleGAN生成器中生成高分辨率人脸图像; 最后, 采用文本重建模块将人脸图像重新生成文本, 计算重建文本和输入文本之间的语义对齐损失, 并将其作为语义监督指导网络训练。在Multi-Modal CelebA-HQ和CelebAText-HQ两个数据集上进行训练与测试, 实验结果表明, 相比其他方法, 该方法能生成更加符合文本描述的高分辨率人脸图像。

**关键词:** 文本生成人脸; 跨模态; CLIP预训练; 文本重建; 文本映射

**中图分类号:** TP391.4

**文献标识码:**

**doi:** 10.3969/j.issn.1671-7449.2024.02.008

**引用格式:** 李源凡, 张丽红. 基于CLIP模型和文本重建的人脸图像生成方法研究[J]. 测试技术学报, 2024, 38(2):154-160.

LI Yuanfan, ZHANG Lihong. Research on face image generation method based on CLIP model and text reconstruction[J]. Journal of Test and Measurement Technology, 2024, 38(2):154-160.

## Research on Face Image Generation Method Based on CLIP Model and Text Reconstruction

LI Yuanfan, ZHANG Lihong\*

(College of Physics and Electronic Engineering, Shanxi University, Taiyuan 030006, China)

**Abstract:** To address the problems of inconsistency between generated images and text descriptions and low image resolution in text-generated face methods, this paper proposes a cross-modal text-generated face image network framework. Firstly, the CLIP pre-training model is adopted to extract features from the text, and the text semantic features are enhanced by the conditional enhancement module to generate hidden vectors; then the hidden vector is projected into the implicit space of the pre-trained model StyleGAN by the mapping network to obtain the untangled hidden vector, which is input to the StyleGAN generator to generate high-resolution face images; finally, the text reconstruction module is adopted to regenerate the face images into text, and the semantic alignment loss between the reconstructed text and the input text is calculated and utilized as semantic supervision to guide the network training. The training and testing are performed on two datasets, Multi-Modal CelebA-HQ and CelebAText-HQ, and the experi-

收稿日期: 2023-03-15

基金项目: 山西省高等学校教学改革创新项目(J2021086); 山西省研究生创新项目(2021Y154)

作者简介: 李源凡(1998-), 男, 硕士生, 主要从事深度学习、人脸生成的研究。E-mail: 2508386569@qq.com。

\* 通信作者: 张丽红(1968-), 女, 教授, 主要从事计算机视觉、机器学习的研究。E-mail: lhzhang@sxu.edu.cn。

mental results show that compared with other methods, the method in this paper can generate high-resolution face images that are more consistent with the text description.

**Key words:** text-generated face; cross-modality; CLIP pre-training; text reconstruction; text mapping

## 0 引言

文本生成人脸旨在根据文本描述生成相应的人脸图像。目前,文本到图像的生成普遍采用多阶段框架。具体来说,利用文本编码器从文本描述中提取文本特征,文本特征与噪声相连接作为生成器的输入,再采用多个生成器逐级生成图像。例如,StackGAN和StackGAN++采用多个生成器和鉴别器逐步堆叠,进行多阶段训练,能够以从粗到细的方式生成图像<sup>[1,2]</sup>。AttnGAN在StackGAN的基础上开发了一种注意力生成网络和跨模态注意力机制,通过关注自然语言描述中的相关单词来合成图像不同子区域的细粒度细节<sup>[3]</sup>。MirrorGAN也采用了AttnGAN类似的框架,加入语义文本重建和对齐模块,进一步提高了生成图像的质量<sup>[4]</sup>。然而,训练一个多阶段的框架过程十分繁琐,计算量也较大。于是,DFGAN通过一

个仿射变换模块注入噪声,使用单个生成器来合成图像<sup>[5]</sup>。但是,上述方法一般生成的图像分辨率较低。本文采用预训练的StyleGAN生成器<sup>[6]</sup>,通过训练映射器将文本描述投射到隐式空间中,以提高生成图像的分辨率和保真度;采用基于对比文本-图像对的预训练方法(Contrastive Language-Image Pre-training, CLIP)提取文本特征,能够充分利用文本信息,考虑更细粒度的文本特征,对文本语义进行更深入的挖掘;另外网络中的文本重建模块增强了生成图像与文本描述的语义一致性。本文模型能根据文本描述达到更好的人脸图像生成效果。

