南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

本科生毕业设计

题　　目：基于扩散模型的文字编辑图像算法

姓　　名：吴泽敏

学　　号：11910934

系　　别：计算机科学与工程

专　　业：计算机科学与技术

指导教师：刘江，李衡

年　　月　　日

诚信承诺书

1.本人郑重承诺所呈交的毕业设计（论文），是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。

2.除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。

3.本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。

4.在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名：

_____年____月____日

基于扩散模型的文字编辑图像算法

吴泽敏

计算机科学与工程系　指导教师：刘江，李衡

[摘要]：白内障是指由于眼中晶状体混浊导致的视觉模糊或失明，开发相关的深度学习算法能实现白内障手术模拟，帮助医生预测手术结果，提高手术成功率。本项目通过生成白内障手术影像，解决深度学习算法缺少数据的问题。以目前在图像生成方面表现优秀的扩散模型为基础，加入插值函数，设计能解析目标文本语义信息并编辑图像的网络框架，并研究了不同插值参数对模型生成效果的影响。模型不仅实现动态地文本编辑图像，生成高质量的白内障手术数据及其语义分割图像，同时由于其能使用在其他数据集上训练的预训练模型，有效降低了扩散模型所需的训练时间。

[关键词]：扩散模型；文字编辑图像；插值实验；数据集生成

[Abstract]: Cataract refers to blurred vision or blindness caused by opacity of the lens in the eye, and the development of related deep learning algorithms can realize cataract surgery simulation, help doctors predict surgical results, and improve the success rate of surgery. This project solves the problem of missing data from deep learning algorithms by generating cataract surgery images. Based on the current diffusion model with excellent performance in image generation, an interpolation function is added, a network framework that can parse the semantic information of the target text and edit the image is designed, and the influence of different interpolation parameters on the model generation effect is studied. The model not only dynamically text-edited images to

generate high-quality cataract surgery data and its semantic segmentation images, but also effectively reduces the time required for diffusion model training because it can use pre-trained models trained on other datasets.
[Keywords]: Diffusion model; text-based image edit; interpolation experiment; data set generate

目录

1.引言

1.1 研究背景与意义

人工智能在医疗方面能极大地便利医生与患者，以算法为核心可以开发初步识别病灶的仪器和云平台，有效地缓解医疗资源紧张，分布不均匀的问题。以白内障手术为例，白内障是全球范围内造成病患失明或视力障碍的主要原因。白内障的唯一治疗方法是临床手术，随着白内障手术流程的优化和手术器械的进步，超声乳化术不仅能有效解决致盲问题，同时其并发症以及手术后的副作用较小，然而该技术由于医疗资源问题不能在全球范围应用。具体而言，发展中国家或人口大国中的白内障失明患者数量庞大，并随着人口老龄化日益增长，然而能实施白内障手术的外科医生与医疗资源并不能满足临床需求。使用深度学习，可以在现有的手术数据集上训练出语义分割模型、病灶检测模型等能有效缓解医疗资源紧张、分布不均匀的问题。

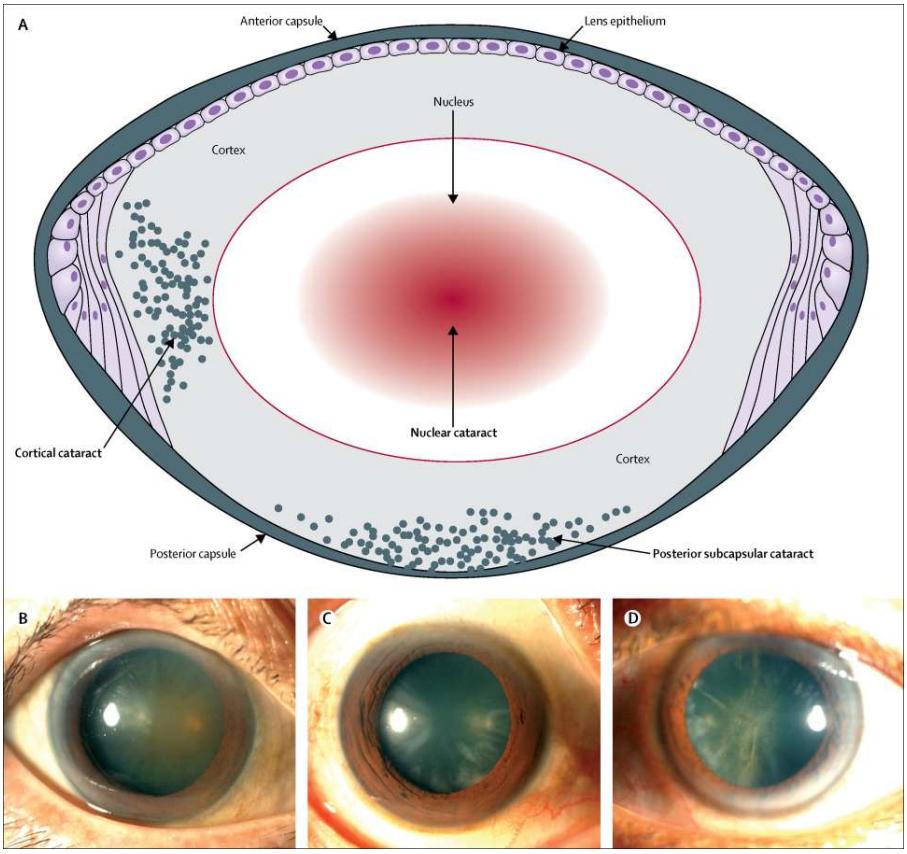

图 1 白内障医学图像[1]

当前医疗人工智能模型开发的普遍难点在于数据集缺乏，没有足够的数据量以及较低质量的标注，数据限制了模型的进一步发展。如下图所示，白内障以及腹腔镜的手术数据集量小，远远不如自然场景数据集，极大地限制了相关算法的研究。为此通过深度学习算法来生成图像，扩充数据集，满足其他模型的训练需求。本项目旨在设计一个能基于语义生成，保真度高且具有多样性的算法。该模型将扩充现有的白内障手术数据集，有助于白内障相关眼科算法的训练，提升其模型的训练效果。

表 1 数据集概述

数据集类型	大小
眼科手术数据集	CADIS: 3490+533
腹腔镜手术数据集	MESAD-real: 23366+2024
自然场景数据集	coco: 33 万

1.2 图像生成的挑战

深度学习中的图像生成模型普遍面临着训练时模型过拟合导致的模型塌陷问题、训练需要的数据量和运算量庞大，对生成结果缺乏客观的质量评估标准等问题。其中，使用文字引导图像生成的深度学习模型存在一下难点：

1. 有效的文字描述方式：模型中的文本编码模块学习的是复杂的语义表述，并且数据集中的文字信息和表述习惯受文化的限制，需要尝试不同的表述以有效地利用模型生成目标图像。

2. 领域偏移：模型在风景、物体等数据集上训练，在其他类的数据集上使用时，其效果会有所下降。模型生成图像的能力收到训练集的影响，若训练集不平衡、或是不包含特定类型的图像，则导致模型在实际应用时的性能波动大

3. 训练模型时间成本高：文字编码模块需要学习海量的文字信息以捕获其中有效的语义部分，同时图像生成模型在生成大尺寸的图像时所需的网络尺寸成倍增大，导致了高额的训练成本。

1.3 本文的创新点

本项目将以 Diffusion 为原理，将文本编码和真实图片作为条件输入，引导模型生成既保留真实图像的细节又贴合语义信息的医学图像，并能同时生成图像对应的语义分割图像。通过大批地生成配对的医学图像和分割图像，将扩充数据集，有助于医学相关算法的工作。与此同时，该项目使用 finetune 思想，使用在风景自然等大数据集上训练的预训练 unet 和 text encoder，在医学图像数据集上进行几分钟的 fintune 便可生成逼真且多样性高的医学数据。这一训练方式，不仅使得模型使用便捷，具有可部署成在线 demo 的能力，同时跳出了真实数据集量小，无法满足生成图像模型训练的制约。通过调节插值和随机种子，本模型可以由一张图像和一串文本生成多样的结果，或是输入不同的文本，对同一张图像进行多样的编辑。

1.4 论文的组织架构

本文将分为五个章节讲述：

第一章：引言。介绍了项目的背景与研究意义，总结了图像生成算法的挑战以及本文的创新点。

第二章：相关工作。该章节将概述近年主流的图像生成算法，分析各自的优缺点，并且对扩散理论发展中的经典模型结构进行介绍。

第三章：方法。该章节将详细介绍本项目设计的模型框架以及训练流程。首先将对扩散理论进行梳理，同时对框架中的文本编码模块、线性插值模块进行进一步阐述。

第四章：实验。在对实验使用的数据集以及对比实验使用的模型进行介绍后，本章节将重点放在对实验过程的说明，并从不同角度分析实验结果。同时列出了实验时的具体参数设置以及评价指标的选择。

第五章：结论。该章节对整体项目进行回顾，对提出的方法以及实验结果进行总结，在分析其局限性的同时展望未来工作。

2.相关工作

本章将介绍主流的图像生成算法 VAE、GAN 和 Diffusion 背后的原理，并对三种模型的优缺点进行分析。同时概括不同模型在不同任务上的表现，接着将详细介绍三种不同的基于 Diffusion 原理的经典模型架构。

2.1 图像生成算法

计算机视觉的任务中的超分任务、语义分割任务和图像修复任务等均可看成不同数据模态的图像生成任务。无监督的生成模型学习了数据分布与概率的关系之后，能从分布中生成新的样本作为输出，加入不同的条件进行引导则能改变输出，进而完成不同的任务。目前主流的图像生成算法可概括为四类，生成对抗网络 Generative Adversarial Network (GAN)[2]、变分自编码器 Variational Auto-Encoder (VAE)[3]、标准化流模型 Normalizing Flow (Flow)和去噪扩散模型 Denoising Diffusion Probabilistic Model (Diffusion)[4]。

