基于方向性CLIP引导的语义调制人脸属性编辑

顾广华[1),2)] 杨远航[1),2)] 伊柏宇[1),2)]

[1)](燕山大学信息科学与工程学院 河北秦皇岛 066004)

[2)](河北省信息传输与信号处理重点实验室 河北秦皇岛 066004)

摘 要 实现高精度且解耦的真实人脸图像编辑仍然是一个具有挑战性的任务。尽管生成对抗网络的隐空间蕴含着丰富且复杂的语义信息,但基于隐空间操作的人脸属性编辑研究常常面临属性纠缠、编辑失效甚至人物身份信息丢失等困境。为了解决这些问题,本文提出了一种基于方向性CLIP引导的语义调制人脸属性编辑模型——SMCI-CLIP。该模型主要包括两大模块:语义调制模块和多通道交互模块。语义调制模块通过调制函数动态调整文本信息在潜码中的权重,并通过正交文本运算去除无关信息,从而生成解耦且精准的编辑向量,实现高精度且解耦的属性编辑。与此同时,多通道交互模块利用交叉注意力机制深入挖掘隐空间潜码各通道间的互信息,以提升模型的全局感知能力。大量实验结果表明,本文所提模型在有效保留人物身份信息的同时,能够实现高精度且解耦的人脸属性编辑,显著优于现有主流方法。

关键词 人脸属性编辑;方向性CLIP;隐空间操作;语义调制;多通道交互

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.00782

Directional CLIP-Guided Semantic Modulation for Facial Attribute Editing

GU Guang-Hua[1),2)] YANG Yuan-Hang[1),2)] YI Bo-Yu[1),2)]

[1)](School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei 066004)

[2)](Hebei Province Key Laboratory of Information Transmission and Signal Processing, Qinhuangdao, Hebei 066004)

Abstract Achieving high-precision and decoupled real facial image editing remains a challenging task. Although the latent space of Generative Adversarial Network (GAN) contains rich and complex semantic information, research on facial attribute editing based on latent space manipulation often encounters dilemmas such as attribute entanglement, editing failure, and even loss of personal identity information. To solve these problems, in this paper, we propose a semantic-modulated facial attribute editing model by integrating the Semantic Modulation and Multi-Channel Interaction (SMCI) based on directional Contrastive Language-Image Pretraining (CLIP) guidance, named SMCI-CLIP. The proposed model consists of two key modules: the Semantic Modulation Module (SMM) and the Multi-Channel Interaction Module (MCI). The Semantic Modulation Module dynamically modulates the weight of text information in the latent code through a modulation function, and removes irrelevant information via orthogonal text operations. Thus, this generates decoupled and accurate editing vectors, enabling high-precision and decoupled attribute editing. Simultaneously, the Multi-Channel Interaction Module employs a cross-attention mechanism to extract mutual information among latent code channels, enhancing global perception. Extensive experimental results demonstrate that the proposed model in this paper can achieve high-precision and decoupled facial attribute editing while effectively preserving

收稿日期:2025-04-01;在线发布日期:2025-11-14。本课题得到国家自然科学基金面上项目(62072394)、河北省自然科学基金面上项目(F2024203049)资助。**顾广华**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为图像理解、人脸属性编辑。E-mail:guguanghua@ysu.edu.cn。**杨远航**,硕士,主要研究领域为人脸属性编辑、图像处理。**伊柏宇**,硕士,主要研究领域为图像处理、计算机视觉。

personal identity information, which is significantly superior to existing mainstream methods.

Keywords　face attribute editing; directional CLIP; latent space manipulation; semantic modulation; multi-channel interaction

1　引　言

目前随着生成对抗网络(Generative Adversarial Network, GAN)[1]在高质量图像生成领域的不断进步,人脸图像编辑技术取得到了显著发展。研究表明[2-4],StyleGAN[5]的潜在空间蕴含丰富的语义信息,其包含的多个语义边界使得通过操纵潜在编码实现精确的语义编辑成为可能。与此同时,预训练语言模型(Contrastive Language-Image Pretraining, CLIP)[6]的出现,极大地推动了基于文本引导的图像编辑方法的研究。这类方法因其交互直观、控制便捷的优势,迅速成为研究热点。现有方法通常利用预训练的CLIP模型作为监督信号,在StyleGAN的潜在空间中搜索与文本描述相匹配的编辑方向,以实现文本驱动的语义编辑。

然而,当前基于CLIP的文本引导图像编辑研究仍面临诸多挑战[7]。首先,目前大多方法[2-4]通过最小化CLIP空间中图像和文本之间的全局相似性来引导模型学习语义向量,这种简化的"引导"方式不仅使模型难以捕捉到文本中的细节信息,且往往需要冗长的训练迭代才能提取有效的语义特征。其次,学习到的编辑向量通常缺乏解耦性,导致图像不相关区域的语义特征发生改变。StyleCLIP[2]和HairCLIP[3]尝试将潜在空间划分为粗、中、细三个层次,并使用独立的映射器生成每个层次的编辑方向。然而,这些方法并非在所有情况下都有效:一方面,HairCLIP局限于发型和发色属性,编辑能力较为单一;另一方面,StyleCLIP虽具备一定的泛化能力,但在表情和年龄等全局属性的编辑表现不佳,缺乏对全局语义的感知能力,加之其依赖单一CLIP全局相似度引导,训练效率较低。DeltaEdit[4]设计了一种基于方向性CLIP引导的映射网络,通过大量的匹配图像训练其模型,实现了解耦且精确的图像编辑。但其对全局属性(如年轻化)的感知能力仍显不足,且对训练数据的多样性和规模要求较高。HyperEditor[8]则提出了一种超网络架构,动态调整StyleGANv2[9]生成器权重,并利用CLIP的跨模态语义对齐能力确保图像与文本之间的语义一致性,以实现真实人脸属性编辑。

近年来,基于扩散模型(Diffusion Models)的图像编辑方法在面部属性编辑领域也引起了广泛关注。例如,Li等人[10]提出的PADA方法,采用扩散模型捕捉年龄编辑过程中的低层次随机变化,有效解决了年龄属性编辑中的不适定问题。Rishubh Parihar等人[11]提出的PreciseControl,借助W^+隐空间的解耦特性,通过训练轻量级Latent Adaptor网络,将潜在编码映射至扩散模型的文本嵌入空间,从而实现更精细的属性编辑控制。尽管此类扩散模型在全局语义一致性和生成质量方面表现优异,但仍面临关键挑战:一方面,多次前向-反向迭代带来较高的计算开销,不易满足实时或交互式编辑的效率需求;另一方面,在发型、发色等局部细节属性的解耦控制能力尚显不足。因此,现有基于扩散模型的方法在实现细粒度属性控制时,往往仍需依赖并融合隐空间的技术。

为此,本文提出了SMCI-CLIP模型,旨在设计一种更高效探索隐空间语义特征的网络结构,以实现更加高精度、可控且解耦的属性编辑。该模型基于StyleGANv2生成器和预训练CLIP模型在文本的引导下实现了解耦合且快速学习属性的图像编辑。与以往基于CLIP引导编辑方法不同,本文设计了一种新颖且简约的语义调制模块(Semantic Modulation Module, SMM),用于在StyleGANv2的潜在空间中生成精确的编辑向量。同时,本文提出一种正交文本分解方法,以精确的文本描述减少不相关区域的修改,从而有效规避属性纠缠问题,确保编辑操作仅影响与文本直接相关的区域。此外,本文还提出了一种多通道交互模块(Muil-channel Interaciton, MCI),用于深层次关联潜在空间与文本空间,通过捕捉潜码多通道交互信息显著增强模型的全局感知力。