## 1 整体框架

本文提出一种跨模态文本生成人脸图像的网络框架,如图1所示,整个网络框架主要由文本映射模块、StyleGAN生成器和文本重建模块组成。

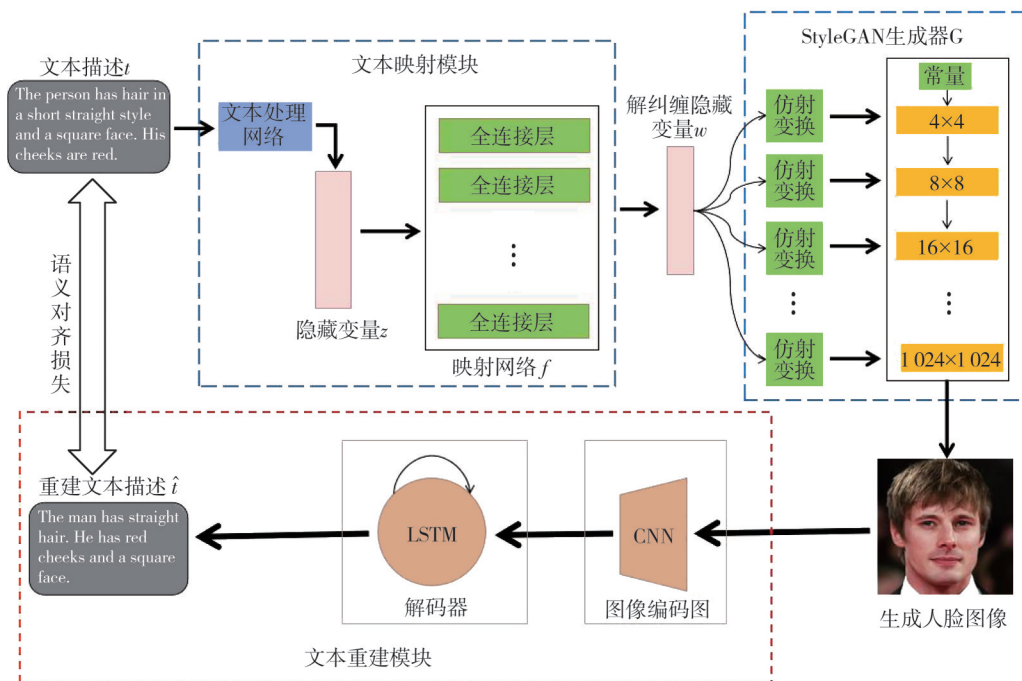


图1 整体网络框架

Fig. 1 Global network framework

其中,文本映射模块包括文本处理网络和映射网络两部分。首先,文本描述 $t$ 通过文本处理网络生成隐藏向量 $z$ ;然后,隐藏变量 $z$ 通过映射网络 $f$ 获得解纠缠隐藏向量 $w$ ,并且将解纠缠隐藏向

量 $w$ 输入StyleGAN生成器生成人脸图像。为了保证隐式空间中图像-文本语义的一致性,引入了文本重建模块来指导网络训练。文本重建模块能够从生成人脸图像中重建文本描述,并与原始文

本描述作比较,可以得到更好的梯度信息,从而保证生成人脸图像更符合文本描述,最终使得模型达到更好的生成效果。

## 2 文本映射模块

文本映射模块主要由文本处理网络和映射网络组成。文本描述通过文本映射模块投影到预训练模型StyleGAN的隐式空间获得解纠缠隐藏向量 $w$ 。

### 2.1 文本处理网络

文本处理网络主要由文本编码器和条件增强模块组成,如图2所示。本文采用CLIP预训练模型作为文本编码器对文本进行特征提取。文本编码器将文本描述编码成文本特征 $s$ ,该特征通过条件增强模块使得语义信息得到进一步增强,并转换成隐藏向量 $z$ 。