模型基本框架如下图所示，其中 VAE 模型使用编码器将图像编码到符合标准正态分布的隐空间中，从隐空间重采样向量，通过 decoder 生成符合训练样本分布的图像。VAE 模型的优势在于其训练速度快，但相对应的缺点在于生成的图像质量往往不高甚至较为模糊[5]。

GAN 则在训练生成器的同时训练一个判别器，判别器的任务是尽可能地将真实的输入图像分类为真，将模型生成的图像分类为假，生成器与判别器博弈，希望生成的图像尽可能符合真实图像的分布，骗过判别器。GAN 训练希望最大化判别器的损失函数，最小化生成器的损失函数，从而无需对不同的任务设计特定的损失函数。通过判别器和生成器对抗学习的方式能生成较高质量的图像，在 Diffusion 模型成熟之前，是生成模型中的 SOTA，在许多任务中都具有优秀的表现。但是需要生成器和分辨器的训练时的更新程度保持一致，往往较难训练，容易出现模式崩塌、梯度消失的状况导致训练失败[6]。

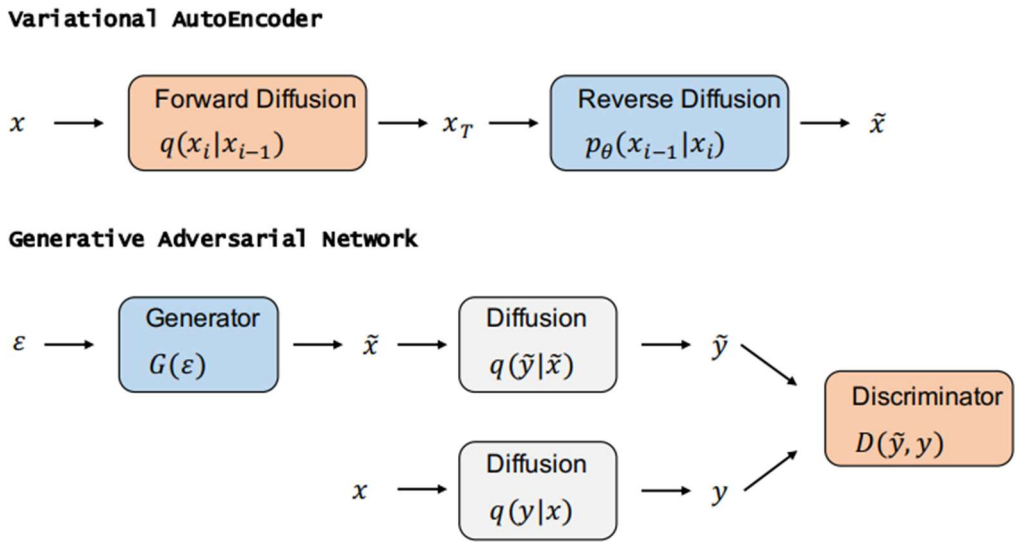

图 2 VAE 和 GAN 生成模型框架[11]

Diffusion 模型的数学原理为基于马尔科夫链的条件概率，在扩散过程中定义每一个时间戳 t 时添加噪声的方式，最终输入的分布将符合正态分布，通过训练简单的网络预测每一时间戳的噪声，逐步从随机采样的纯噪声中降噪，生成高清高质量的图像。扩散模型则具有前两者的优点，扩散模型只需要训练对纯噪声的去噪过程，所以训练简单且容易收敛，同时生成的图像质量高，丰富多样。不足之处在于训练所需资源较多，Diffusion 模型的采样速度较慢，通常需要采样数千个时间戳噪声，以生成一张质量较好的图像数据，图片生成的时间和训练时间从而长于 VAE 和 GAN 模型。

2019 年扩散理论发表，Denoising Diffusion Probabilistic Models 论文[4]对扩散理论进行了严谨的数学推导，展示易于实现的代码，完善了整个推理过程以及训练流程，为扩散模型的发展奠定了基础，后续的发展都继承了文中提出的前向加噪声、反向降噪生成的框架。随着近两年的急速发展，研究人员不断完善了扩散模型在训练时的细节[7]，提出了针对不同任务的训练框架[8–10]。目前 Diffusion Models 俨然成为生成模型中新的 SOTA，取代了原 SOTA：GAN，并且在诸多应用领域都有出色的表现，如计算机视觉、波形信号处理、多模态建模、分子图建模、时间序列建模、对抗性净化等[11]。为提高 diffusion 模型的应用价值，研究者们提出了 Discretization Optimization，Non-Markovian Process，Partial Sampling 三种提高采样效率，优化生成时间的方法[12]。在去噪过程中输

入其他信息作为条件引导生成过程，可以使得扩散模型实现基于文字、图像的条件生成[13–15]。

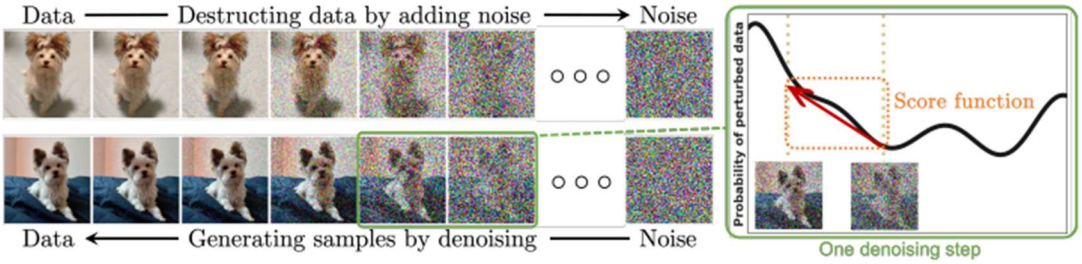

图 3 Diffusion 模型的正逆向过程示例

2.2 基于文字生成图像算法

基于文字编辑图像是近几年的热门研究领域，其主要任务是从一句描述性文本修改图像生成一张与文本内容相对应的图片。目前大多数基于文字内容编辑图像方法，局限于特定的编辑类型，例如物体叠放、风格迁移，或者应用于合成图像，或者需要多个输入图像。

2.2.1 clip 模型

clip（Contrastive Language-Image Pre-training）是 open-AI 发布的基于对比学习的多模态预训练模型，使用图像以及对应的文本描述作为训练数据，在捕捉图像语义信息与风格特征上具有良好表现，同时能将文本与图像进行匹配[16]。CLIP 主要由两个部分组成，基于 NLP 中 text transformer 的 text encoder 将文本编码成文本嵌入，提取文本的语义特征；基于 CNN 或是 vision transformer 的 image encoder 将图像编码成图像嵌入，捕获图像的语义特征与风格信息，如图 4 所示。通过对每一个 batch 的文本-图像对进行对比学习，计算图像嵌入和文本嵌入余弦相似性（cosine similarity），最大化正样本的相似性，最小化负样本的相似性作为损失更新 text encoder 和 image encoder 的参数。

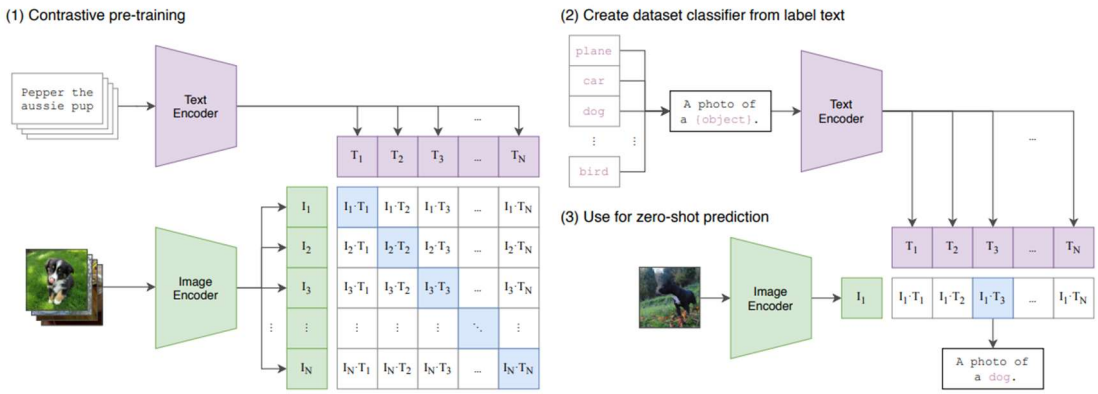

图 4 CLIP 模型训练以及生成示意图

该预训练模型通过在互联网上抓取 4 个亿的文本-图像对作为训练使用的数据集。将 text encoder 和 image encoder 作为模块加入具体任务的模型框架，只需要微调便能将 CLIP 中学习到的文本图像对迁移到具体任务中，从而实现图像文本相关的多模态任务，而不需要额外的训练，达到 one shot。论文作者使用 CLIP 模型和 ResNet50 在 27 个不同的数据集上对比其分类能力，其中 CLIP 在 26 个数据集上的表现远超 ResNet50，如图 5 所示，但是在一些特别的数据集上，clip 的表现较差，例如常用的 MNIST 数据集，研究者发现这是由于 CLIP 训练的数据集中并没有与 MNIST 相似的数据，即 MNIST 对于 CLIP 而言是域外数据，从而可知如果测试数据集的分布和训练集相差较大，CLIP 会表现较差。

