arXiv:2304.00964v1 [cs.CV] 3 Apr 2023

一种鲁棒的文本驱动图像编辑方法：自适应探索 StyleGAN 和 CLIP 潜在空间中的方向

Tsuyoshi Baba*† Kosuke Nishida*‡ Kyosuke Nishida‡
†警察学院警察信息通信研究中心
‡NTT株式会社人类信息学实验室{tsuyoshi.baba.um, kosuke.nishida.ap, kyosuke.nishida.rx}@hco.ntt.co.jp

摘要

自动图像编辑有着广泛的应用需求，而使用自然语言指令对于实现用户所设想的灵活、直观的编辑至关重要。作为文本驱动图像编辑领域的先驱工作，*StyleCLIP* 在 *CLIP*空间中寻找编辑方向，然后将该方向映射到 *StyleGAN*空间，以此编辑图像。同时，除了原始图像和文本指令之外，很难为图像编辑调整出合适的输入。在本研究中，我们提出了一种利用支持向量机构建方法，可在 *StyleGAN*空间和 *CLIP*空间中自适应地构建编辑方向。我们的模型将编辑方向表示为 *CLIP*空间中的法向量，该法向量通过训练支持向量机对正样本图像和负样本图像进行分类而获得。这些图像从一个大规模图像语料库中检索得到；该语料库最初用于 *StyleGAN* 的预训练，检索依据是图像与文本指令之间的 *CLIP* 相似度。我们确认，尽管该模型允许使用简单的输入而不会增加计算时间，但其性能与 *StyleCLIP* 基线相当。



图 1. 使用 StyleCLIP-FEU 和 StyleCLIP 编辑的图像示例。StyleCLIP 的中性文本 t_0 分别为 (a)“面部”、(b)“男性面部”或“女性面部”，以及 (c) 详细图注。StyleCLIP-FEU 无需 t_0 即可根据指令编辑更多属性。StyleCLIP-FEU 中明显失败的部分以红色标出。

1. 引言

自动图像编辑因其众多应用而需求广泛。生成对抗网络（GANs） [6] 是合成高分辨率图像最常用的方法之一，并通过使用预定义属性 [2, 8] 和参考图像 [4, 32]被广泛用于自动图像编辑。

使用自然语言指令进行图像编辑，对于开发易用的应用程序至关重要，这类应用能够按照用户的设想实现灵活、直观的编辑。这项具有挑战性的任务涉及对图像和自然语言的多模态理解。在此，图文对比预训练（CLIP）模型将视觉和语言概念嵌入多模态特征空间 [19]。因此，CLIP 为探索完成这项任务的方法提供了契机。

开创性工作 StyleCLIP [18]在 StyleGAN 特征空间中向原始图像表示添加一个向量，而该空间已知具有良好的解耦性 [3, 7, 26, 30]。该工作提出了两类方法：第一类包括潜变量优化和潜变量映射器，通过最大化输出图像与给定自然语言指令之间的 CLIP 相似度，计算穿过神经网络的反向传播。因此，这些方法针对每次用户输入都需要较高的计算成本。

第二种类型称为全局方向，它在 CLIP空间中获取一个编辑方向，然后将其映射到 Style-

*共同贡献。此项工作完成于 Tsuyoshi Baba 在 NTT 实习期间。

GAN 特征空间。编辑方向被定义为使用 CLIP 文本编码器从原始图像的文本描述指向指令的向量。因此，用户必须手动指定原始图像的描述。

为了打造一款快速易用（FEU）的图像编辑应用，我们提出了 StyleCLIP-FEU，通过解读 StyleGAN 和 CLIP 空间中的潜在语义，采用第二种方法。受 Interface-GAN 启发 [25], ，为避免手动描述原始图像，我们将编辑方向定义为一个法向量，该向量通过支持向量机（SVM）学习，用于对利用 CLIP 即时检索到的正、负图像进行分类。尽管有些论文扩展了 StyleCLIP [1, 9, 27, 33–35], ，但除 [14],采用第一种方法并加快了优化过程，以及 [24], 采用第二种方法并进一步进行预训练，以对齐 StyleGAN 空间和 CLIP 空间之外，这些论文并未关注上述问题。