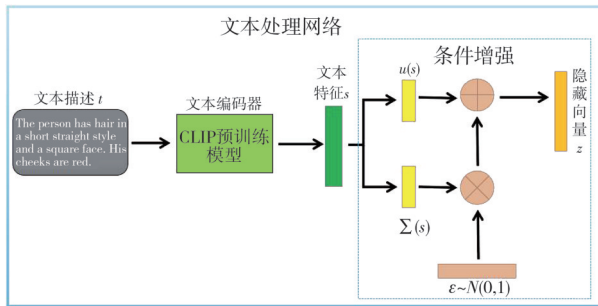


图2 文本处理网络

Fig. 2 Text processing network

#### 2.1.1 CLIP模型

本文采用CLIP预训练模型作为文本编码器,从文本描述中提取文本特征。该模型能够处理更加复杂的文本描述,融合文本信息和视觉信息,生成与文本语义一致的人脸图像。

CLIP预训练模型<sup>[7]</sup>将自然语言和视觉信息进行结合。在训练过程中,模型的输入是文本-图像对,图片通过图像编码器得到图像特征,同理,文本通过文本编码器得到文本特征,CLIP预训练模型通过文本特征和图像特征作对比学习,而对比学习需要定义正样本和负样本。定义配对的文本-图像对为正样本,其余的为负样本,如图3所示,特征矩阵里沿对角线元素都是正样本,非对角线元素都是负样本。

首先,文本描述 $t$ 通过文本编码器编码成512维特征向量,即文本特征

$$s = CLIP(t), \quad (1)$$

式中: $t = \{t_1, \dots, t_i, \dots, t_C\} \in \mathbb{R}^{S \times C}$ 为文本描述编码,CLIP为文本编码器, $t_i$ 为一个单词的独热编码,

$C$ 为句子长度, $S$ 为词汇表大小。

CLIP预训练模型具有以下优势:很好地将视觉语义和文本语义联系到一起,学习到较强的语义特征,从而达到对文本语义进行更深入的挖掘;具有很好的迁移学习能力,能够在任意数据集上取得非常好的效果。

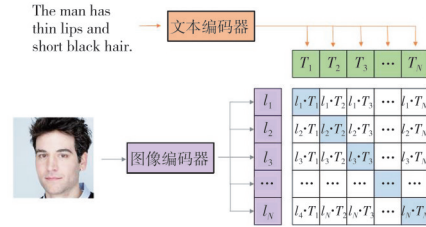


图3 CLIP预训练模型

Fig. 3 CLIP pre-training model

#### 2.1.2 条件增强模块

条件增强模块可以丰富文本特征的语义信息,增强文本语义特征。如图2中条件增强模块所示,将文本特征 $s$ 输入条件增强模块 $F^{ca}$ 生成一个隐藏向量 $z \in \mathbb{R}^{512}$

$$z = F^{ca}(s). \quad (2)$$

隐藏向量 $z$ 从高斯分布 $N(u(s), \Sigma(s))$ 中随机采样,其中 $u(s)$ 和 $\Sigma(s)$ 是文本特征 $s$ 的均值和协方差。Karras等<sup>[6]</sup>已经证明了 $w$ 解纠缠隐藏向量可以通过学习非线性映射 $F: z \rightarrow w$ 得到。 $z$ 需要遵循先验分布,例如标准的高斯分布 $\epsilon \sim N(0, 1)$ 。为了使隐藏向量 $z$ 的近似分布 $N(u(s), \Sigma(s))$ 与标准高斯分布对齐 $\epsilon \sim N(0, 1)$ ,引入KL散度,KL散度为近似分布 $N(u(s), \Sigma(s))$ 对标准高斯分布 $\epsilon \sim N(0, 1)$ 的相对熵,用 $L_{KL}$ 表示,其公式为

$$L_{KL} = D_{KL}(N(\mu(s), \Sigma(s)) // N(0, 1)). \quad (3)$$

同时,隐藏向量 $z$ 从高斯分布中随机采样,也有助于生成更多样化的图像。

### 2.2 映射网络

映射网络由8个全连接层组成,将文本处理网络中输出的隐藏向量 $z$ 作为映射网络 $f$ 的输入,得到解纠缠隐藏向量 $w \in \mathbb{R}^{512}$ 。并将向量传入StyleGAN生成器得到18个控制向量,使得每个控制向量能够控制不同的视觉特征。映射网络能使得输入向量不受数据集分布的影响,更好地实现属性解耦合。