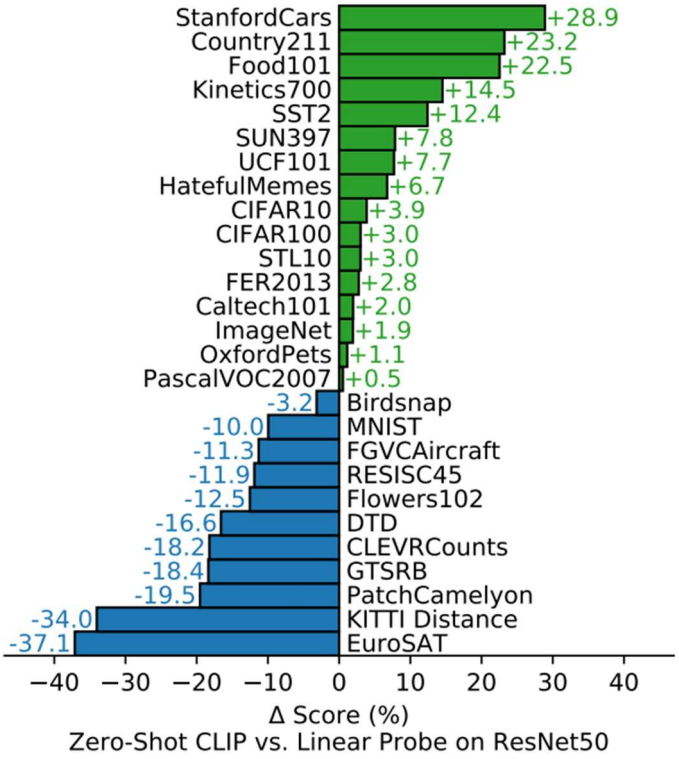

图 5 CLIP 模型和 ResNet50 在不同数据集上的分类效果对比

2.2.2 dalle2 模型

以文本描述语义和风格引导图像生成，不仅能实时地使用不同文本引导生成真实图像，而且由于语义信息只描述了关键的信息，输出的图像在非必要的细节上能呈现出多样性，极大地提高了数据库的丰富程度。DALLE2 模型摒弃了 DALLE 中使用的 transformer 模型库，而是以扩散理论为基础，结合了 CLIP 模型的优势，层级式地根据文本生成图像。文本不不仅能引导图像的语义、概念，同时可以修改图像的光线、纹理[17,18]。

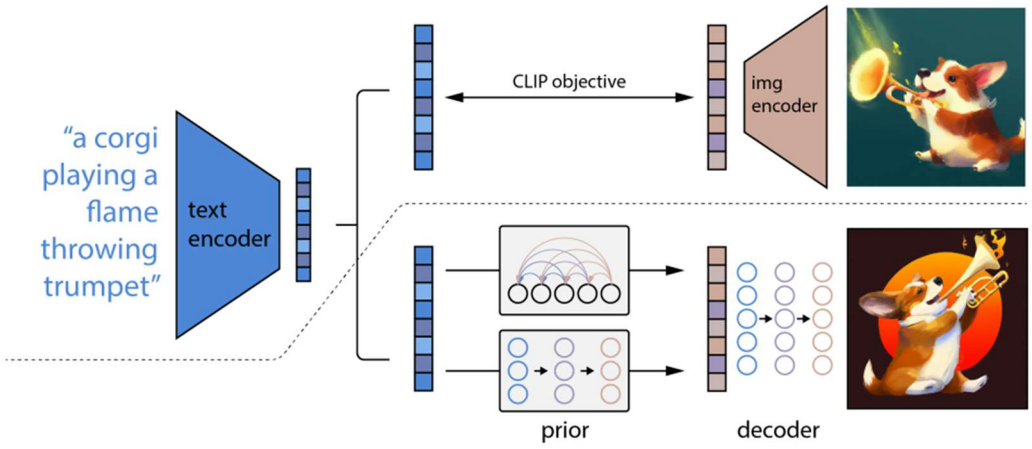

图 6 DALLE2 模型训练流程

该模型根据 CLIP 编码生成的文本特征和图像特征作为训练的先验，可以看成 CLIP 的反向过程，模型包括两个 Diffusion 模块，分别完成从文本嵌入生成图像嵌入的任务和从图像嵌入生成真实图像的任务。具体的训练过程分为三部分，首先使用已训练的 clip 将文本和图像对编码成文本嵌入和图像嵌入。其次训练以 Diffusion 为原理的 prior 模块，倒置 clip 中的映射，得到图像嵌入。最后将图像嵌入作为条件，引导同样以 Diffusion 为原理的 decoder 模块，生成指定 64*64 尺寸的真实图像，虽然牺牲了生成的多样性，但是提高了生成图像的真实性。在研究中发现，DALLE2 不能有效识别图像中文字的顺序，物体上下左右等空间关系，并且不能生成特别复杂的场景图像，细节的缺失较为严重。

图 7 DALLE2 模型的采样结果，目标文字为“A sign that says deep learning”

2.2.3Stable Diffusion

相较于之前的以扩散理论为基础的 text2image 模型框架 Latent Diffusion[19,20]，

Stable Diffusion 的特点在于其使用了较 CLIP 而言更大的在纯文本上训练的自然语言模型（T5）作为文字编码器，并指出训练效果比使用尺寸更大的 Diffusion Model 作为生成模块效果更显著，该模型能达到比之前模型更高的保真度，同时引导文本与生成图片的匹配程度更高[21]。

该模型的训练过程是，首先使用使用 T5-XXL 提取文本特征，将文本嵌入作为条件输入以 Diffusion 为原理的图像生成模型，引导其生成大小为 64*64 的真实图像。下一步是将第一步生成的图像作为条件，通过两个级联的超分模块，将图像放大到 1024*1024 的尺寸，超分模块同样以 Diffusion 为原理。Stable Diffusion 舍弃了图像文本编码器对，而是只使用了文字编码器，并且在预训练的基础上进行 fine tune 操作。

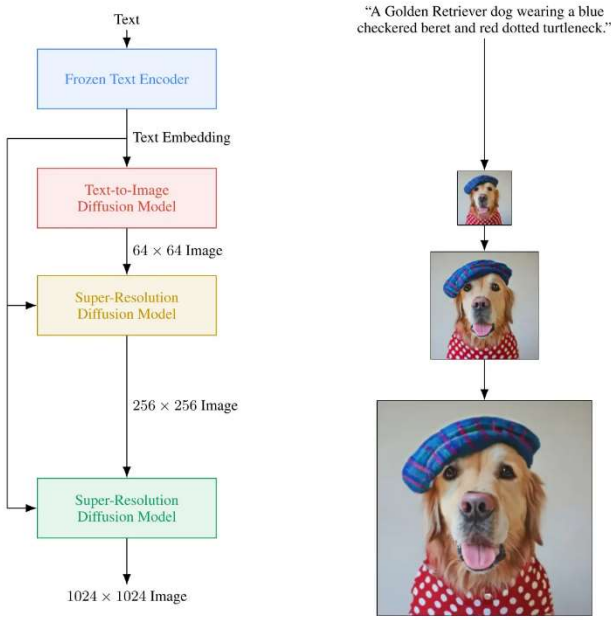

图 8 Stable Diffusion 模型以及训练流程

Stable Diffusion 的特点之一在于其训练时使用了更大的引导参数，以使得生成的图像与文本更加匹配。为了解决参数超过[-1,1]区间所造成的图像失真问题，作者提出了两种解决方法，一种为静态阈值，在生成图像采样噪声时将引导参数数值裁剪到[-1,1]范围内，可以避免生成过多空白图，另外一种是动态阈值，在每一个时间戳采样后生成的图像中随机选一个像素值 s，若 s>1，便将该步骤的图像的全部像素点裁剪到[-s,s]区间内，防止像素饱和。该模型的特点之二在于在生成图像时使用了联级的超分模块，可以在低分辨率时观察到模型的

表现，减少训练成本，同时在超分模块中加入了噪声增强，使得超分效果更鲁棒，提高生成的图像的质量。

2.3 本章小结

本章中对本项目研究领域的算法进行了概括。在 2.1 中将生成算法分为三类，阐述了三个经典模型的基本训练框架以及实验流程，分析其各自的优劣，同时对 Diffusion 的发展进行了总结。在 2.2 小节中详细介绍了多模态模型 CLIP 以及两个文字引导图像生成的扩散模型框架，描述了模型的训练细节。上述的两个模型为后续本项目的方法提供了设计思路，并指导了实验实践过程中的实现细节。

3.方法

本章将详细地描述基于扩散理论提出的文本生成图像模型框架，并解释插值函数背后的数学原理以及实现细节，对图像生成模块背后的扩散理论和无分类器的扩散引导理论进行进一步说明，最后提出训练流程以及超参数的选择。

3.1 text encoder

实现由文本生成图像的模型背后是在庞大数据库上训练的文本语义编码器，用以捕获复杂度高的语言表述中的信息。目前的文本编码器通常在文本图像对上进行训练，生成图像模型可以额外地在自己的数据库上训练文本编码器，或是使用预训练模型，减少运算消耗的资源。研究发现文本编码器有效地将文字中的重要信息编码成文本嵌入，其中的重要信息可以描述图像中的视觉信息。另一种能对文本进行编码的是大型语言模型，同样可以应用到文字生成图像任务中。(如 BERT b， GPT， T5)等模型的发展迅速，其对文本解读解读能力和生成能力的有了质的飞跃[21]。这是因为这些模型使用的数据集只包括了文本，数据量远远大于文本图像对数据集，进而导致模型能在训练过程中学习到更加广泛的分布。

本实验将使用 clip 模型中的作为 text encoder 模块作为预训练的文本编码器，其训练细节在相关工作已经做出了详细的介绍，clip 模型不仅可以完成 zero-shot 的分类任务，同时利用其编码器以及交叉熵的运算，可以完成文本引导的图像生成任务。