实验结果表明，在自然语言指令的引导下，StyleCLIP 能够自然地编辑人脸图像。此外，我们观察到 StyleCLIP 会受到原始图像描述的影响。

2. 预备知识

2.1. 任务定义

我们将通过自然语言指令进行图像编辑定义为一项任务：模型根据原始图像 x ，按照自然语言指令 t 生成图像 x'。此外，我们假设有一个可用的图像语料库（不含任何标注） I 。

2.2. CLIP

CLIP [19] 会将提及相似概念的图像和文本嵌入到多模态特征空间 M 中，使它们彼此接近。CLIP 包含一个图像编码器 $\text{CLIP}^{\text{image}}$ 和一个文本编码器 $\text{CLIP}_{\text{text}}$。它分别通过两个编码器计算图像嵌入 $\boldsymbol{h}_{\text{image}}$ 和文本嵌入 $\boldsymbol{h}_{\text{text}}$ 。CLIP 使用从网站收集的 4 亿组成对图像及其图像说明文字进行训练，训练目标是使正样本对之间的 $\cos(\boldsymbol{h}_{\text{image}}, \boldsymbol{h}_{\text{text}})$ 高于批内负样本对之间的值。

2.3. StyleGAN

StyleGAN [11,12]能够生成由风格控制的图像，这是因为它具有解耦特征空间。它由映射模块和合成模块组成。每个模块都由一个 N_{map} 层、 N_{syn} 层神经网络构成。StyleGAN 首先通过映射模块，从高斯噪声 $\boldsymbol{z}$ 中获得潜在特征 $\boldsymbol{w} \in W$ 。随后，它将 $\boldsymbol{w}$ 仿射变换为 $\boldsymbol{s}_l (l = 1, \cdots, N_{\text{syn}})$，并将这些特征输入合成模块的第 l 层。所有层中的 $\boldsymbol{s}_l$ 拼接后得到的 $\boldsymbol{s} \in S$，称为风格特征。

2.4. StyleCLIP

[18]通过结合 StyleGAN 和 CLIP，开发了 StyleCLIP，以便使用自然语言指令进行图像编辑。StyleCLIP 包含三种方法：潜变量优化、潜变量映射器和全局方向。其中，我们重点关注全局方向方法，因为其他方法需要针对每个 t优化神经网络，而且作者报告称，全局方向中的特征具有最高的解耦程度。

[18] 将编辑定义为 CLIP空间中的一个方向 M，从中性文本（描述原始图像的文本，例如，“有头发的面部”） t_0 指向目标文本（指令：“有金发的面部”） t。因此，除了我们任务定义中规定的输入外，StyleCLIP 还需要中性文本。该模型将 M 中的编辑方向映射为一个向量 $\Delta \boldsymbol{s} \in S$。随后，合成模块 G 生成编辑后的图像$x' = G(\boldsymbol{s} + \alpha \Delta \boldsymbol{s})$，其中 $\boldsymbol{s}$ 是通过 E4E 等图像编码器对原始图像 x 进行编码得到的风格特征，而 α 是一个超参数。

Δs_c 表示 M 中的编辑方向与解耦风格空间 S中第 c个通道对应属性的相关程度。因此，StyleCLIP 根据 M中的两个向量$\Delta \boldsymbol{i}_c$ 和 $\Delta \boldsymbol{t}$的点积来定义Δs_c 。其中，$\Delta \boldsymbol{i}_c$ 是将 S中通道c的方向映射得到的向量。 $\Delta \boldsymbol{t}$ 是表示编辑的向量。