2　相关工作

2.1　基于隐空间的图像操作

近年来,生成对抗网络GAN在图像处理领域取得了显著的进展。尤其是ProgressiveGAN[12]、

StyleGAN、StyleGANv2等模型，在生成细致且逼真的人脸图像方面表现尤为突出。StyleGAN模型因其潜在空间蕴含丰富的人脸语义信息，被广泛应用于编辑任务。许多研究表明，通过调整StyleGAN系列模型潜在空间中的潜在编码，可以实现对图像语义属性的有效操控。如文献[13-15]以监督学习方式，利用带有属性标签的样本或预训练分类器来寻找有意义编辑方向；InterfaceGAN[16]利用预训练分类器学习一个支持向量机（Support Vector Machines，SVM）的分类边界，将潜在空间分离为相反的语义标签，并将支持向量边界的法向量作为语义方向输出；Latent-Transformer[15]针对特定属性训练变换网络，并在属性标签的监督下进行优化；Adatrans[14]则通过设计一组自适应学习模型，以标签互信息的学习策略来逐步预测语义方向；此外，StyleCLIP通过将不同细粒度的人脸风格潜在编码进行三层划分，并利用CLIP预训练模型引导预测网络进行语义学习；HairCLIP则通过CLIP模型的文本嵌入将发型、发色文本信息分别注入不同预测网络中，以预测发色及发型语义向量；DeltaEdit提出了一个“增量-文本空间”，基于该空间DeltaEdit模型将CLIP文本特征差异映射至StyleGAN的潜在空间中，以预测编辑方向。

2.2 基于文本引导的图像操作

基于文本的图像操作方法核心思想是利用自然语言文本描述引导模型学习与文本描述一致的语义向量，从而实现图像编辑[17-19]。目前，许多研究工作采用条件生成对抗网络（Conditional GAN，cGAN）[20]结构，将输入的文本表示为GAN的条件信息，并通过引入多种正则化项，以实现可控的图像生成。然而，这些方法通常采用端到端的方式训练整个框架，且仅依赖判别器来提升图像质量，这使得生成的图像可能无法与给定文本完全匹配，同时图像生成的多样性也会受到限制。近年来，基于StyleGAN系列模型的研究工作逐渐得到广泛关注，并取得显著进展。例如，TediGAN[21]提出了视觉-语言相似性模块，在StyleGAN的潜在空间中对齐视觉和文本模态信息。CI-GAN[22]则借鉴TediGAN的思想，提出在GAN反演过程中使用循环一致性（Cycle Consistency）来提升图像编辑效果。其中最具代表性的工作StyleCLIP开创性地将StyleGANv2的生成能力与CLIP的图像-文本表示能力相结合，用于探索图像的编辑方向。StyleCLIP提出了三种方法：潜在优化（Latent Optimization，StyleCLIP-OP）、潜在映射器（Latent Mapper，StyleCLIP-LM）和全局方向（Global Directions，StyleCLIP-GD）。StyleCLIP-OP是通过计算生成图像与文本描述在CLIP空间中的相似度损失，并对潜码进行迭代优化，实现逐步逼近目标语义的图像编辑；StyleCLIP-LM则构建了一个潜在映射网络，能够在保持编辑语义一致性的前提下，快速预测目标潜码，从而实现高效图像编辑；StyleCLIP-GD将图像映射至StyleSpace[23]的解耦空间中，并通过计算CLIP空间中源文本与目标文本间的距离来预测空间中的编辑向量，从而实现解耦的属性编辑。HyperEditor通过引入一个条件超网络，根据输入的文本语义生成权重调整因子，直接对预训练的StyleGANv2生成器的网络权重进行重新分配，从而实现图像编辑。表1展示了本文方法与现有主流方法在关键指标上的对比结果。

表1 本文方法与主流方法在关键性能指标上的对比

方法 指标	图像视觉自然性	身份保持能力	人脸属性解耦机制	模块可解释性
StyleCLIP	较为自然	○部分	×无	×无
TediGAN	○较一般	○部分	×无	×无
DeltaEdit	✓较为自然	✓较好	✓方向性引导	○部分
本文方法	✓较高	✓较好	✓正交性引导	✓支持

3 提出方法

3.1 模型框架

本文提出的模型结构如下图1所示。首先，通过e4e（Encoder for Editing）反演编码器[24]将输入图像投影至StyleGANv2的W^+空间中，以获取初始潜码w_{init}。同时，利用CLIP的文本编码器将目标文本映射至文本特征空间，并通过正交分解运算得到向量Δe_{rele}。随后，将Δe_{rele}和w_{init}输入至语义调制模块中进行语义调制，生成初步的语义向量Δw_1，接着，再将Δw_1和Δe_{rele}输入至多通道交互模块，用于生成通道语义向量Δw_2。然后，将Δw_2作为输入，重复两次调制-交互操作，最终输出预测编辑向量$\Delta w_2''$。最后，通过StyleGANv2的生成器G生成编辑后的图像$G(w_{init}+\lambda_s*\Delta w_2'')$，其中$\lambda_s$为属性强度控制参数，用于调节编辑效果的强弱。

3.2 e4e反演编码器

在基于StyleGANv2隐空间的人脸属性编辑

流程中，准确地将真实图像映射至生成器潜在空间是实现高保真重建与可编辑性的关键。e4e反演编码器正是针对这一需求而设计的预训练模型，其核心目标是将输入的真实人脸图像高精度地投影到StyleGANv2的W^+空间中，同时保留足够的可编辑信息。该编码器通过限制不同层风格向量之间的方差，使生成的潜在编码更接近StyleGANv2映射网络的分布。其次，引入潜在空间判别器，通过对抗训练机制进一步约束编码输出，使其分布更加贴近真实的W^+样本分布。这样的设计能够在很小的重建精度损失下，实现对输入图像的高质量重建并保留其语义细节，从而在失真程度与编辑可控性之间取得平衡。鉴于到e4e编码器生成的潜码兼顾了较高的重建精度和良好的可编辑性，本文采用该模块作为SMCI-CLIP默认的反演编码器模块。

3.3　正交方向性CLIP引导

针对当前成对图像数据集中特定属性变化样本相对稀缺所导致的监督信息不足问题，本文利用CLIP模型的强大跨模态语义对齐能力，以自监督学习的方式来实现原始图像向目标图像的转化。现有研究（如HairCLIP、StyleCLIP等）主要采用全局语义对齐策略，通过最小化生成图像与目标文本在CLIP联合嵌入空间的余弦距离实现属性控制，其目标函数如公式(1)所示。其中，E_i表示CLIP的图像编码器，E_t表示CLIP的文本编码器，$\cos(\cdot,\cdot)$表示余弦相似度，x_{tar}是生成的图像，$text$表示为文本信息。

$$L_{CLIP}^{globe}=1-\cos(E_i(x_{tar}),E_t(text)) \tag{1}$$

然而，这种全局对齐范式容易引发图像全局特征空间的不必要扰动，特别是在处理局部属性编辑任务时，会导致优化方向模糊化与模型收敛效率低等次优问题。为克服这一局限性，本文提出了一种正交方向性CLIP机制。具体而言，通过构建源域-目标域文本对的残差特征向量$\Delta e_t\in\mathbb{R}^{512}$，并采用正交投影剔除无关信息，以建立具有目标方向约束的语义编辑空间，具体过程如式(2)～(6)所示。

$$e_y=E_t(t_{tar}),e_x=E_t(t_{ori}) \tag{2}$$

$$\Delta e_t=e_y-e_x \tag{3}$$

其中，t_{tar}和t_{ori}分别表示目标域和源域的语义描述（如t_{tar}=“face with laugh”，t_{ori}=“face”）。由于CLIP特征空间存在语义纠缠，残差Δe_t作为引导向量仍包含正交噪声分量。为此，本文引入了PCA算法[25]对源域特征进行本征分解，通过人工合成文本集$\{T_x^{(k)}\}_{k=1}^N$计算e_x的主成分空间，提取主导基向量$v_{face}=\arg\max_{v\in\mathbb{R}^{512}}\mathbb{E}\left[\left(e_x^Tv\right)^2\right]$，其表征面部属性主方向。将原始残差向量$\Delta e_t$进行正交分解：

$$\Delta e_t=\Delta e_{rele}+\Delta e_{irrele} \tag{4}$$

并计算Δe_t在“face”方向的投影，在图1中用$proj_{v_{face}}$符号表示投影操作，过程如式(5)所示。

$$\Delta e_{irrele}=\frac{<\Delta e_t,v_{face}>}{\|v_{face}\|^2}v_{face} \tag{5}$$