3.2 DDPM

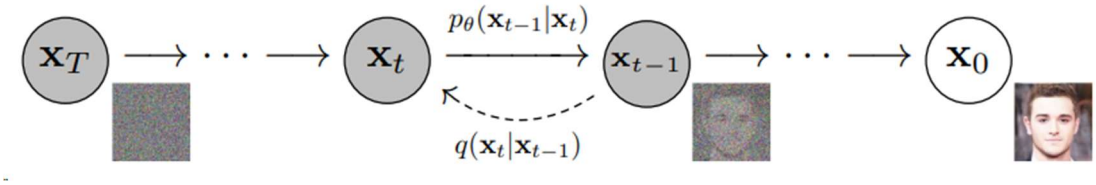

图 9 DDPM 采样过程

高斯扩散理论背后的数学原理主要基于马尔可夫链和条件概率分布，于 2015 年提出，2020 年得到进一步的完善， Denoising Diffusion Probabilistic Model(DDPM)。如上图所示，扩散模型分为两个过程，从图像逐步加噪声变成纯噪声的正向/扩散过程（forward/diffusion process）和从纯噪声逐步去噪变成图像的逆向/去噪过程（reverse/denoised process）。对于一个随机的数据样本，$\mathbf{x}_0$，通过定义加噪声的方式α_t，如公式(1)所示，对图像逐步添加高斯噪声，得到马尔科夫链$\mathbf{x}_0\ \mathbf{x}_1 \ldots \mathbf{x}_t$。同时，利用递推关系，可以得到由$\mathbf{x}_0$到任意$\mathbf{x}_t$的公式(2)，即实现添加噪声到输入图像$\mathbf{x}_0$得到任意时间戳的图像$\mathbf{x}_t$，便于后续的训练。随着 t 的增大，往图像上覆盖的噪声越多，$\mathbf{x}_t$越接近标准正态分布。

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{1}$$

$$q(x_t \mid x_{t-1}) := \mathcal{N}\left(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}\right) \tag{2}$$

训练过程的伪代码如图所示，使用的参数 T 为 2000，即生成时采样 2000 步，对输入的 batch，随机取 t 值，通过训练 Unet 对 t 时刻的噪声进行预测，利用模型公式 3，则可以从随机采样的x_t逐步去噪$x_{t-1}\ x_{t-2} \ldots x_1\ x_0$，得到逼真的图像$x_0$。训练时的损失函数如公式(4)所示，目标为预测的噪声与时间戳 t 时公式(2)计算出的噪声一致，其中ϵ为采样的随机高斯噪声，均值μ_θ的定义如公式(5)所示。

$$p_\theta(x_{t-1} \mid x_t) := \mathcal{N}\left(\mu_\theta(x_t), \Sigma_\theta(x_t)\right) \tag{3}$$

$$|| \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\right) - \epsilon || \tag{4}$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)\right) \tag{5}$$

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

图 10 DDPM 实现细节

3.3 Classifier-free Guidance

在生成任务中加入条件引导生成过程能得到更好的输出结果，这是因为加入条件后限制了模型输出的多样性。OpenAI 的 Guided Diffusion 提出了一种简单地引导扩散模型生成指定类别图像的方法[13]。其主要思想是在采样的每一个时间戳，将图像输入至分类网络，以网络的交叉熵损失作为梯度，引导下一次采样。该方法不需要对已有的网络进行再次训练，但是延长了采样每一张图片所需的时间。通过使用不同类型的分类网络，可以实现基于文本、图像、多模态引导生成的扩散模型。

Algorithm 1 Joint training a diffusion model with classifier-free guidance

Require: p_{uncond}: probability of unconditional training
1: **repeat**
2: $(\mathbf{x}, \mathbf{c}) \sim p(\mathbf{x}, \mathbf{c})$ ▷ Sample data with conditioning from the dataset
3: $\mathbf{c} \leftarrow \varnothing$ with probability p_{uncond} ▷ Randomly discard conditioning to train unconditionally
4: $\lambda \sim p(\lambda)$ ▷ Sample log SNR value
5: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
6: $\mathbf{z}_\lambda = \alpha_\lambda \mathbf{x} + \sigma_\lambda \boldsymbol{\epsilon}$ ▷ Corrupt data to the sampled log SNR value
7: Take gradient step on $\nabla_\theta \left\|\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - \boldsymbol{\epsilon}\right\|^2$ ▷ Optimization of denoising model
8: **until** converged

图 11 Classifier-free 训练细节

由于 Guided Diffusion 不仅增加了采样的时间，同时使用的分类网络与生成模块是分开训练的，不能通过同步训练提升模型效果，谷歌的 Jonathan Ho 等人

提出了不依赖额外分类网络的引导扩散模型生成的方法，Classifier-free guidence[22]。如图所示，训练时将条件 y 与时间 t 作为 Unet 预测噪声网络的输入，将无条件生成与条件生成结合，将标签 y 以设定的概率替换成空标签∅，从而预测网络能同时支持无条件和有条件的噪声估计。

Algorithm 2 Conditional sampling with classifier-free guidance

Require: w: guidance strength
Require: $\mathbf{c}$: conditioning information for conditional sampling
Require: $\lambda_1, \ldots, \lambda_T$: increasing log SNR sequence with $\lambda_1 = \lambda_{\min}, \lambda_T = \lambda_{\max}$
1: $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = 1, \ldots, T$ **do**
▷ Form the classifier-free guided score at log SNR λ_t
3: $\tilde{\boldsymbol{\epsilon}}_t = (1 + w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t)$
▷ Sampling step (could be replaced by another sampler, e.g. DDIM)
4: $\tilde{\mathbf{x}}_t = (\mathbf{z}_t - \sigma_{\lambda_t}\tilde{\boldsymbol{\epsilon}}_t)/\alpha_{\lambda_t}$
5: $\mathbf{z}_{t+1} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_{\lambda_{t+1}|\lambda_t}(\mathbf{z}_t, \tilde{\mathbf{x}}_t), (\tilde{\sigma}^2_{\lambda_{t+1}|\lambda_t})^{1-v}(\sigma^2_{\lambda_t|\lambda_{t+1}})^v)$ if $t < T$ else $\mathbf{z}_{t+1} = \tilde{\mathbf{x}}_t$
6: **end for**
7: **return** $\mathbf{z}_{T+1}$

图 12 Classifier-free 采样细节

在采样的过程中，通过公式 6，结合有条件和无条件估计的噪声去噪，使用引导参数 s 引导 xt 的生成沿 θ (xt|y)，远离 θ (xt|∅)方向。

$$\hat{\epsilon}_\theta(x_t \mid y) = \epsilon_\theta(x_t) + s \cdot \big(\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t)\big) \tag{6}$$

其梯度可以写成

$$\nabla_{x_t} \log p^i(y \mid x_t) \propto \nabla_{x_t} \log p(x_t \mid y) - \nabla_{x_t} \log p(x_t) \propto \epsilon^*(x_t \mid y) - \epsilon^*(x_t) \tag{7}$$

无分类器引导有两个优势第一，首先，生成模型在训练过程能学习到条件相关的分布情况，而不依赖于其他分类器。当条件信息复杂，分类器难以预测时，无分类器引导能简化训练过程。其次，模型训练完成后，在采样时能通过动态地调节引导参数的取值从而得到不同的生成结果。其缺陷在于对于不同的任务需要重新训练生成模块，同时预测时每一步都需要运算两次噪声估计网络，需要更多计算资源。

3.4 插值函数

使用文本编码器将文本编辑成文本嵌入后，将文本嵌入作为条件，引导图像的生成使得模型能完成基于文字引导的图像生成任务。文本嵌入中包含了生

成模型重建数据分布所需的语义信息和风格特征，而使用支持无条件和有条件预测噪声的生成网络，可以通过插值的方式能实现生成不同类型的图像[18]。以文本嵌入为条件，使用不同的引导参数，可以生成具有文本内容的不同图像，随着 s 增大，图像更匹配编码器保留的文本信息。为实现文本编辑图像任务，可以使用额外的编码模块将原图像编码为对应的文本嵌入 p0，将目标文本编码为 p1，在两者之间使用插值函数，得到介于两者之间的文本嵌入，作为条件输入引导图像生成，通过使用不同的插值函数，可以得到既保留原图像细节又符合文本内容的逼真图像，达到文本编辑图像的效果。本项目将使用线性插值函数和球面插值进行实验。

$$\alpha \cdot e_{text}+(1-\alpha) \cdot e_{image} \tag{8}$$

$$Slerp(p,q,t)=\frac{\sin[(1-t)\theta]\cdot p+\sin(t\theta)\cdot q}{\sin\theta} \tag{9}$$

3.5 模型架构

模型框架和训练流程如图所示，E1 和 E2 为预训练的图像和文本编码器，能将文本和图像进行有效地对应；F 为使用的插值函数，能在生成时动态地选择生成图像的分布；DM 为基于扩散原理的预训练生成模块，以插值后的文本嵌入为条件生成 64*64 大小的逼真图像；SR 为基于扩散原理的超分模块，以生成模块的输出为条件，生成 512*512 大小的数据。

模型能实现动态地文本编辑图像的功能，在预训练模型的基础上，对每一对目标文本和待编辑图像进行微调。训练流程为，首先 E1 将待编辑图像编码为对应的嵌入 e image，fine tune 一定步数生成模块，使得生成模块能更好地学习输入图像的细节。其次 E2 将文本编码为 e text，使用插值函数得到介于两者之间的目标嵌入 e input，作为条件引导生成模块的采样过程，从而找到既能与原图像有较高的保真度，又与目标文本描述一致的图像。最后经过初步的筛选以及指标的计算，使用 SR 模块将生成图像放大至 512*512 的尺寸。

该模型在有效利用了预训练模块的同时，通过短时间的优化步数，能达到实时编辑图像的操作。同时由于是使用插值后的嵌入向量作为条件引导，既可以生成不同分布的图像，同时可以满足输出的多样性。