构建 $\Delta \boldsymbol{i}_c$ 首先，StyleCLIP 获取与风格空间的通道 c 对应的 CLIP 空间方向 $\Delta \boldsymbol{i}_c$ 。StyleCLIP 从图像语料库 I中采样 100 张图像 $x_i (i = 1, \cdots, 100)$，并使用 E4E 编码器提取其风格特征 $\boldsymbol{s}_i$ 。模型计算标准差 $\boldsymbol{\sigma} \in \mathbb{R}^C$，其中 C 是通道数。模型扰动图像特征 $\boldsymbol{s}_i$ 的通道 c ，并生成成对图像：

$$x_{i,c}^{\pm} = G(\boldsymbol{s}_i \pm 5\sigma_c \boldsymbol{e}_c), \tag{1}$$

其中， $\boldsymbol{e}_c \in S$ 是通道 c中的单位向量。模型获取 CLIP 空间中的方向 M

$$\Delta \boldsymbol{i}_c = \text{Ave}_i \left(\text{CLIP}_{\text{image}}(x_{i,c}^{+}) - \text{CLIP}_{\text{image}}(x_{i,c}^{-}) \right). \tag{2}$$

构建 $\Delta \boldsymbol{t}$ 随后，StyleCLIP 会获取与编辑对应的 CLIP空间方向 $\Delta \boldsymbol{t}$ 。在 Style- CLIP 中，编辑由目标文本 t 和中性文本 t_0定义。此外，模型会准备 80 个提示词 p （例如，“一张.....的照片”），并将每个提示词与文本拼接，记为 $[\cdot; \cdot]$。模型会获取 CLIP空间中的方向 M

$$\Delta \boldsymbol{t} = \text{Ave}_p \left(\text{CLIP}_{\text{text}}([p; t]) - \text{CLIP}_{\text{text}}([p; t_0]) \right). \tag{3}$$

构建 $\Delta \boldsymbol{s}$ 预期 Δs 只会影响目标文本中提及的属性。因此，StyleCLIP 不采用简单的 $\Delta \boldsymbol{i}_c$ 与 $\Delta \boldsymbol{t}$点积，而是引入一个超参数 β，使 $\Delta \boldsymbol{s}$ 变得稀疏。此外，模型会对 $\Delta \boldsymbol{s}$进行归一化，使其最大值为 1：

$$\Delta s_{0,c} = \begin{cases} \Delta \boldsymbol{i}_c \cdot \Delta \boldsymbol{t} & \text{if } |\Delta \boldsymbol{i}_c \cdot \Delta \boldsymbol{t}| \geq \beta \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

$$\Delta \boldsymbol{s} = \frac{1}{\max_c(s_{0,c})} \Delta \boldsymbol{s}_0. \tag{5}$$

局限性 StyleCLIP 需要将中性文本 t_0 作为额外输入。t_0 要求用户通过反复试错，根据指令中提及的属性来描述原始图像。本质上，哪些属性会被编辑、哪些不会被编辑，可以根据指令 t 和原始图像 x推断出来。为了高效使用编辑系统，理想情况下，模型不应依赖这类冗余输入。

3. 提出的方法

为消除中性文本，我们使用通过 SVM 学到的法向量来处理 $\Delta \boldsymbol{t}$，而非使用从t_0 到 t的编辑方向，如图2所示。首先，我们根据指令的文本嵌入与 $\text{CLIP}_{\text{text}}(t)$ 的余弦相似度，从 I 中检索 100 张正样本图像 X^+ 和负样本图像 X^-。然后，我们训练一个线性 SVM 对图像进行分类，以获得超平面 $\hat{\boldsymbol{w}}^\top \boldsymbol{h}_{\text{image},\ i} + \hat{b} = 0$ 及其法向量 $\boldsymbol{n} = \hat{\boldsymbol{w}}$，其中 $\hat{\boldsymbol{w}}$和 b 是学习得到的参数，而$\boldsymbol{h}_{\text{image},\ i} = \text{CLIP}^{\text{image}}(x_i)$，其中 x_i 是 X^+ 或 X^-中的一张图像。SVM 的公式如下：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{x_i,y_i} \max\left(\left(1 - y_i(\boldsymbol{w}^\top \boldsymbol{h}_{\text{image},i} + b)\right), 0\right), \tag{6}$$