然后，减去无关向量，得到“纯净”文本向量：

$$\Delta e_{rele}=\Delta e_t-\Delta e_{irrele} \tag{6}$$

该正交运算在数学本质上实现了对文本特征空间的局部正交分解，有效剥离与目标属性无关的语义成分。当且仅当Δe_{rele}与目标属性变化方向保持最大相关性时，该引导向量可约束模型在潜在空间的优化轨迹，从而确保非目标区域的特征保持性与编辑区域的特征特异性之间的动态平衡。

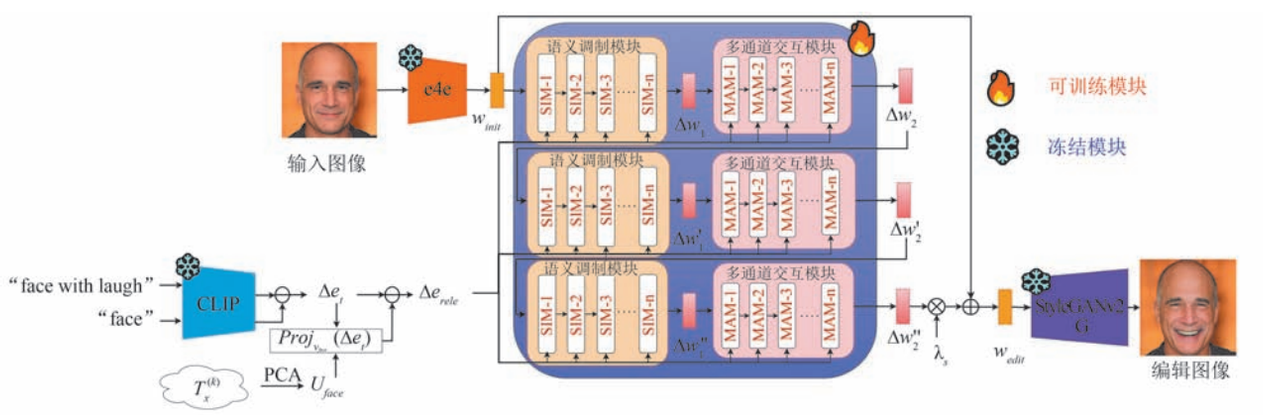

图1　SMCI-CLIP结构图

3.4 语义调制模块

受到StyleCLIP的Latent-mapper网络启发，本文提出了一种语义调制模块，用于动态调制文本信息在潜码中的权重，该模块将语义信息嵌入至语义调制模块中的每个子模块中，而子模块会将文本语义信息纳入潜码空间中以指导图像编辑。

调制模块由n个语义注入子模块(Semantic Inject Modulation，SIM)堆叠而成。每个子模块的结构设计如图2所示，其包含一个简单全连接模块(FC)、一个调制模块以及一个泄露$Relu$激活层。子模块的核心操作是语义特征$\boldsymbol{\Delta}e_{rele}$通过两个函数$f_\alpha(\boldsymbol{\Delta}e_{rele})$和$f_\beta(\boldsymbol{\Delta}e_{rele})$分别生成调制系数$\alpha_t$和偏置系数$\beta_t$，用于调整输入潜码。之后，通过对潜码进行线性变换和归一化操作，输出语义向量$\boldsymbol{\Delta}w_1$，$\boldsymbol{\Delta}w_1\in\{\boldsymbol{\Delta}w_1', \boldsymbol{\Delta}w_1'', \boldsymbol{\Delta}w_1'''\}$。过程表示如公式(7)、(8)、(9)所示。

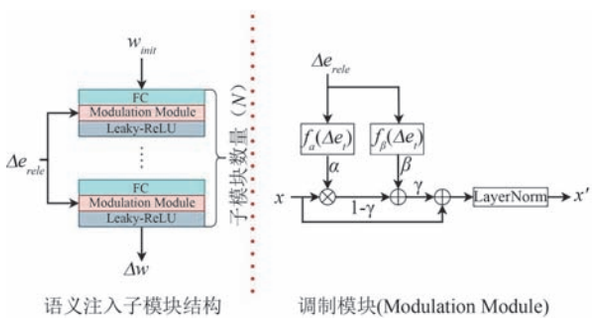

图2 语义调制模块结构

$$\alpha_t=f_\alpha(\boldsymbol{\Delta}e_{rele}),\beta_t=f_\beta(\boldsymbol{\Delta}e_{rele}) \tag{7}$$

其中，f_α和f_β是两个不同的可学习函数，这两个函数均采用多层全连接层结构实现。之后，使用调制系数α_t和β_t对输入潜码w进加权处理：

$$w_{mod}=\alpha_t\cdot w+\beta_t \tag{8}$$

调制模块通过参数γ来调节原始潜码w与文本分支调制后的潜码w_{mod}的融合比例：

$$w'=\gamma\cdot w_{mod}+(1-\gamma)\cdot w \tag{9}$$

其中，γ是一个可学习的标量，用于控制文本分支与原始潜码的加权组合。最后再经过层归一化操作后输出语义潜码向量$\boldsymbol{\Delta}w_1$，最终输出公式为

$$\boldsymbol{\Delta}w=LN\left[\gamma\cdot(\alpha_t\cdot w+\beta_t)+(1-\gamma)\cdot w\right] \tag{10}$$

其中，LN为层归一化操作(LayerNorm)。该模块通过参数控制文本分支的权重，使得文本特征的影响得以灵活调节。

3.5 多通道交互模块

语义调制模块采用调制系数进行加权运算操作，增强了模型对于局部属性的感知力，如口红、发色、发型等，这些属性通常局限于面部图像的局部区域并表现为显著的像素级变化。然而，经过大量实验发现，该模块在处理如年龄、生气、惊吓等全局属性表现欠佳。这类属性通常涉及整张人脸的多区域协同变化，语义调制模块缺乏对其的充分建模能力。

基于此，本文提出了多通道交互模块，用于将潜码与文本中的全局信息进行多层次交互，以增强模型的全局感知。该模块由m个结构相同的多重注意力子模块(Multi Attention Module，MAM)级联而成。该子模块借鉴了Transformer编码器结构[26]，包含一个多注意力层和两层前馈网络。形成“上下双层前馈网络＋多头注意力层”夹心设计，结构图3所示。具体而言，将语义调制模块的输出Δw_1输入至由全连接层(FC)以及$Relu$激活函数构成的上层前馈网络，得到隐层表示，并将该表示作为下层注意力层的键($\boldsymbol{K}$)和值($\boldsymbol{V}$)，而$\boldsymbol{\Delta}e_{rele}\in\mathbb{R}^{k\times512}$作为注意力层的查询($\boldsymbol{Q}$)，计算过程如公式(11)、(12)、(13)所示：

$$\text{Muil}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(head_1,\cdots,head_h)\boldsymbol{W}^o \tag{11}$$

$$head_i=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{12}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)\text{V} \tag{13}$$

其中，$\boldsymbol{W}^o,\boldsymbol{W}_i^Q,\boldsymbol{W}_i^K,\boldsymbol{W}_i^V$是可学习的权重矩阵，$h$是注意力头的数量，$d_k$是潜在空间的维度。之后通过下层前馈网络进行非线性变换，输出预测潜码。经过多个MAM子模块级联后，残差文本信息与潜码在多通道上完成充分的特征交互，输出的潜码$\boldsymbol{\Delta}w_2\in\mathbb{R}^{k\times512}$，$\boldsymbol{\Delta}w_2\in\{\boldsymbol{\Delta}w_2', \boldsymbol{\Delta}w_2'', \boldsymbol{\Delta}w_2'''\}$具有足够的全局语义表达能力。