## 3 StyleGAN生成器

本文采用StyleGAN生成器生成高分辨率人

脸图像。StyleGAN 生成器具有以下优势: 用  $4 \times 4 \times 512$  常量代替传统输入, 可以降低生成不正常照片的概率, 减少特征纠缠, 使得网络更容易学习; 由 9 个生成阶段组成, 从生成  $4 \times 4$  分辨率图像, 到生成  $8 \times 8$  分辨率图像, 最终生成  $1024 \times 1024$  高分辨率图像; 对噪声和扰动的鲁棒性较好, 可以在不同的条件下生成稳定的图像。

## 4 文本重建模块

为了保证文本描述与生成图像之间的语义一

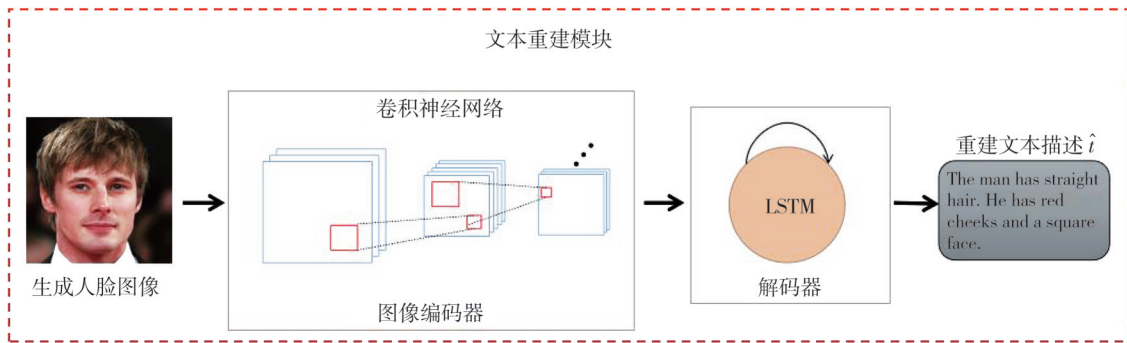


图 4 文本重建模块

Fig. 4 Text reconstruction module

采用在 ImageNet 数据集上预训练的卷积神经网络作为图像编码器, 将人脸图像嵌入到特征向量  $\mathbf{a}$  中

$$\mathbf{a} = \{a_1, \dots, a_i, \dots, a_L\}, a_i \in \mathbb{R}^D, \quad (4)$$

式中:  $L$  表示向量的个数;  $D$  表示向量的维数。通过软注意力机制从特征向量  $\mathbf{a}_i$  中计算上下文向量  $\hat{z}_k$

$$\hat{z}_k = \sum_{i=1}^L (\beta_i a_i), \quad (5)$$

式中: 下标  $k$  表示 LSTM 的第  $k$  个时间步长,  $\hat{z}_k \in \mathbb{R}^D$  则是图像在第  $k$  个时间步长的相关区域的动态表示。  $\alpha = \{\beta_1, \dots, \beta_i, \dots, \beta_L\}$  表示特征  $a$  的权重。例如, 时间步长  $k$  时的权重  $\beta_{ki}$  计算为

$$\hat{e}_{ki} = f_{att}(\mathbf{a}_i, h_{k-1}), \quad (6)$$

$$\beta_{ki} = \frac{\exp(\hat{e}_{ki})}{\sum_{j=1}^L \exp(\hat{e}_{kj})}, \quad (7)$$

式中:  $f_{att}$  代表多层感知机组成的注意力模型, 而  $h_{k-1} \in \mathbb{R}^n$  为 LSTM 前一时刻单元的隐藏输出。根据上下文向量  $\hat{z}_k$ 、隐藏状态  $h_k$  和之前生成的单词  $t_{k-1} \in \mathbb{R}^S$ , 计算输出的单词在当前时间步长  $k$  处的概率

$$p(\hat{t}_k | a, \hat{t}_1^{k-1}) \propto \exp(L_o(E\hat{t}_{k-1} + L_h h_k + L_z \hat{z}_k)), \quad (8)$$

式中:  $L_o \in \mathbb{R}^{S \times m}$ ,  $E \in \mathbb{R}^{m \times S}$ ,  $L_h \in \mathbb{R}^{m \times n}$  和  $L_z \in \mathbb{R}^{m \times D}$  是随机初始化的参数。为了丰富生成文本的