其中，$y_i = \pm 1$ 是 x_i的标签。最后，我们定义 $\Delta \boldsymbol{t} = \boldsymbol{n}$。

[25] 提出了 InterfaceGAN，该方法在风格空间中训练支持向量机，以根据预定义属性进行编辑。我们使用该技术，针对用户的任意指令即时获取编辑方向。

4. 实验

我们在面部领域开展了基于自然语言指令的图像编辑实验。

4.1. 数据集

StyleCLIP-FEU 无需使用图像-描述配对数据集进行训练。我们使用 Multi-Modal-CelebA-HQ 作为验证集和测试集来选择超参数。我们还使用 FlickrFaces-HQ (FFHQ) 作为图像语料库 I。

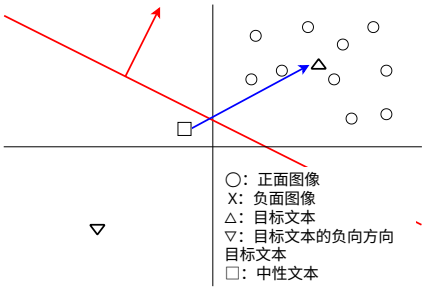


图2. StyleCLIP（蓝色）与所提方法（红色）中 $\Delta \boldsymbol{t}$ 的构建。StyleCLIP 将 $\Delta \boldsymbol{t}$ 定义为从 t_0 到 t的向量。所提方法使用通过支持向量机学习得到的超平面的法向量。

Multi-Modal-CelebA-HQ [31] 中的 CelebA 数据集 [10] 为每张图像自动标注了十条图像说明文字。我们从原始训练集中随机抽取 100 张图像和 10 条图像说明文字，并将其用作验证集。此外，我们从原始测试集中随机抽取 50 张图像和 10 条图像说明文字，并将其用作测试集。

FFHQ [11] 包含 70,000 张人脸图像，并用于预训练 StyleGAN。

4.2. 实现

我们使用 StyleGAN2 [1] 和 CLIP-ViT-B/32 [2] 作为预训练模型。

我们通过采用 0.5 的步长搜索超参数 $\alpha \in [2.0, 6.0]$，并通过采用 0.05 的步长搜索 $\beta \in [0.1, 0.2]$，以最大化编辑后的图像与目标文本 t 在 CLIP空间中的余弦相似度。所有方法均选择 $\alpha = 6.0$ 和 $\beta = 0.1$。其他超参数的选择遵循 [18]中的设定。

4.3. 对比模型

我们使用了 StyleCLIP全局方向 [18]。我们评估了三种中性文本：(a)“面部”，(b)“男性面部”或“女性面部”，以及 (c) 从 Multi-Modal-CelebA-HQ 的十条标注图像说明文字中随机抽取的、对原始图像的详细描述。

4.4. 结果与讨论

对中性文本的敏感性。如图1所示，StyleCLIP生成的图像取决于中性文本。尤其是，如图3所示，当原始图像的性别与指令中指定的性别不同时，使用“男性面部”或“女性面部”作为 t_0 能够提升生成图像的质量。在某些情况下，使用详细图注的 StyleCLIP 成功编辑了图像，但

[1] https://github.com/NVlabs/stylegan2
[2] https://github.com/openai/CLIP



图3. 在原始图像的性别与指令中指定的性别不同时，使用 StyleCLIP-FEU 和 StyleCLIP 编辑图像的示例。StyleCLIP-FEU 明显失败的结果以红色标出。

也有一些情况出现了嘴巴或面部形状不自然的现象。相比之下，StyleCLIP-FEU 执行的编辑效果稳定，且没有 t_0。

对超参数的敏感性。图4展示了将 α 从 2.0 调整到 6.0 的效果示例。与 StyleCLIP-FEU 相比，我们观察到，StyleCLIP 在 α 较大时会导致过度编辑，且难以调整 α 。图5 展示了将 β 从 0.1 调整到 0.2 的效果示例。当 β 值较大时，StyleCLIP 和 StyleCLIP-FEU 均因归一化发散而无法生成图像。然而， β值较小时，则会编辑与指令无关的属性。自适应选择 β 将是我们未来的研究方向。