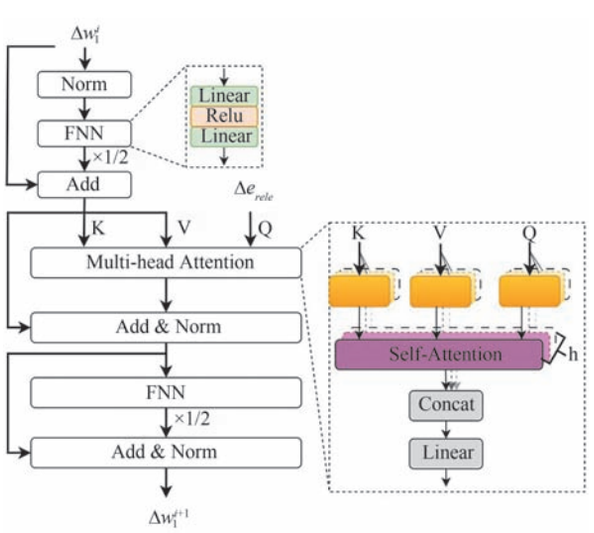

图3 多重注意力子模块结构

3.6　损失函数

文本引导的图像编辑旨在使编辑后的图像与输入文本相匹配，同时保持与文本无关的区域不变。为此，本研究引入三类损失函数，基础重损失、身份保持损失和语义一致损失。

（1）基础重损失函数：

为了提升图像整体重建质量，本文引入像素损失[27]和图像特征损失[28]。像素损失通过逐像素比较生成图像$G(w')$与反演图像$G(w)$之间的差异，实现对图像细节的细粒度监督。特征损失则是在特征层面比较图像。具体公式定义如(14)、(15)所示：

$$L_{pixel}=\left\| G(w)-G(w') \right\|_2^2 \tag{14}$$

$$L_{lpips}=\left\| F_{VGG}\left(G(w)\right)-F_{VGG}\left(G(w')\right) \right\|_2^2 \tag{15}$$

其中，$G(w)$，$G(w')$分别表示原始反演图像及编辑图像，F_{VGG}为用于提取图像特征的预训练VGG网络[29]。

为了在编辑过程中有效保留背景内容，本文引用了背景重构损失函数[30]。表达式如公式(16)所示。

$$L_{bg}=\left\| (G(w')-G(w))*(P(G(w'))\cap P(G(w))) \right\|_2 \tag{16}$$

其中，$P(\cdot)$是面部解析网络[31]，$P(G(w'))$，$P(G(w))$分别表示输入图像和输出图像中的非面部区域。

（2）身份保持损失函数：

为保留图像人物信息，本文在训练过程中引入了身份损失函数，具体公式如下：

$$L_{ID}=1-\cos\left(R(G(w)),R(G(w'))\right) \tag{17}$$

其中，G表示为生成器，$G(w)$，$G(w')$分别表示原始反演图像及编辑图像，R表示ArcFace[32]网络，该网络用于提取人脸身份特征。

（3）语义一致损失函数：

为实现精确属性编辑本文引入方向性CLIP损失函数[27]对模型进行监督。计算公式如下：

$$L_{CLIP}^{differ}=1-\cos\left(E_i(x_{tar})-E_i(x_{ori}),\boldsymbol{\Delta e}_{rele}\right) \tag{18}$$

其中，x_{ori}为原始图像，x_{tar}为编辑图像，E_i为CLIP图像编码器，E_t为CLIP文本编码器，$\boldsymbol{\Delta e}_{rele}$为残差文本特征。

为避免编辑过度问题。本文引入范数损失[33]，以限制编辑向量幅度，具体如下：

$$L_{l2}=\left\| \boldsymbol{\Delta} w \right\|_2 \tag{19}$$

综上，整体损失函数可以表示为如下公式：

$$\mathcal{L}_{total}=\lambda_{clip}\cdot L_{CLIP}^{differ}+\lambda_{id}\cdot L_{\mathrm{ID}}+\lambda_{w2}\cdot L_{l2}+\lambda_{pixel}\cdot L_{pixel}+\lambda_{\mathrm{lpips}}\cdot L_{\mathrm{lpips}}+\lambda_{bg}\cdot L_{bg} \tag{20}$$

4　实验结果与分析

4.1　实验设置

本实验在FFHQ数据集[5]和Celeba-HQ验证数据集[12]上进行了训练和验证，并收集了40对通用面部描述词，涵盖发色、胡须、年龄、眼睛等多个类别属性。实验采用了预训练的e4e模型作为反演编码器，并利用StyleGANv2生成器生成图像。为保持生成图像质量整体稳定，反演编码器与生成器的全部参数在训练过程中保持冻结，仅训练本文设计的多通道交互模块与语义调制模块。在训练配置方面，SIM子模块数量设置为5，多通道交互模块中的注意力头数量设置为8，MAM模块数量设置为2，采用Adam优化器对可训练模块进行优化，并设置迭代次数为500，初始学习率设置为0.001，批次大小设置为4，超参数设置为$\lambda_{clip}=1$，$\lambda_{id}=0.8$，$\lambda_{w2}=0.2$，$\lambda_{pixel}=0.5$，$\lambda_{\mathrm{lpips}}=0.1$和$\lambda_{bg}=0.5$。本研究模型是在Pytorch框架下使用单张NVIDIA GeForce RTX 2080 Ti进行模型训练。

4.2　属性编辑实验

为了验证所提方法的有效性，本实验展示了对13种不同人脸属性的编辑结果，结果如图4、5所示。在图4中，本文方法能够精确地调整面部表情、发色、年龄、眼睛颜色、眼镜等目标属性，同时有效保留人物身份、背景和光照等无关特征。图5进一步验证了本文方法在多属性编辑任务中的适用性。本文方法在编辑口红属性（+口红）后再编辑肥胖属性（+肥胖）、编辑微笑属性（+微笑）后再编辑卷发属性（+卷发）以及编辑妆容属性（+妆容）后再编辑眼镜属性（+眼镜），生成图像仍能保持主观上的自然性与身份一致性，没有出现图像失真或属性纠缠等问题。这些实验结果表明本文方法不仅在单一属性的编辑上表现优异，也能够有效应对更为复杂的多属性编任务。

4.3　对比实验

本实验展示了本文方法与当前主流文本引导图像编辑方法（TediGAN、StyleCLIP-GD、DeltaEdit和HairCLIP）之间的面部图像操控对比结果，实验结果如图6所示。从图中可以看出，TediGAN的图像编辑效果相对较差，无法成功完成对妆容、年龄及肥胖等属性的编辑。同时，编辑后的人脸周围出现