致性, 本文采用了文本重建模块, 该模块由图像编码器和解码器组成, 使用卷积神经网络 (Convolutional Neural Networks, CNN) 作为图像编码器, 使用长短期记忆人工神经网络 (Long and Short-Term Memory Artificial Neural Networks, LSTM)<sup>[8]</sup> 作为解码器。如图 4 所示, 通过图像编码器将生成人脸图像编码为视觉特征, 再将视觉特征输入解码器进行解码重新生成文本, 最后计算重建文本描述和输入文本描述之间的语义对齐损失。

语义信息, 采用双重随机正则化, 使得  $\sum_{i=1}^L \beta_{ji} = 1$  和  $\sum_{j=1}^C \beta_{ji} \approx 1$ 。文本重建模块损失函数

$$L_{TRL} = -\sum_{j=1}^C \log p(\hat{t}_j) t_j + \sum_{i=1}^L (1 - \sum_{j=1}^C \beta_{ji})^2. \quad (9)$$

文本重建模块用于将 StyleGAN 生成器中生成的人脸图像重新生成新的文本描述, 再计算重建文本和输入文本之间的语义对齐损失, 并将其作为语义监督指导网络训练, 使得生成的人脸图像更符合文本描述。

## 5 实验结果及分析

### 5.1 数据集

采用 Multi-Modal CelebA-HQ 和 CelebA Text-HQ 两个数据集。Multi-Modal CelebA-HQ 和 CelebA Text-HQ 数据集的人脸图像都来自 CelebA-HQ 数据集, 包含 30 000 幅图像, 每幅图像有 10 种语言描述, 其中 24 000 幅图像作为训练集, 6 000 幅图像作为测试集。前者人脸图像的文本描述由属性标签合成, 而后者人脸图像的文本描述则是由手动标注完成。

### 5.2 损失函数

本文采用预训练的 StyleGAN 模型作为人脸

图像生成器,在训练过程中不需要计算鉴别器损失和生成器损失。同时,本文采用文本重建模块作语义监督,引导网络更快地收敛。整个网络的损失函数定义为

$$L = \lambda_1 L_{KL} + \lambda_2 L_{TRL} = \lambda_1 D_{KL}(N(\mu(s), \sum(s) \parallel N(0, 1)) - \lambda_2 \sum_{j=1}^C \log p(\hat{t}_j) t_j + \lambda_2 \sum_{i=1}^L \left(1 - \sum_{j=1}^C \beta_{ji}\right)^2). \quad (10)$$

设置参数  $S=30\ 000$ ,  $C=50$ ,  $L=196$ ,  $D=2\ 048$ ,  $n=512$  和  $m=512$ 。式中:  $\lambda_1$  和  $\lambda_2$  是控制各项权重的超参数。在整个训练过程中,设置损失函数的超参数  $\lambda_1=1$  和  $\lambda_2=5$ 。

### 5.3 评价指标

本文通过弗雷歇初始距离(Fréchet Inception Distance, FID)和学习感知图像块相似度(Learned Perceptual Image Patch Similarity, LPIPS)<sup>[9]</sup>来评估文本生成人脸图像模型的性能。

FID表示在特征空间中生成图像的特征向量和真实图像的特征向量之间的距离。FID分数越低,表示生成的图像视觉质量越好。如式(11)

$$FID(r, g) = \left\| \mu_r - \mu_g \right\|_2^2 + \text{trace} \left( \sum_r + \sum_g - 2 \left( \sum_r \sum_g \right)^{\frac{1}{2}} \right), \quad (11)$$

式中:  $r$  表示真实人脸图像;  $g$  表示生成人脸图像;  $\mu$  和  $\sum$  分别表示相应人脸图像数据分布的平均值和协方差。  $\|*\|_2^2$  表示  $L_2$  范数的平方,  $\text{trace}$  为矩阵的迹。

LPIPS表示真实样本和生成样本的特征差异,这个差异是在每个通道内计算  $L_2$  范数,再在所有通道上加权平均求和。LPIPS主要用于度量两张图像之间的差别, LPIPS的值越低,表示两张图像越相似,反之,则差异越大。如式(12)

$$d(r, g) = \sum_l \frac{1}{H_l W_l} \sum_{h, w} \left\| \mathbf{w}_l \odot (y_{hw}^l - y_{hw}^{l'}) \right\|_2^2, \quad (12)$$