用于确定分离超平面的正样本图像和负样本图像。图 6 展示了 StyleCLIP-FEU 在 CLIP空间中检索到的正样本图像和负样本图像。在此示例中， X^+ 包含满足“女性”“双下巴”和“微笑”指令的图像。

推理时间。 在 NVIDIA Tesla T4 GPU 上，使用 StyleCLIP 根据一条指令编辑一张图像耗时 0.37 秒。StyleCLIP 与 StyleCLIP-FEU 的差异在于图像检索和 SVM训练，这两项耗时不到 0.01 秒。

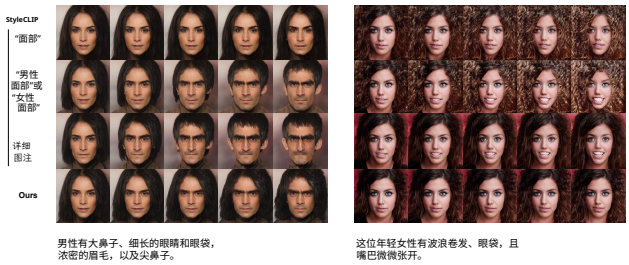


图4. α 从 2.0（左）变为 6.0（右）时编辑后的图像示例。 β 为 0.1。



图5. 当 β 从 0.1（左）变为 0.2（右）时编辑后的图像示例。 α 为 6.0。



图6. 排名前 16 的正样本图像 X^+ 和负样本图像X^-。

主观评估。 我们请两位工作人员根据两项二元标准评估 StyleCLIP 和 StyleCLIP-FEU 生成的图像：性别编辑的准确性，以及图像的自然度，例如面部和头发的形状。若指令 t if t 提及性别，则以此作为真实性别；否则，真实性别与原始图像相同。标注者之间的 Cohen’s κ 为 0.92 和 0.63。

StyleCLIP-FEU 在性别和

表 1. 主观评估结果。

模型	t_0	性别	自然度
StyleCLIP	“面部”	86.6	82.0
	“男性面部”或“女性面部”	87.1	76.9
	详细图注	80.8	66.0
Ours	–	**87.5**	**95.7**

自然度方面优于 StyleCLIP；因此，我们确认 StyleCLIP-FEU 能够遵循 t 并稳健地生成图像。此外，中性文本的调整较为困难，因为详细描述降低了图像的自然度。

5. 相关工作

在 CLIP 发布之前，一些研究已尝试使用自然语言指令进行图像编辑 [5,15,16]。这些研究假设存在图文配对语料库，以便获得多模态理解能力。

StyleCLIP 提出了两种结合 StyleGAN 和 CLIP 的方法。第一种方法利用神经网络进行优化，但计算成本较高。 [9], [34],和 [33] 提高了输出质量。第二种方法是一种快速算法，通过映射编辑方向来实现。 [35] 采用这种方法处理交互式图像编辑任务。然而，这类方法需要额外的输入。 [24]扩展了这类方法；他们还对 StyleGAN 进行了额外训练，使样式空间与 CLIP空间对齐，这一过程在 64 块 NVIDIA A100 GPU 上耗时五周。相比之下，StyleCLIP-FEU 无需额外训练，只需进行轻量级的支持向量机训练。

CLIP 也被用于其他指令式图像编辑任务设置中。 [31]允许使用草图等多模态指令，而 [27] 和 [28] 则假设已有图文配对语料库可用于改进 StyleCLIP。