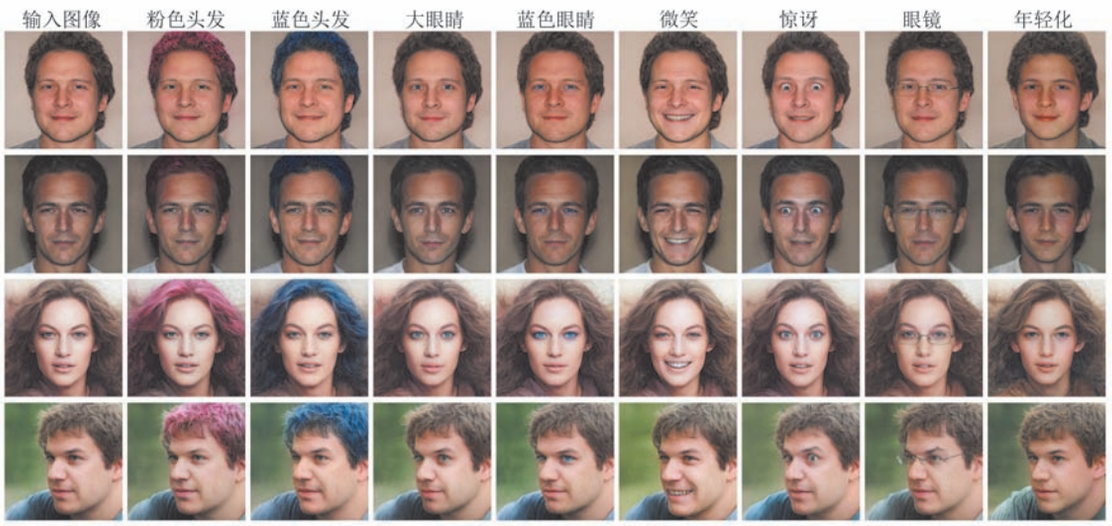

图4 单属性编辑结果图

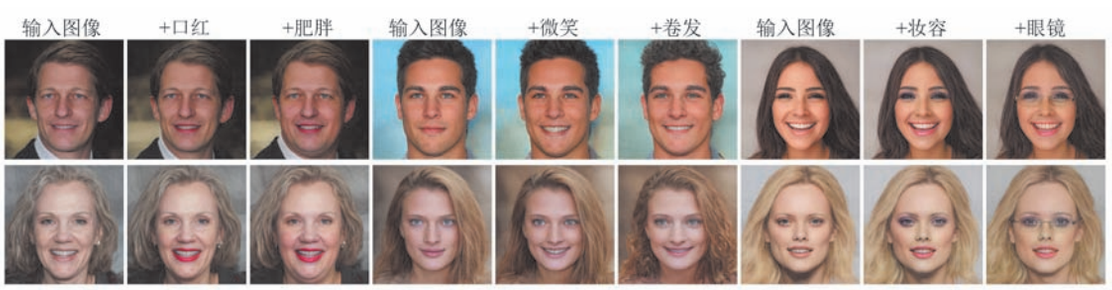

图5 多属性编辑结果图

较多“伪影”，影响了图像质量。StyleCLIP-GD虽然能够完成发型、肥胖等目标区域的编辑，但在编辑卷发属性过程中，人物身份信息丢失较为严重，并且无法实现蓝色发色属性的编辑。HairCLIP对于发色和发型属性编辑的有着较高编辑精确度，但同时也改变了其余无关属性。例如，在编辑灰色发色或红色发色时，人物的衣物颜色也随发色的变化而发生了改变。此外，DeltaEdit在发色编辑方面的效果与StyleCLIP-GD类似，无法很好地完成对发色的准确编辑，并且在编辑肥胖属性时，人物姿态发生了改变，影响了编辑的稳定性。

与上述方法相比，本文方法在属性编辑的精度和解耦性上表现更优。例如，在编辑灰色发色属性时，本文方法能够完整地保留与目标属性无关的区域，同时准确地实现了特定属性修改。此外，在年龄编辑任务中，本文方法成功完成了属性变换。而StyleCLIP-GD、DeltaEdit及TediGAN均未成功完成年龄属性的编辑。对于头发区域，本文方法呈现了一种非常自然的编辑效果，同时也较好地保留了人物发型信息。

4.4 评价指标

为评估本文方法在属性编辑任务中的性能与优越性，本实验采用了峰值信噪比(PSNR)[34]、结构相似性(SSIM)[35]、人物身份相似度(ID)以及CLIP Sorce得分(CS)指标[36]，来全面衡量处理后图像的质量以及图像编辑前后文本描述之间的一致性。

(1) 结构相似性(SSIM)：该指标用于衡量两幅图像在亮度、对比度和结构三个方面的相似度，其值域为[0, 1]，其数值越高，表示两张图像的结构越接近。

(2) 峰值信噪比(PSNR)：该指标通过比较图像的最大像素值与图像误差信号的均方误差之间的比值，来衡量图像的失真程度。PSNR值越大，代表图像失真越小，即生成图像与原图越接近。

(3) 人物身份相似度(ID)：该指标用于衡量编辑图像与原图在人物身份上的一致性，通常通过预训练的人脸识别模型(如ArcFace)提取特征，并计算两张图像之间的余弦相似度，分数越高表示两者身份特征越一致。

(4) CLIP Score(CS)：该指标基于CLIP模型，

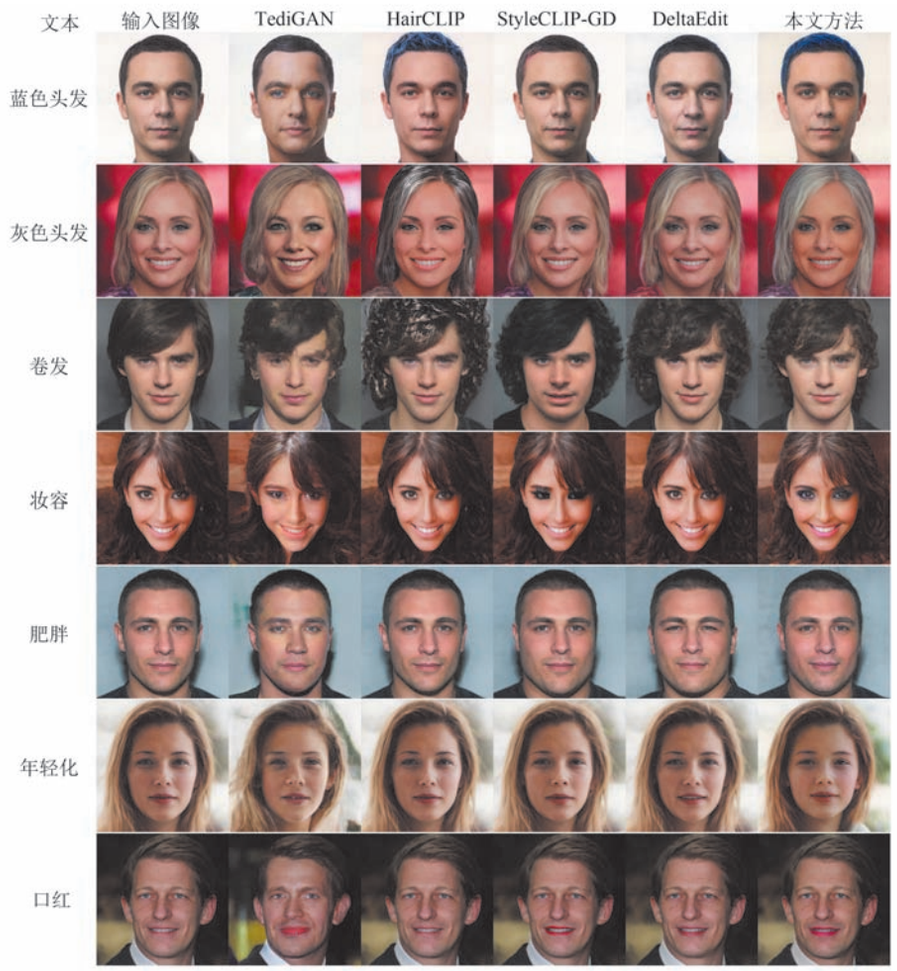

图6　对比实验结果图

通过计算生成图像与目标文本描述之间的相似度得分，用于评估图像内容与目标语义的一致性。得分越高，说明编辑后的图像与所期望的属性描述更契合。

4.5　定量评估实验

本实验选取SSIM、PSNR、ID以及CS作为评价指标，对于发型发色属性，选取了黑色发色、棕色发色、蓝色发色、灰色发色、卷发、粉色发色、紫色发色、红色发色八种发色属性进行量化评估。而对于局部及全局属性，则选取了大笑、肥胖、惊讶、大眼、口红、年轻化、妆容七种属性。实验选取Celeba-HQ数据集的测试集作为评估数据。评估结果如表2所示，本文方法在多个评估指标上取得了显著优势。具体而言，在发色属性编辑的CS指标上，HairCLIP表现略优，这反映了其在发色调控上的敏感性。然而，从整体视觉一致性与保真度角度来看，本文方法在发质、光泽保持及编辑自然度方面展现出更平衡的表现。与此同时，本文方法在SSIM、PSNR、ID得分方面，均优于包括HairCLIP在内的其它对比模型。