式中:  $d$  表示真实人脸图像和生成人脸图像的特征差异;  $l$  表示卷积层;  $H, W$  表示特征空间两个通道维度;  $\sum$  表示求和;  $\mathbf{w}_l$  表示激活向量;  $\odot$  表示作点积运算;  $y_{hw}^l, y_{hw}^{l'}$  表示从  $l$  层提取特征堆栈并在通道维度中进行单位归一化的结果。

### 5.4 实验结果

本文采用SGD优化器,学习率设置为

0.0001, 动量参数  $\beta$  设置为0.9。对Multi-Modal CelebA-HQ数据集进行500个轮次的训练,批量处理数量为12,对CelebAText-HQ数据集进行300个轮次的训练,批量处理数量为6。对本文网络进行训练和测试,部分实验结果如图5、表1~表3所示。

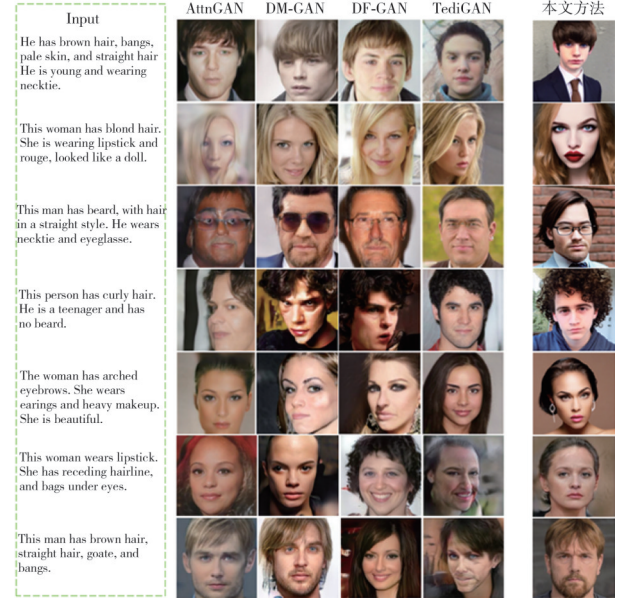


图5 人脸生成结果对比图

Fig. 5 Face generated result comparison map

表1 本文方法与其他方法FID指标比较

Tab. 1 Comparison of FID index of the method in this paper with other methods

| 模型      | FID指标                 |               |
|---------|-----------------------|---------------|
|         | Multi-Modal CelebA-HQ | CelebAText-HQ |
| AttnGAN | 41.33                 | 70.47         |
| DM-GAN  | 38.35                 | 70.89         |
| DF-GAN  | 31.54                 | 64.25         |
| TediGAN | 28.62                 | 60.58         |
| 本文方法    | 22.81                 | 60.52         |

表2 本文方法与其他方法LPIPS指标比较

Tab. 2 Comparison of LPIPS index of the method in this paper with other methods

| 模型      | LPIPS指标               |               |
|---------|-----------------------|---------------|
|         | Multi-Modal CelebA-HQ | CelebAText-HQ |
| AttnGAN | 0.548                 | 0.524         |
| DF-GAN  | 0.482                 | 0.457         |
| TediGAN | 0.456                 | 0.448         |
| 本文方法    | 0.446                 | 0.431         |

图5为人脸生成结果对比图,对AttnGAN, DM-GAN<sup>[10]</sup>, DF-GAN, TediGAN<sup>[11]</sup>和本文方法生成的人脸图像进行比较,可以看出现有的一些

文本生成人脸图像方法难以生成高质量且符合文本描述的人脸图像。例如, AttnGAN 第 2 行生成的人脸面孔比较模糊, 缺少了一些细节; AttnGAN 和 DF-GAN 第 4 行生成的人脸图像头部姿势过度旋转; DM-GAN 第 4 行和 TediGAN 第 6 行、第 7 行生成的人脸面部形状扭曲; DF-GAN 第 7 行生成的人脸性别错误, 同时还可以观察到 AttnGAN 第 7 行、DM-GAN 第 6 行、DF-GAN 第 3 行以及 TediGAN 第 3 行和第 5 行生成的人脸图像不符合文本描述, 与语义不一致。相比之下, 本文方法生成的人脸图像质量更好, 更符合文本描述。