与我们的研究类似， [1] 引入了支持向量机，以利用法向量进行编辑，不过其任务设置与我们不同：他们将指令限定为预定义属性。他们在 CLIP空间中计算主成分向量，并根据图像与各主成分向量的相似度检索正样本图像和负样本图像。随后，他们学习预定义属性与主成分向量之间的对应关系。与 [25] 和 [1]，相比，我们会根据任意自然语言指令形式的查询，即时检索正样本图像和负样本图像。

最近，扩散模型已被用于合成照片级真实感图像 [21–23]并编辑图像 [13,17,20]。然而，由于它们使用时间序列模型生成图像，计算开销比生成对抗网络更大。 [24]报告称，GLIDE [17] 生成一张 64x64 图像需要 10.9 秒，而 StyleGAN-T [24] 在 A00 GPU 上仅需 0.06 秒。

6. 结论

我们提出了 StyleCLIP-FEU，只需输入一条自然语言指令即可编辑图像。与 StyleCLIP 和 InterfaceGAN 不同，StyleCLIP-FEU 会获取一个超平面，用于分隔即时检索到的正样本图像和负样本图像，并将其法向量用作编辑方向。人脸图像编辑是一项重要任务，应用广泛，例如虚拟形象创建和交流支持。我们让图像编辑方法更易于使用且更高效，以提升其实用性。

参考文献

[1] Rameen Abdal、Peihao Zhu、John Femiani、Niloy J. Mitra 和 Peter Wonka。Clip2stylegan：无监督提取 StyleGAN 编辑方向。CoRR，abs/2112.05219，2021。2，5[2] Yunjey Choi、Minje Choi、Munyoung Kim、Jung-Woo Ha、Sunghun Kim 和 Jaegul Choo。Star GAN：用于多领域图像到图像转换的统一生成对抗网络。载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集，2018年6月。1[3] Edo Collins、Raja Bala、Bob Price 和 Sabine Susstrunk。风格编辑：揭示生成对抗网络的局部语义。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2020年。1[4] Brian Dolhansky 和 Cristian Canton Ferrer。使用示例生成对抗网络进行眼部修复。载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集，2018年。1[5] Hao Dong、Simiao Yu、Chao Wu 和 Yike Guo。通过对抗学习进行语义图像合成。载于 IEEE 国际计算机视觉会议（ICCV）论文集，2017年10月。5[6] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。载于 NIPS，第27卷，2014年。1[7] Erik H¨ark¨onen、Aaron Hertzmann、Jaakko Lehtinen 和 Sylvain Paris。GANSpace：发现可解释的生成对抗网络控制方式。载于《神经信息处理系统进展》，第33卷，第9841–9850页，2020年。1[8] Zhenliang He、Wangmeng Zuo、Meina Kan、Shiguang Shan 和 Xilin Chen。AttGAN：只改变你想改变的内容，实现面部属性编辑。《IEEE图像处理汇刊》，28(11)：5464–5478，2019年。1[9] Xianxu Hou、Linlin Shen、Or Patashnik、Daniel Cohen-Or 和 Hui Huang。FEAT：带注意力的面部编辑。CoRR，abs/2202.02713，2022年。2，5[10] Tero Karras 等。通过渐进式增长提升生成对抗网络的质量、稳定性和多样性。arXiv预印本 arXiv:1710.10196，2017年。3