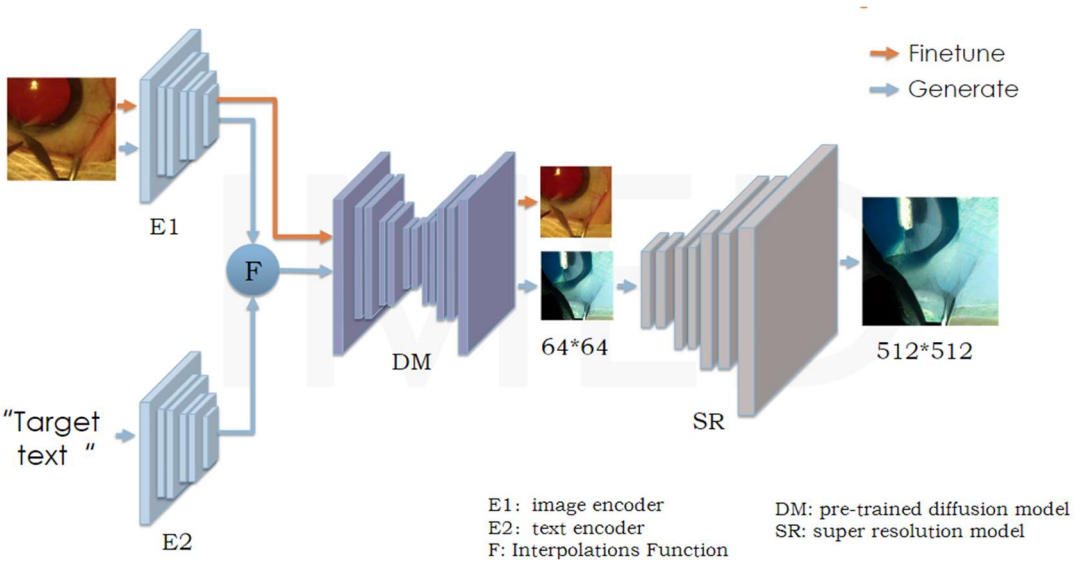

图 13 模型架构与训练流程

3.5 本章小结

本章详细介绍了提出的模型框架的原理以及思路。在 3.1 小节中介绍了不同的文本编码器，在 3.2 小节中详细描述了扩散原理和 DDPM 的训练方式，为本项目的模型提供了理论依据。在 3.3 小节中介绍了两种条件引导生成的方式，本项目使用了无分类器引导扩散的方式。3.4 小节解释了对插值函数在文本引导图像生成任务上的意义，最后在 3.5 小结展示了本项目设计的训练流程以及实现方式，解释了其可行性。

4.实验

本章将介绍实验的过程并分析实验结果。首先对使用的白内障手术数据集进行概述，接着介绍实验的详细流程，包括使用的硬件，预训练模型的参数，对比模型的选择，以及评估指标。最后将展示可视化的实验结果并对其分析。

4.1 数据集

4.1.1 数据集概况

本项目的目标之一是使用扩散模型生成文本编辑后的逼真白内障手术数据，同时生成对应的语义分割标签，平衡数据集，有助于提升白内障手术图像分割模型结果。数据集的质量在模型训练中起着重要的作用，需要准确的手术数据集标签图用于后续的指标运算。其中 CADIS[23]数据集有 3550 张训练集、534 张验证集和 586 张测试集，标签种类为 36 类，其语义分割图像标签齐全，包括瞳孔、虹膜以及各类手术器材，标注质量较高。InSegCat[24]数据集有 237 张训练集、61 张验证集和 88 张测试集，不仅数据量较少，而且标注质量不高，其标签只包含了不同的器械，缺少瞳孔虹膜等信息，同时其标签存在错误，需要进一步的处理。CATARACT 数据集是无标签的公开白内障手术数据集，其数据量为 1120 且并未分成训练集和测试集，有待后续处理。

表 2 白内障数据集概括

数据集	样本量	标签种类	现状
CADIS	Training:3550 Validation:534 Test: 586	36 类	标签齐全，包括瞳孔、虹膜、手术器材
InSegCat	Training：237 Validation:61 Test: 88	11 类	标签不齐全，只有器械
CATARACT	1120，未分组	15 类	组内使用 Roboflow 分批际注

4.1.2 数据集处理

为了完善 InSegCat 和 CATARACT 数据集的标签，提高数据集的标注质量，本项目对数据进行了进一步的处理。其中使用了 Pair 标注软件，在组内成员的合作下，将 InSegcat 的标签种类由 11 类扩充至 15 类，并且将不清晰的图像剔除，修改了错误标注的标签。在标注过程中，首先使用软件的智能标注工具，

该工具以像素层面对图像进行分割，在边缘上比多角形标注工具更贴合。由于手术中出现气泡以及光线的不足，导致部分图像标注前需要调高对比度。其次使用了 Roboflow 在线标注网站，将 CATARACT 分成两个批次进行标注，为其设计了 15 类标签，以便后续的模型训练。

图 14 数据集图例

4.2 实验细节

4.2.1 软硬件环境与参数配置

训练所使用的 GPU 为 VIDIA GeForce RTX 3060，12GB 显存，8 卡，深度学习框架 pytorch 版本为 1.8.1+cu101，文本编码器、图像编码器和基于 Diffusion 原理的图像生成模块均载入 huggingface 上公开的 diffuser 库，checkpoint 为 stable-diffusion-v1-5。

训练时使用的 Pipeline 为 StableDiffusionPipeline，优化图像嵌入生成的步数为 500，优化生成模块的的步数为 1000，学习率分别为 1e-3 和 1e-6，输入图像中心裁剪至 128*128 分辨率，训练一对目标文本和待编辑图像并保存其对应的文本嵌入以及图像嵌入，所需的时间约为 15min。在完整数据集上训练生成模块时，使用统一的文本表述：“a photo of cataract surgery with eye ball and pupil and surgical instrument”，学习率为 1e-05，epoch 为 15000。

生成时使用 DDIM 提高采样效率，采样 50 步，该模块的 bata 初始值为 0.00085，结束值为 0.012。在文本和图像嵌入之间做插值函数，生成目标嵌入，目标嵌入对生成模块的引导参数设置为 3，生成批大小（batch size）为 30。导入优化后的模型并生成图像需要时间约为 1min。

4.2.2 对比方法介绍

对比实验可以分为两部分，首先对比不同的图像生成算法类型在生成白内障图像上的直观表现。使用的算法为 DIP VAE[25]、Pix2Pix[26]和 Cycle GAN[27]，在 CADIS 数据集上训练。若 Diffusion 模型在白内障手术图像上的生成效果更优，则表明本实验具有可行性。其次对比了基于 Diffusion 的不同网络框架，分别从无条件生成图像、文本引导图像生成、文本编辑图像生成三种图像生成方式展现扩散模型的性能，若本项目提出的网络框架在生成图像方面更符合医学图像，则进一步说明本项目选择文本编辑图像生成这一方式具有合理性。

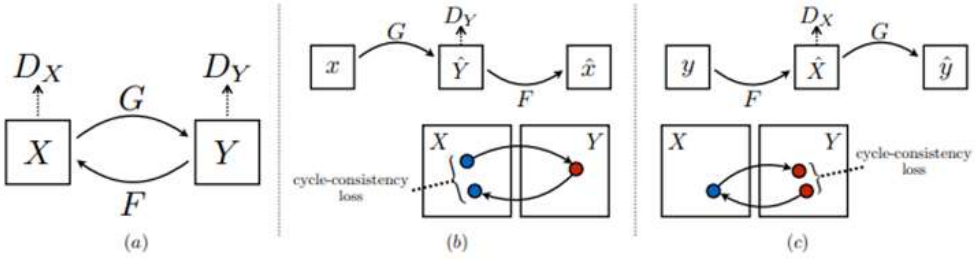

图 15 cycleGAN 训练过程图

4.2.3 实验任务介绍

在完成对比实验后，本项目的主体实验将分为三个步骤进行。

首先在 CADIS 数据集中随机抽取样本，目标文本为“a black and white photo of cataract surgery with eye ball and pupil”，对预训练模型进行短步数的优化，使用线性插值函数，对不同的α取值进行试验，对比插值函数对生成结果的影响，同时验证训练过程的可行性。

第二步则在验证了模型能根据文本编辑图像后，将样本从随机单张真实图像变为随机真实图像和对应语义标签联合后的图像，验证模型是否能同时编辑真实图像以及对应的语义标签，同时加入了球面插值函数，对比线性插值函数和球面插值函数在引导生成图像的效果。

最后使用完整的 CADIS 和 InSegCat 数据集对生成模型进行再训练，提高生成模型对白内障数据集的生成多样性，观察在提高生成多样性的同时，生成图像和文本的匹配程度是否被影响。

4.2.4 评价指标

PSNR ↑: 峰值信噪比，Peak Signal-to-Noise Ratio。用于衡量两张图像之间差异，例如压缩图像与原始图像，评估压缩图像质量；复原图像与 ground truth，评估复原算法性能等。

$$\mathrm{PSNR}(f,g) = 10\log_{10}(255^2/MSE(f,g)) \tag{10}$$

$$\mathrm{MSE}(f,g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(f_{ij}-g_{ij})^2 \tag{11}$$

SSIM ↑: 结构相似性，Structural Similarity。基于人眼会提取图像中结构化信息的假设，比传统方式更符合人眼视觉感知。

$$\mathrm{SSIM}(x,y) \quad = (l(x,y))^{\alpha}(c(x,y))^{\beta}(s(x,y))^{\gamma} \tag{12}$$

其中 l 表示亮度（luminance component）亮度以平均灰度衡量，通过平均所有像素的值得到。

$$l(x,y) = \frac{2\mu_{\mathbf{A}}\mu_{\mathbf{B}} + C_1}{\mu_{\mathbf{A}}^2 + \mu_{\mathbf{B}}^2 + C_1} \tag{13}$$

c 表示对比度，通过灰度标准差来衡量。

$$c(x,y) = \frac{2\sigma_{\mathbf{A}}\sigma_{\mathbf{B}} + C_2}{\sigma_{\mathbf{A}}^2 + \sigma_{\mathbf{B}}^2 + C_2} \tag{14}$$

s 表示结构，可以用相关性系数衡量。

$$s(x,y) = \frac{\sigma_{\mathbf{AB}} + C_3}{\sigma_{\mathbf{A}}\sigma_{\mathbf{B}} + C_3} \tag{15}$$