表2　定量分析结果表

方法	发型编辑				局部及全局属性编辑			
	SSIM ↑	PSNR ↑	ID ↑	CS ↑	SSIM ↑	PSNR ↑	ID ↑	CS ↑
StyleCLIP-GD α=5	0.804	21.16	0.932	23.48	0.880	25.25	0.815	22.74
TediGAN	0.660	20.47	0.618	25.65	0.661	20.51	0.620	24.20
DeltaEdit	0.785	20.28	0.916	23.79	0.839	23.96	0.853	22.19
HairCLIP	0.791	18.47	0.930	**29.73**	-	-	-	-
本文方法	**0.833**	**22.94**	**0.938**	25.73	**0.889**	**26.19**	**0.879**	**25.09**

4.6 属性强度控制实验

为了进一步验证本文方法的灵活性，本文设计了属性强度控制实验。通过引入属性强度控制参数$\lambda_s \in (0,2)$对属性方向向量的强度进行缩放和调节。具体公式为$w_{init} + \lambda_s * \Delta w_2''$，其中，$\lambda_s$参数控制属性变化的强弱，$\Delta w_2''$表示为模型预测属性向量。在具体实验中，本文选取了肥胖、妆容、年轻化、微笑、大笑等较为典型属性进行评估，实验结果如图7所示。在不同强度参数控制下，本文方法能够实现平滑且自然的属性编辑过渡。例如，在微笑属性编辑过程中，随着参数λ_s的增加，人物微笑特征逐渐加深但人物表情仍保持自然。此外，在大笑、年轻化、妆容、肥胖属性编辑过程中，模型同样保持良好的强度控制能力。

图7 属性强度控制实验结果图

4.7 时空消耗实验

本文方法与当前主流的图像编辑方法，包括StyleCLIP-LM、HaieCLIP和DeltaEdit进行了训练时长收敛对比实验，实验结果如图8所示。本文方法在处理每个文本描述对时仅需500次迭代即可收敛(图8(d))，而StyleCLIP-LM的训练中至少需要8,000次迭代才能达到稳定状态(图8(c))，HairCLIP需要150,000次迭代(图8(b))，DeltaEdit则需要250,000次迭代才能收敛(图8(a))。

此外，从评估结果表3中也可以看出，本文方法在推理和训练时间消耗上表现出明显优势。TediGAN虽然能以对话文本的形式编辑图像，但推理所需时间在表中最长。HairCLIP是用多个文本输入来训练单个模型，这样虽然能简便推理过程，但需要花费大约1天的时间来训练模型。并且，HairCLIP只能编辑头发区域，如果添加新文本则需要重新训练。StyleCLIP-LM至少需要6小时来训练单个文本的映射器。DeltaEdit在单张图像推理时间上表现最为优异，但在模型结构上与HairCLIP类似，新增文本也需重新训练。相比之下，本文方法仅需10分钟即可完成单个文本的学习，推理时间也仅需0.16 s，综合推理时间和训练时长来看，本文方法优于其它方法。

4.8 消融实验

为了验证本方法模型设计的合理性，本文设计消融实验，分析语义调制模块(SMM)和多通道交互模块(MCI)对模型性能的影响。实验分为四组：(1) 不使用SMM和MCI的对照实验(类似于StyleCLIP-OP)，记为“w/o SMM& w/o MCI”；(2) 不使用SMM但引入MCI模块，评估通道交互的贡献，记为“w/o SMM”；(3) 使用SMM但不使用MCI，验证语义调制模块的作用，记为“w/o MCI”；(4) 同时使用SMM和MCI，记为本文方法。

4.8.1 定性分析

实验结果如图9所示，在移除了SMM模块的第(1)组和第(3)组实验中，模型均未能准确完成对

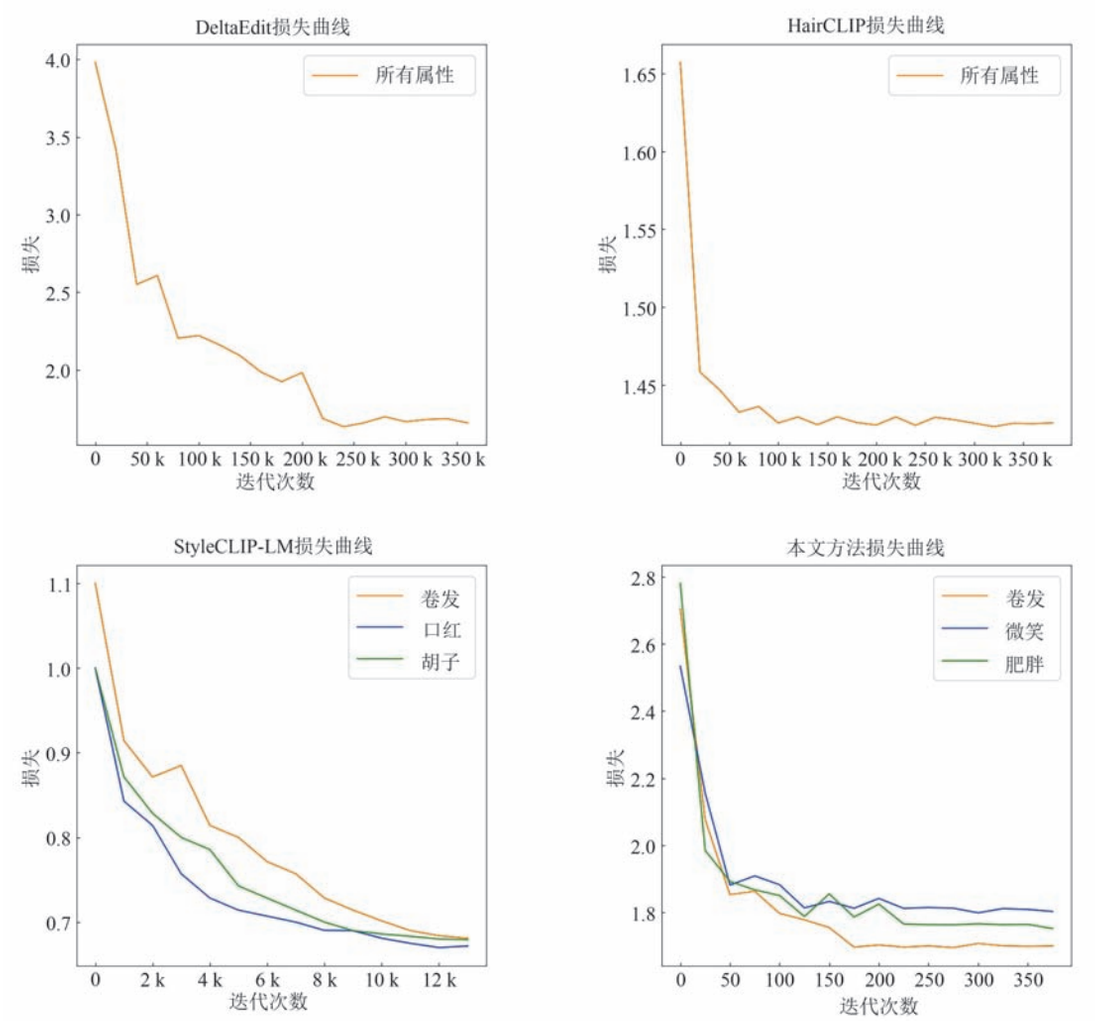

图8　训练收敛对比实验结果图

表3　时空消耗对比表

方法	训练耗时	推理耗时	文本类型
StyleCLIP-LM	8 h+	0.32 s	单文本
TediGAN	15 h+	22 s	长句文本
HairCLIP	1 days+	0.23 s	多文本
DeltaEdit	2 h+	**0.14 s**	多文本
本文方法	**<15 min**	0.16 s	单文本