[11] Tero Karras 等。用于生成对抗网络的基于风格的生成器架构。载于 CVPR，第4401–4410页，2019年。2，3[12] Tero Karras 等。分析并提升 StyleGAN 的图像质量。载于 CVPR，第8110–8119页，2020年。2[13] Gwanghyun Kim、Taesung Kwon 和 Jong Chul Ye。DiffusionCLIP：用于稳健图像操控的文本引导扩散模型。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第2426–2435页，2022年6月。5[14] U mut Kocasari、Alara Dirik、Mert Tiftikci 和 Pinar Yanardag。StyleMC：基于多通道的快速文本引导图像生成与操控。载于 IEEE/CVF冬季计算机视觉应用会议（WACV）论文集，第895–904页，2022年。2[15] Bowen Li、Xiaojuan Qi、Thomas Lukasiewicz 和 Philip H.S. Torr。ManiGAN：文本引导的图像操控。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2020年6月。5[16] Seonghyeon Nam、Yunji Kim 和 Seon Joo Kim。文本自适应生成对抗网络：用自然语言操控图像。载于 S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett 编，《神经信息处理系统进展》，第31卷。Curran Associates, I nc.，2018年。5[17] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。GLIDE：利用文本引导扩散模型实现逼真图像生成与编辑。CoRR，abs/2112.10741，2021年。5[18] Or Patashnik 等。StyleCLIP：由文本驱动的 StyleGAN 图像操控。载于 CVPR，第2085–2094页，2021年。1，2，3[19] Alec Radford 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议（ICML），第8748–8763页，2021年。1，2[20] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。基于 CLIP 潜变量的分层文本条件图像生成，2022年。5[21] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。零样本文本生成图像。CoRR，abs/2102.12092，2021年。5[22]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。利用潜在扩散模型进行高分辨率图像合成。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第10684–10695页，2022年6月。5[23]Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S. Sara Mahdavi、Rapha Gontijo Lopes、Tim Salimans、Jonathan Ho、David J Fleet 和 Mohammad Norouzi。通过深度语言理解实现逼真文本生成图像的扩散模型，2022年。5[24] Axel Sauer、Tero Karras、Samuli Laine、Andreas Geiger 和 Timo Aila。StyleGAN-T：释放生成对抗网络的潜力，实现快速大规模文本生成图像。第 abs/2301 09515卷，2023年。2，5

[25] Yujun Shen 等. Interfacegan：解读生成对抗网络学习到的解耦人脸表征。《IEEE模式分析与机器智能汇刊》，2020年。2, 3, 5[26] Yujun Shen、Jinjin Gu、Xiaoou Tang 和 Bolei Zhou. 解读生成对抗网络的潜在空间以实现语义人脸编辑。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，2020年。1[27] Jianxin Sun、Qiyao Deng、Qi Li、Muyi Sun、Min Ren 和 Zhenan Sun. Anyface：自由风格的文本到人脸合成与操控。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第18687–18696页，2022年6月。2, 5[28] Ming Tao、Bing-Kun Bao、Hao Tang、Fei Wu、Longhui Wei 和 Qi Tian. DE-Net：动态文本引导的图像编辑对抗网络，2022年。5[29] Omer Tov 等. 设计用于StyleGAN图像操控的编码器。《ACM图形学汇刊（TOG）》，40(4):1–14，2021年。2[30] Zongze Wu、Dani Lischinski 和 Eli Shechtman. StyleSpace分析：用于StyleGAN图像生成的解耦控制。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第12863–12872页，2021年。1[31] Weihao Xia 等. TediGAN：文本引导的多样化人脸图像生成与操控。载于CVPR，第2256–2265页，2021年。3, 5[32] Taihong Xiao、Jiapeng Hong 和 Jinwen Ma. ELEGANT：利用生成对抗网络交换潜在编码，以迁移多种人脸属性。载于《欧洲计算机视觉会议（ECCV）论文集》，2018年9月。1[33] Zipeng Xu、Tianwei Lin、Hao Tang、Fu Li、Dongliang He、Nicu Sebe、Radu Timofte、Luc Van Gool 和 Errui Ding. 预测、预防与评估：由预训练视觉语言模型赋能的解耦文本驱动图像操控。载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，第18229–18238页，2022年6月。2, 5[34] Ziqi Zhang、Cheng Deng、Kun Wei 和 Xu Yang. 通过语义感知知识迁移实现文本驱动的图像操控，2022年。2, 5[35] Yufan Zhou、Ruiyi Zhang、Jiuxiang Gu、Chris Tensmeyer、Tong Yu、Changyou Chen、Jinhui Xu 和 Tong Sun. TIGAN：基于文本的交互式图像生成与操控。《AAAI人工智能会议论文集》，36(3):3580–3588，2022年。2, 5