Clip-score ↑：生成图像与目标文字的吻合程度。

$$\mathrm{CLIP}-\mathrm{S}(\mathbf{c},\mathbf{v}) = w * \max\left(cos(\mathbf{c},\mathbf{v}),0\right) \tag{16}$$

1′LPIPS ↑：生成图像与原图像的吻合程度。

$$d(x,x_0) = \sum_{l}\frac{1}{H_l W_l}\sum_{h,w} || \, w_l \odot (\hat{y}_{hw}^{l} - \hat{y}_{0hw}^{l}) \, ||_2^2 \tag{17}$$

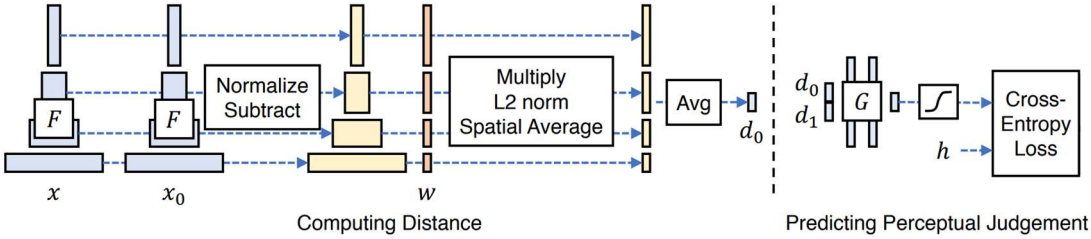

图 16 1′LPIPS 模型图

4.3 实验结果

4.3.1 图像生成任务对比实验

如图所示，无条件生成图像的基于 sr3 的 DDPM 模型[28]在图像的真实度上效果优秀，但是图像的风格变化不明显，并且图像直接相似度高。该模型能支持同时无条件生成手术图像与语义分割标签，且标签与图像对应，质量较高，能改善数据集的分布问题。

Stable Diffusion 在以文字为条件生成的图像的任务上，生成的图像具有丰富性、真实度高，但是模型对于描述白内障手术图像的文本并不敏感，大部分生成结果并不符合医学图像的需求，需要多次尝试，并且生成过程容易塌陷，全黑图像出现频率高。

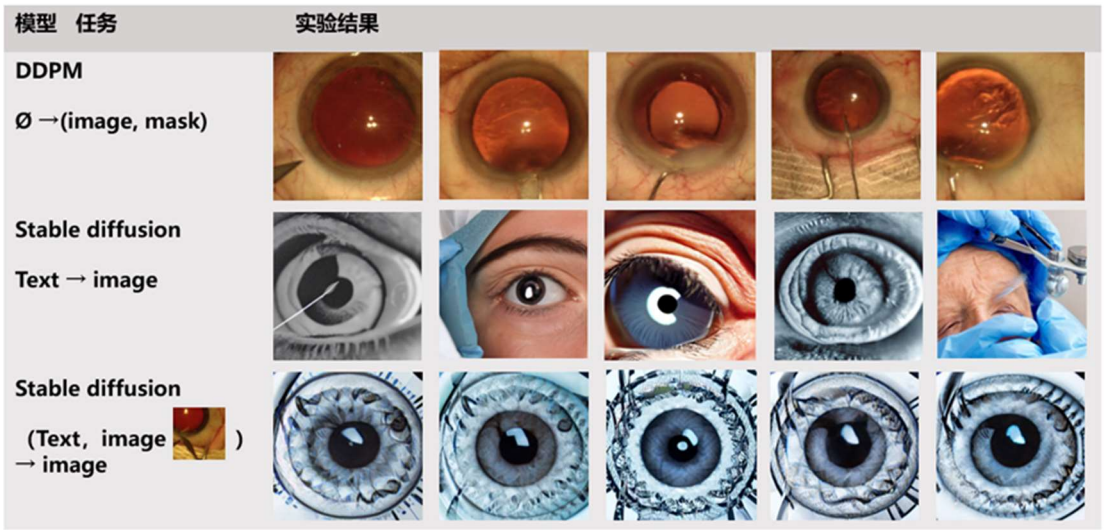

图 17 基于 Diffusion 的图像生成任务对比实验

而 Stable Diffusion 具有多任务的优势，可以支持多模态的文字和图像作为

输入条件，引导生成的图像在清晰度和风格转换上具有突出效果，通过不同的参数选择能达到实时文本编辑图像的效果，但是生成图像与原图像相差过大，无法直接用于扩充数据集，且生成过程塌陷概率高。

4.3.2 图像生成模型对比实验

VAE 复现了[25]论文中的方法，生成任务为无条件生成白内障手术图像，可以生成大小为 64*64 的多样图像，清晰度不加的同时生成图像中出现粗边黑框，在生成效果上不如 Diffusion 完成的无条件生成图像。Pix2Pix 以语义分割图像为条件，生成对应的真实图像，图像在语义上贴合输入，但是在细节上有所扭曲。CycleGAN 在训练过程中出现过拟合情况，对于不同的语义分割输入图像，模型均生成与引导语义不符合的，相似度很高的图像作为输出。

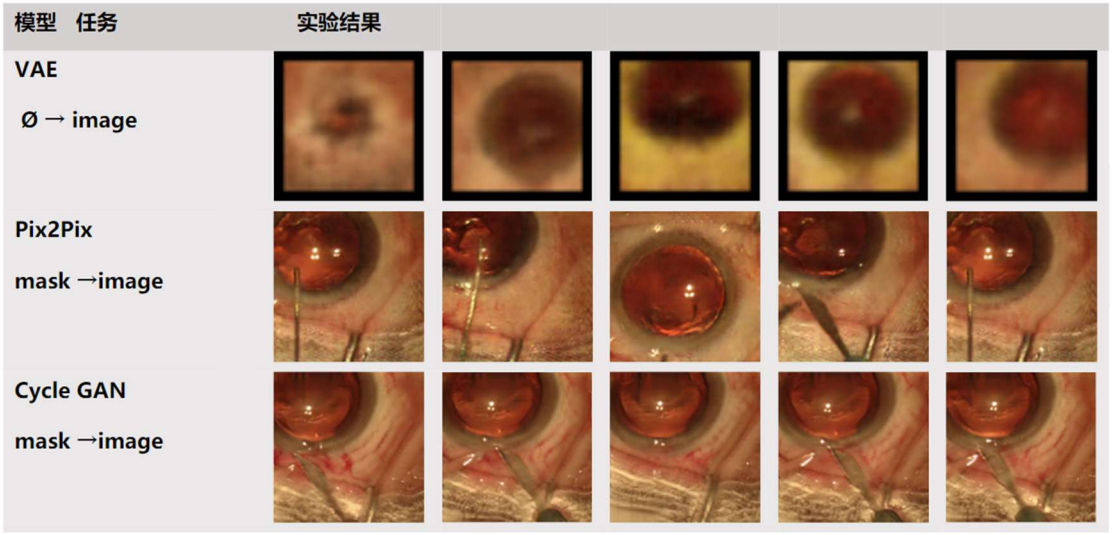

图 18 不同生成模型的对比实验

4.3.3 文字编辑图像实验

实验的第一步是搭建模型之后使用随机图像和文本“a black and white photo of cataract surgery with eye ball and pupil”作为模型的输入，对验证提出方法的可行性，使用的插值函数是线性插值，并探索了在α取值对生成结果的影响，使用的评价指标为 psnr 和 ssim。

在结果的可视化展示中，直观反映了在α=0.6 时生成的结果与原图差别不大，在 1.2 和 1.5 时生成的图像具有了风格的变化的同时并不会极大地改变图像

中的语义信息，但是并不是所有输出图像均符合文本的需求。当α=1.7 时生成的图像有了较大的扭曲。生成结果对于不同 seed 有不同的结果，说明一对文本图像输入可以生成无数的结果。实验过程中发现，随着α的取值超出[-1,1]，生成过程中塌陷频率变高，且随着α取值逐渐极端，塌陷频率逐渐升高。

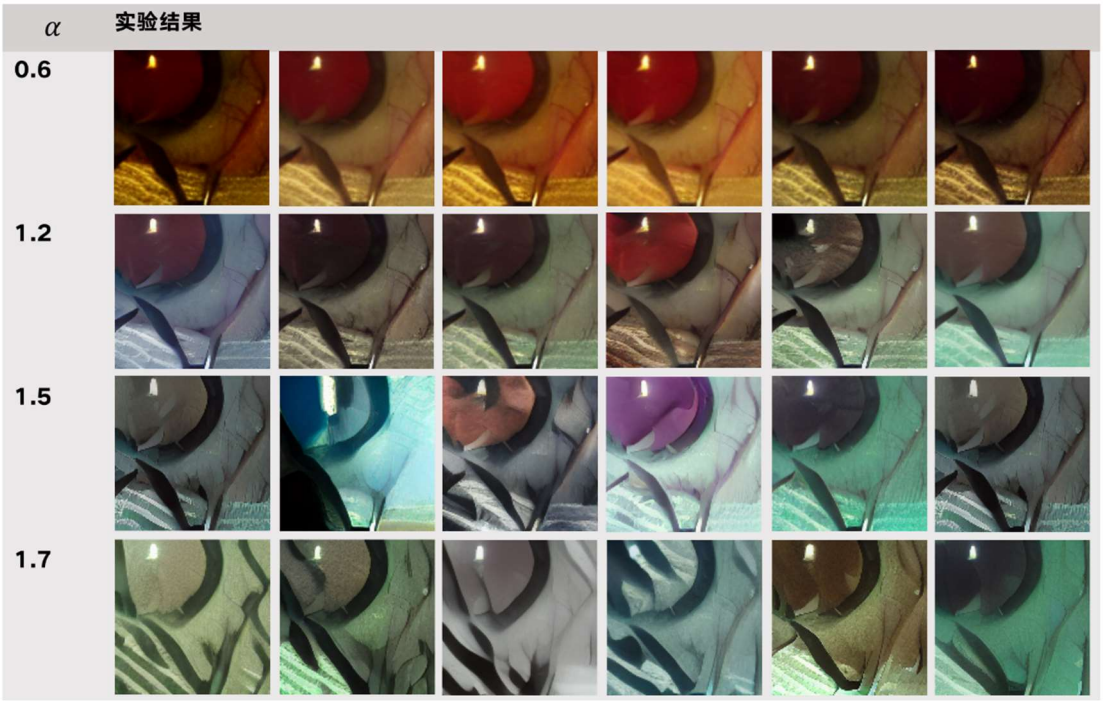

图 19 图像生成任务下的线性差值实验

如图 20 所示，本项目探索了插值参数对生成图像的文本对齐程度和图像保真程度的影响，直观对比发现，当参数取值范围在[1. 4, 1. 7]内，生成图像既能保持原图像的特征，目标文本又能对图像进行多样的编辑。