发色的编辑。比如，在编辑黄色发色和绿色发色属性任务中，第(1)组实验产生的发色结果偏白，存在显著偏差；而(3)组实验虽然在一定程度上实现了发色编辑，但同时引发了人物的衣物颜色的变化，产生了属性纠缠现象。相比之下，在保留SMM模块的(2)组和(4)组实验中，模型能够实现对发色属性的准确调控，生成结果更加自然且语义解耦。这表明SMM模块能够有效调制文本语义信息，并增强语义方向向量的选择性与准确性，从而在潜空间中找到更具分离性和表达力的方向，实现精细且可控的局部属性编辑。

此外，若移除MCI模块，模型在全局属性编辑上表现明显不足。比如，在编辑年轻化和伤心属性时，(1)组和(2)组实验结果未展现出显著变化，而具有MCI模块的(3)组和(4)组实验则能够精准捕捉目标属性并成功完成编辑。

综上所述，SMM模块有助于提升语义调制的精度与解耦能力，适用于细粒度的局部属性编辑；而MCI模块则通过跨通道信息整合增强了模型的全局感知能力。

4.8.2　定量分析

为验证所设计模型的合理性及有效性，本文进行了模块消融定量评估实验。评估数据为CelebA-HQ的2000张高清人脸测试集，选取了八个典型属性(蓝色发色、绿色发色、粉色发色、紫色发色、黄色发色、年轻化、伤心、惊吓)作为测试属性。实验采用峰值信噪比(PSNR)、结构相似性指数(SSIM)、人

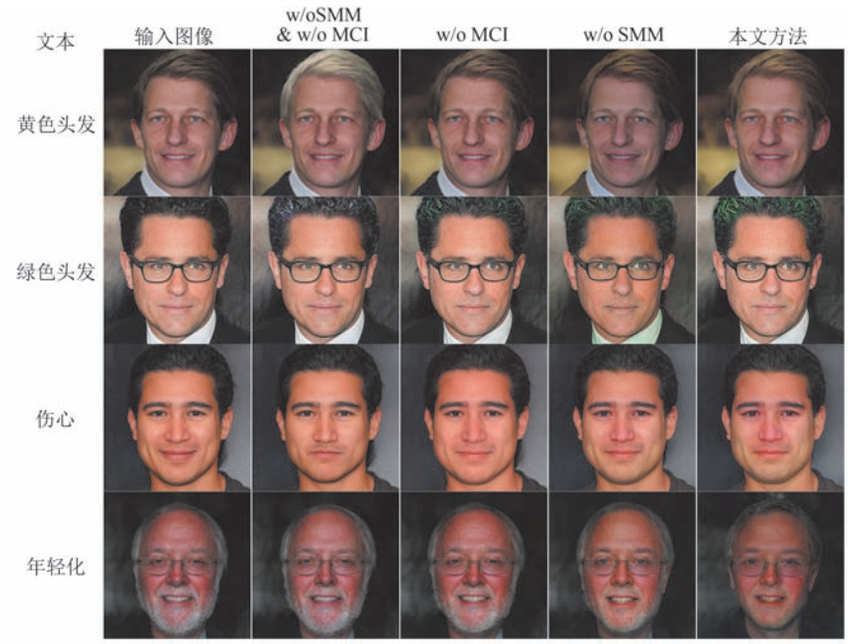

图9 定性分析实验结果图

物身份信息保留度(ID)以及CLIP Score(CS)四项指标,对各模块在图像编辑任务中的表现进行全面分析,结果如表4所示。

表4 单模块定量评估表

方法	SSIM↑	PSNR↑	ID↑	CS↑
w/o SMM & w/o MCI	**0.928**	**28.99**	**0.966**	22.22
w/o MCI	0.881	26.07	0.964	24.37
w/o SMM	0.825	22.81	0.928	25.34
本文方法	0.843	24.20	0.933	**25.97**

(1) 移除SMM和MCI模块的实验组(组1,记w/o SMM & w/o MCI)。在移除了语义调制模块(SMM)和多通道交互模块(MCI)后,该实验组的CS得分最低,SSIM、PSNR和ID得分最高。从指标定义上分析看,这表明在缺少这两个关键模块时,模型无法有效完成属性编辑,图像的整体结构及细节并未发生变化,虽然在SSIM、PSNR和ID分数上表现良好,但语义编辑效果不佳。

(2) 仅移除MCI模块的实验组(组2,记w/o MCI)。移除MCI模块后,该组实验的CS得分高于组1,而SSIM、PSNR和ID得分出现了明显下降。结合其定性分析表现可知,该组实验能够相较于组1能较好地完成部分属性编辑(如发色),但对表情、年龄等全局属性的编辑能力不足,因此在整体语义编辑上表现较弱。

(3) 仅移除SMM模块的实验组(组3,记w/o SMM)。移除SMM模块后,实验组的CS得分高于组2,同时SSIM、PSNR和ID得分低于组2。结合其定性分析表现可知,在加入了多通道交互模块后,模型在解耦度上的表现有所下降,尤其在发色编辑任务上不够准确,但全局属性(如表情和年龄)的语义编辑效果有所提升,因此CS得分较高,但SSIM、PSNR和ID整体得分偏低。

(4) 完整模型实验组(组4,记本文方法)。完整模型实验组的CS得分最高,且SSIM、PSNR和ID得分高于组3。结合定性分析表现表现可知,在SMM和MCI模块的共同作用下,模型能够在实现高精度编辑的同时确保解耦性,成功完成发色、表情以及年龄等全局属性的编辑,实现了准确的可控编辑以及图像失真之间的折中平衡。该结果进一步证明了所设计模块在提升编辑图像质量与语义一致性方面的有效性。

通过以上分析可知,SMM模块在减少属性纠缠方面起到了关键作用,而MCI模块则显著改善了模型对全局属性编辑的效果。完整模型实验结果验证了模块设计的合理性与有效性,表明SMCI-CLIP能够显著提升编辑图像的质量和语义一致性。

4.9 解耦性分析

为进一步验证正交解耦机制在属性编辑中的有效性及优越性,本文设计了语义解耦过程的可视化实验,展示本文模型在不同引导方法下的编辑结果。图10展示了三类典型属性(肥胖、胡须、卷发)的编辑效果对比。每组示例依次包括原始输入图

像、CLIP引导、方向性CLIP引导以及正交方向性CLIP引导下的编辑结果。

以肥胖属性为例，使用CLIP引导的编辑结果目标属性几乎未发生明显变化；方向性CLIP引导虽成功修改了目标属性，但人物姿态属性也发生显著改变，存在明显的属性纠缠现象。相较之下，模型采用正交方向性CLIP引导机制剔除了无关语义干扰，编辑结果在保持身份一致性的同时，呈现出了自然、聚焦的属性变化，体现了优越的解耦能力。

在胡须和卷发属性上也观察到类似现象。未解耦的向量往往导致图像其他区域结构发生偏移，如人物面部表情（微笑）变化或人物胡须密度异常；而经正交处理后的引导向量则在局部区域实现更精准的属性控制，编辑效果更加连贯自然。