表 3 消融结果对比  
Tab. 3 Comparison of ablation results

| 模型              | <i>FID</i> | <i>LPIPS</i> |
|-----------------|------------|--------------|
| Baseline        | 44.31      | 0.442        |
| Baseline+文本重建模块 | 22.81      | 0.431        |

表 1 为本文方法与其他方法在 Multi-Modal CelebA-HQ 数据集和 CelebA-Text-HQ 数据集上的 *FID* 指标比较, 由表 1 可知, 本文方法在 Multi-Modal CelebA-HQ 数据集上具有较低的 *FID* 评分, 在 CelebA-Text-HQ 数据集上 *FID* 评分与 TediGAN 的 *FID* 评分相近。结果表明, 本文方法能够生成质量较好的图像。

表 2 为本文方法与其他方法在 Multi-Modal CelebA-HQ 数据集和 CelebA-Text-HQ 数据集的 *LPIPS* 指标比较, 结果表明, 本文模型生成的人脸图像更符合文本描述。

在 CelebA-Text-HQ 测试集上进行了消融研究, 验证了文本重建模块的有效性。Baseline 表示不引入文本重建模块的网络模型, Baseline+文本重建模块表示引入该模块的网络模型。如表 3 所示, 引入该模块显著提升了网络的性能。

由于本文方法在基线模型基础上增加了文本重建模块, 使得网络模型参数量增多, 相比基线模型, 该方法运算时间较长。但实验结果表明, 本文方法生成的人脸图像质量更好且更符合文本描述。

### 5.5 损失函数曲线

如图 6 所示, 本文网络在 Multi-Modal CelebA-HQ 数据集上迭代了 50 000 次, 损失函数值趋向于 1.2, 网络趋向收敛。

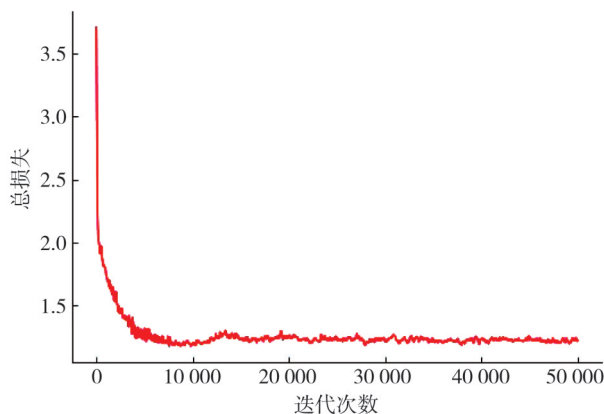


图 6 损失函数曲线  
Figure. 6 Loss function curve

## 6 结语

本文设计了一种基于 CLIP 预训练模型和文本重建模块的文本生成人脸图像网络, 网络主要由文本映射模块、StyleGAN 生成器和文本重建模块组成。首先, 文本描述通过文本映射模块投影到预训练模型 StyleGAN 的隐式空间, 获得解纠缠隐藏向量, 再将向量输入 StyleGAN 生成器, 生成一个高分辨率的人脸图像。采用文本重建模块作语义监督, 增强了生成图像与文本描述的语义一致性。实验结果表明, 基于 CLIP 预训练模型和文本重建模块相结合的方法优于其他网络模型。

### 参考文献:

[ 1 ] ZHANG H, XU T, LI H, et al. Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks [C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 5907-5915.

[ 2 ] ZHANG H, XU T, LI H, et al. Stackgan++: realistic image synthesis with stacked generative adversarial networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018, 41(8): 1947-1962.

[ 3 ] XU T, ZHANG P, HUANG Q, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1316-1324.

[ 4 ] QIAO T, ZHANG J, XU D, et al. Mirrorgan: learning text-to-image generation by redescription [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 1505-1514.

- [ 5 ] TAO M, TANG H, WU F, et al. Df-gan: a simple and effective baseline for text-to-image synthesis[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16515-16525.
- [ 6 ] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4401-4410.
- [ 7 ] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning, PMLR, 2021: 8748-8763.
- [ 8 ] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016(2): 207-212.
- [ 9 ] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4681-4690.
- [10] ZHU M, PAN P, CHEN W, et al. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 5802-5810.
- [11] XIA W, YANG Y, XUE J H, et al. Tedigan: Text-guided diverse face image generation and manipulation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 2256-2265.