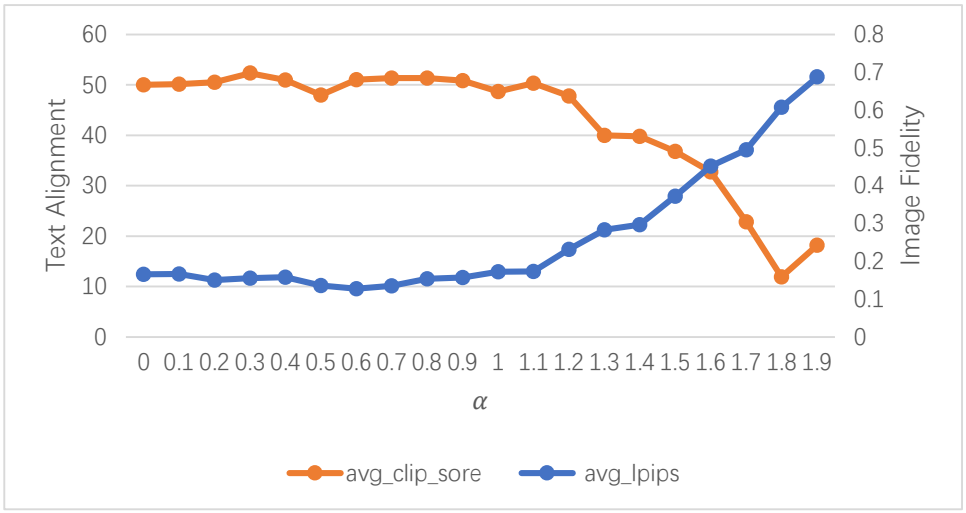

图 20 文本对齐度与图像保真度权衡

若以生成多样性且合理的数据集作为实验的目的，可以不考虑输出图像与文本的对齐程度，以其他指标为准。从第一步实验中可以看出 fine tune 思想的可行性，仅通过数十分钟的 fine tune 可以使得在其他数据集上预训练的编码模块和生成模块有效地学到文本嵌入代表的图像，但是模型对于不同的器械，以及图像的多样性、合理性方面有所不足，表现为对原图像的有效编辑只局限于色调。

在验证了模型的可行性以及线性插值函数的作用后，进入实验的第二阶段，将插值函数从线性插值修改为球面插值，研究球面插值函数对生成结果多样性的提升效果，同时尝试训练模型同时学习图像以及对应的语义分割模块的信息。

图 22 中直观展示了部分实验结果，实验发现球面插值与线性插值不同，不需要大幅度修改插值参数便可以极大地影响生成的结果，球面插值能以一对文本图像作为输入，对图像的风格做出各个方向的修改，在提高数据集的丰富性上比线性插值更有效。但是于此同时，生成过程中模型塌陷生成黑色图的频率较高，对不同插值参数的取值实验中发现插值参数为 0.1 0.6 0.4 0.8 时生成的图像过于扭曲，通过图 21 可视化插值参数对 PSNR 以及 SSIM 影响，无法表达语义信息，说明球面插值函数的不稳定，以牺牲模型稳定性换来了生成结果的多样性。

图 21 球面实验结果

为了探索模型在数据生成上的能力，本项目尝试训练模型同时生成逼真的医学图像以及其对应的语义分割标签。模型使用的生成模型继承了 Latent Diffusion 的思想，使用 encoder 将图像编码至隐空间中作为先验，在隐空间中使用扩散理论生成图像在隐空间中的表示，最后使用 decoder 将图像表示恢复为图像。由此模型能通过简单修改 encoder 和 decoder 的网络尺寸，在不替换隐空间的生成模块的同时，支持不同尺寸的数据生成任务，实验的其他细节与第一步的具体实现方式一致。

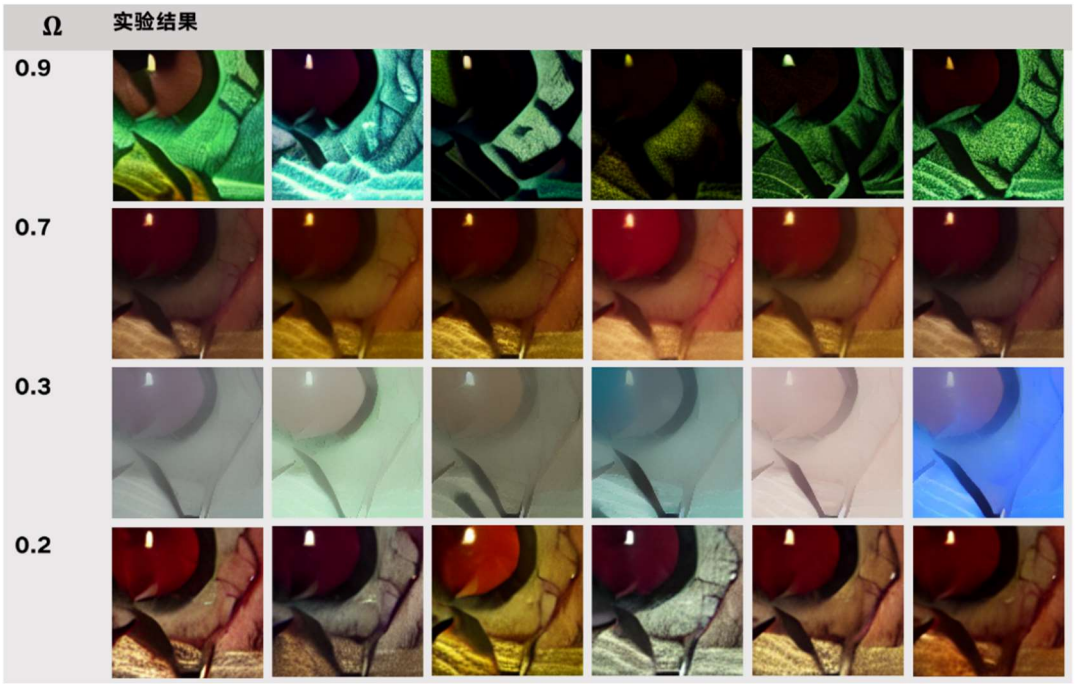

图 22 图像生成任务下的球面差值实验

在图像标签对的生成任务中，实验尝试了线性插值和球面插值两种插值函数，可以发现模型能有效地生成图像标签对，但是当微调插值参数时，生成模块仍能维持生成的逼真图像的细节以及语义信息，但是标签的风格被剧烈修改，生成的结果中线性插值函数会扭曲标签的语义信息，而球面插值对标签的语义信息捕捉更好，但是图像的细节质量下降。

此时由于未训练生成模块，白内障手术数据集对于生成模块以及语义编码模块而言属于域外数据，模块不能很好地利用其在预训练中学习到的分布，反映在结果上是文本编辑图像的结果只有色调的改变，同时模型不能很好识别

mask 信息，mask 的像素值代表着其标签值，mask 的灰度不统一会导致标签无效，无法加入数据集中训练其他模型。

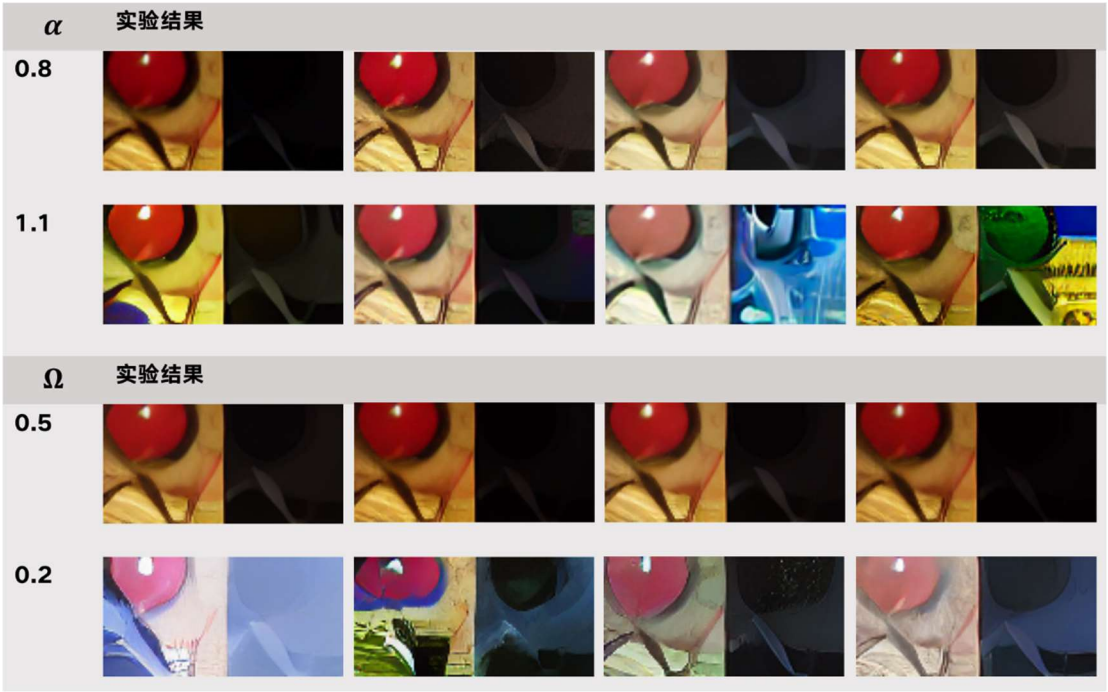

图 18 图像标签对生成任务下的差值实验

在前两步的基础上，实验进入第三步，即在白内障手术数据集上批量训练生成模块和编码模块。本项目在 CADIS 和 CATARACT 两个数据集上分别处理数据集为文本图像对，训练步骤分为三步：首先使用数据集训练 Diffusion 的 Unet 生成模块和 text encoder 模块，接着输入 target 文本和初始 image，由 encoder 得到两者的 embedding，同时再次 fine tune Diffusion 的生成模块，最后在两个 embedding 之间做线性或球面插值，得到由文本编辑过的图像。在插值实验中发现对生成模块的再训练有效提高了语义分布的多样性，并且生成的语义分割图像在风格上与原标签相似度高，并且语义分割图像中的器械以及瞳孔信息与逼真图像一致，标注质量高于第二步中的实验结果。