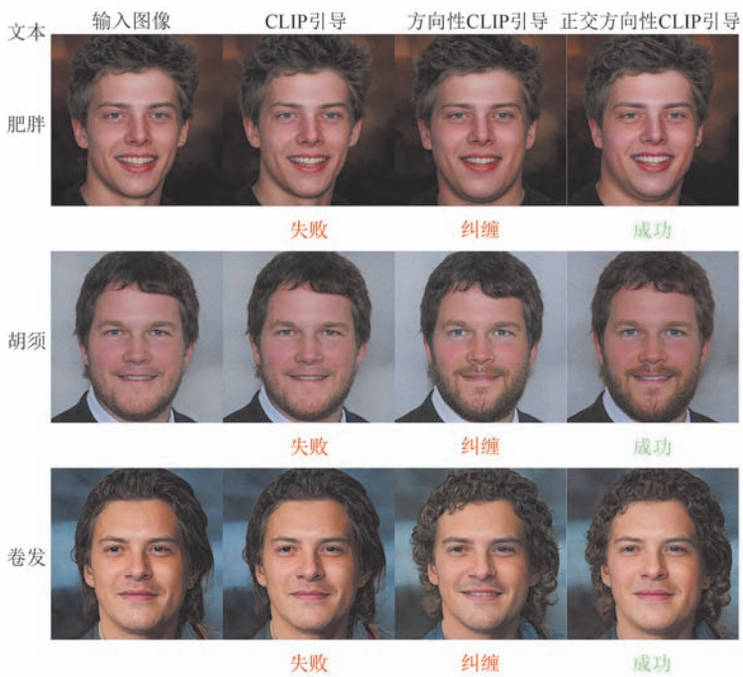

图10　解耦策略对比实验结果图

5　结　论

在本文中，我们提出了一种基于方向性CLIP引导的图像编辑方法—SMCI-CLIP。其核心是设计了语义调制模块（SMM），通过映射函数动态调整文本权重，灵活调节潜码以生成精准且高度解纠缠的编辑方向。此外，本文提出了正交方向性CLIP引导机制，通过在CLIP文本嵌入空间中构造目标属性与非目标属性的正交分量，有效剔除无关语义干扰，进一步提升了属性编辑的独立性与语义纯度。同时，本文还提出了一个多通道交互模块（MCI），实现潜码各通道与文本的深度交互，从而增强模型的全局感知能力。大量实验表明，SMCI-CLIP不仅能够在较短的训练时间内实现高效收敛，还能完成更精准、解耦性更强的编辑效果，为文本引导图像编辑提供了高效且解耦的解决方案。

参　考　文　献

[1] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139-144

[2] Patashnik O, Wu Z, Shechtman E, et al. StyleCLIP: Text-driven manipulation of StyleGAN imagery//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 2085-2094

[3] Wei T, Chen D, Zhou W, et al. HairCLIP: Design your hair by text and reference image//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 18072-18081

[4] Lyu Y, Lin T, Li F, et al. DeltaEdit: Exploring text-free training for text-driven image manipulation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 6894-6903

[5] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4396-4405

[6] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 8748-8763

[7] Chen Beijing, Zhang Haitao, Li Yuru. Active defense algorithm for face attribute editing via three-stage adversarial perturbation generation. Chinese Journal of Computers, 2024, 47(3): 677-689 (in Chinese)
(陈北京, 张海涛, 李玉茹. 面向人脸属性编辑的三阶段对抗扰动生成主动防御算法. 计算机学报, 2024, 47(3): 677-689)

[8] Zhang H, Wu C, Cao G, et al. HyperEditor: Achieving both authenticity and cross-domain capability in image editing via hypernetworks//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 7051-7059

[9] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of StyleGAN//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8110-8119

[10] Li P, Wang R, Huang H, et al. Pluralistic aging diffusion autoencoder//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 22613-22623

[11] Parihar R, Sachidanand V S, Mani S, et al. Precise Control: Enhancing text-to-image diffusion models with fine-grained attribute control //Proceedings of the European Conference on Computer Vision. Milan, Italy, 2024: 469-487

[12] Karras T, Aila T, Laine S, et al. Progressive growing of GANs for improved quality, stability, and variation//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1000-1009

[13] Härkönen E, Hertzmann A, Lehtinen J, et al. GANSpace: Discovering interpretable GAN controls//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2020: 9841-9850

[14] Huang Z, Ma S, Zhang J, et al. Adaptive nonlinear latent transformation for conditional face editing//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 21022-21031

[15] Yao X, Newson A, Gousseau Y, et al. A latent transformer for disentangled face editing in images and videos//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 13789-13798

[16] Shen Y, Gu J, Tang X, et al. Interpreting the latent space of GANs for semantic face editing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9243-9252

[17] Zhou Y, Zhang R, Chen C, et al. Towards language-free training for text-to-image generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 17907-17917

[18] Xu T, Zhang P, Huang Q, et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1316-1324

[19] Cheng J, Wu F, Tian Y, et al. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10911-10920

[20] Mirza M, Osindero S. Conditional generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[21] Xia W, Yang Y, Xue J H, et al. TediGAN: Text-guided diverse face image generation and manipulation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 2256-2265

[22] Wang H, Lin G, Hoi S C H, et al. Cycle-consistent inverse GAN for text-to-image synthesis//Proceedings of the 29th ACM International Conference on Multimedia. Virtual, 2021: 630-638

[23] Wu Z, Lischinski D, Shechtman E. StyleSpace analysis: Disentangled controls for StyleGAN image generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 12863-12872

[24] Tov O, Alaluf Y, Nitzan Y, et al. Designing an encoder for styleGAN image manipulation. ACM Transactions on Graphics, 2021, 40(4): 1-14

[25] Abdi H, Williams L J. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433-459

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[27] Gal R, Patashnik O, Maron H, et al. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. ACM Transactions on Graphics, 2022, 41(4): 1-13

[28] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 694-711

[29] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 586-595

[30] Zhu Y, Wu Y, Liu S, et al. One model to edit them all: Free-form text-driven image manipulation with semantic modulations//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 25146-25159

[31] Lee C H, Liu Z, Wu L, et al. MaskGAN: Towards diverse and interactive facial image manipulation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5549-5558

[32] Deng J, Guo J, Xue N, et al. ArcFace: Additive angular margin loss for deep face recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4690-4699

[33] Zhao H, Gallo O, Frosio I, et al. Loss functions for image restoration with neural networks. IEEE Transactions on

Computational Imaging, 2016, 3(1): 47-57
[34] Huynh-Thu Q, Ghanbari M. Scope of validity of PSNR in image/video quality assessment. Electronics Letters, 2008, 44(13): 800-801
[35] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612
[36] Hessel J, Holtzman A, Forbes M, et al. CLIPScore: A reference-free evaluation metric for image captioning// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic, 2021: 7514-7528

GU Guang-Hua, Ph. D., professor. His research interests include image understanding and facial attribute editing.

YANG Yuan-Hang, M. S. candidate. His research interests include facial attribute editing and image processing.

YI Bo-Yun, M. S. candidate. His research interests include image processing and computer vision.

Background

With the development of human society, people's social interactions have become increasingly broad and frequent. Due to the significant social attributes of the face, facial attribute research has attracted considerable attention from researchers. The face contains a wealth of valuable information and plays a crucial role in social processes by conveying rich emotional and personal identity information. These complex facial attributes represent each unique individual, thereby underscoring the importance of facial attribute editing research. The goal of facial attribute transfer is to replace a specific attribute in a given facial image with a target attribute while preserving the identity information. Traditional facial attribute editing tasks employed Generative Adversarial Network (GAN) architectures to modify facial attributes, and later evolved to modifying facial attributes in the latent space. In this paper, the facial attribute editing model adopts a text-guided latent space manipulation approach to edit facial image attributes, ensuring that while the target attribute is accurately modified, the remaining attribute details remain unchanged, resulting in more vivid and natural generated images.

This work was partly supported by the National Natural Science Foundation of China (Grant No. 62072394) and the Natural Science Foundation of Hebei Province (Grant No. F2024203049).