但是模块对目标文本的敏感度降低了，表现在于使用与前两步同样的文本图像对输入时，生成图像的风格并不会因插值参数的取值不同而有多差异。同时，由于对 text encoder 模块训练时使用的文本并不包含白内障手术图像中的具体器械类型以及位置信息，所以无法通过文本描述目标图像中出现的器械指导生成模块生成所需图像。

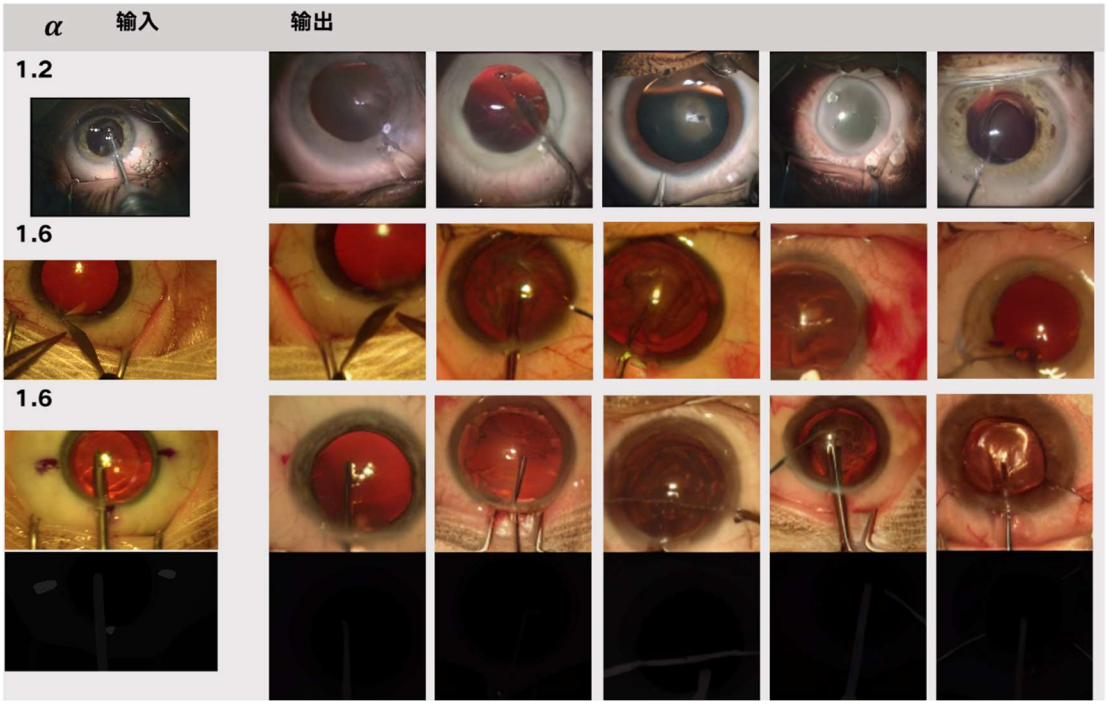

图 19 在训练生成模块后的差值实验

4.4 本章小结

本章阐述了本项目的实现细节以及实验结果，证明了提出方法的可行性和合理性。在 4.1 节中概括了实验使用数据集，描述了实验前对数据集的处理方式。4.2 中罗列了模型实现时对预训练模型的选择、超参数的设置和实验使用的硬件信息。在介绍了对比实验中使用的模型和评价指标的数学公式后，规划了主体研究的实验步骤。4.3 小节展示了对比实验结果和本文提出模型的在不同插值函数以及生成任务上的实验结果，可以看到本模型具备根据文本编辑图像，生成高质量的逼真图像和对应语义分割图像的数据对的能力。

5.结论

白内障手术的临床需求远超医疗资源所及，通过开发相关的医疗辅助诊断深度学习算法可以帮助医生对患者进行治疗，而基于医学图像训练的模型均受到数据集标注质量不高，数据量小的困扰。本项目基于扩散模型与插值函数的原理，构建了文本编辑医疗图像的算法框架，具备生成质量高、多样丰富的逼

真手术数据以及对应的语义分割图像，达到扩充数据集、平衡数据集的效果。实验将本模型与其他类型的生成模型进行了对比，并且证明了文字编辑图像在生成任务上的优势。通过使用不同的插值函数，模型能基于一对目标文本和图像生成风格不同，保真度高的医学数据，在病人数据的隐私保护和提高模型鲁棒性上具有重大意义。

同时工作在以下方面具有不足之处，由于对比方法的任务不同，难以选择合适的评价指标客观评价模型的性能，其次由于模型中使用的 text encoder 的局限性，模型对文本输入的语义信息捕获不准确，未来可以通过使用多角度的评估方式和对比不同的 text encoder 在编辑任务上的性能两个方面完善实验。

参考文献

[1] LIU Y C, WILKINS M, KIM T, et al. Cataracts[J/OL]. The Lancet, 2017, 390(10094): 600-612. DOI:10.1016/S0140-6736(17)30544-5.

[2] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J/OL]. Communications of the ACM, 2020, 63(11). DOI:10.1145/3422622.

[3] KINGMA D P, WELLING M. Auto-Encoding Variational Bayes[J]. 2013.

[4] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Advances in Neural Information Processing Systems: Vols. 2020-December. 2020.

[5] LARSEN A B L, SØNDERBY S K, LAROCHELLE H, et al. Autoencoding beyond pixels using a learned similarity metric[J]. 2015.

[6] MIRZA M, OSINDERO S. Conditional Generative Adversarial Nets[J]. 2014.

[7] NICHOL A, DHARIWAL P. Improved Denoising Diffusion Probabilistic Models[J]. 2021.

[8] SAHARIA C, CHAN W, CHANG H, et al. Palette: Image-to-Image Diffusion Models[C/OL]. 2022. DOI:10.1145/3528233.3530757.

[9] NICHOL A, DHARIWAL P, RAMESH A, et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models[J]. 2021.

[10] KAWAR B, ZADA S, LANG O, et al. Imagic: Text-Based Real Image Editing with Diffusion Models[J]. 2022.

[11] YANG L, ZHANG Z, SONG Y, et al. Diffusion Models: A Comprehensive Survey of Methods and Applications[J]. 2022.

[12] KONG Z, PING W. On Fast Sampling of Diffusion Probabilistic Models[J]. 2021.

[13] DHARIWAL P, NICHOL A. Diffusion Models Beat GANs on Image Synthesis[C]//Advances in Neural Information Processing Systems: Vol. 11. 2021.

[14] COUAIRON G, VERBEEK J, SCHWENK H, et al. DiffEdit: Diffusion-based semantic image editing with mask guidance[J]. 2022.

[15] LIU X, PARK D H, AZADI S, et al. More Control for Free! Image Synthesis with Semantic Diffusion Guidance[C/OL]//Proceedings - 2023 IEEE Winter Conference on

Applications of Computer Vision, WACV 2023. 2023. DOI:10.1109/WACV56688.2023.00037.

[16] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models From Natural Language Supervision[J]. 2021.

[17] REDDY M, BASHA M, CHINNAIAHGARI H. DALL-E: CREATING IMAGES FROM TEXT[J]. Dogo Rangsang Research Journal , 2021, 8(14).

[18] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical Text-Conditional Image Generation with CLIP Latents[J]. 2022.

[19] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-Resolution Image Synthesis with Latent Diffusion Models[J]. 2021.

[20] WU C H, DE LA TORRE F. Unifying Diffusion Models' Latent Space, with Applications to CycleDiffusion and Guidance[J]. 2022.

[21] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding[J]. 2022.

[22] HO J, SALIMANS T. Classifier-Free Diffusion Guidance[J]. 2022.

[23] GRAMMATIKOPOULOU M, FLOUTY E, KADKHODAMOHAMMADI A, et al. CaDIS: Cataract Dataset for Image Segmentation[J]. 2019.

[24] FOX M, TASCHWER M, SCHOEFFMANN K. Pixel-Based Tool Segmentation in Cataract Surgery Videos with Mask R-CNN[C/OL]//2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2020: 565-568. DOI:10.1109/CBMS49503.2020.00112.

[25] KUMAR A, SATTIGERI P, BALAKRISHNAN A. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations[J]. 2017.

[26] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-Image Translation with Conditional Adversarial Networks[J]. 2016.

[27] ZHU J Y, PARK T, ISOLA P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks[J]. 2017.

[28] SAHARIA C, HO J, CHAN W, et al. Image Super-Resolution Via Iterative Refinement[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

DOI:10.1109/TPAMI.2022.3204461.