南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

本科生毕业设计（论文）

题　　目：基于语义标签图的手术场景图像生成

姓　　名：马子晗

学　　号：11912732

系　　别：计算机科学与工程系

专　　业：计算机科学与技术

指导教师：刘 江

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

基于语义标签图的手术场景图像生成

马子晗

（计算机科学与工程系　指导教师：刘江）

[摘要]：随着医疗人工智能的发展，医疗数据集的缺乏成为制约医疗算法模型的一个重要因素。目前，图像生成领域发展迅速，将其成果引入医学图像生成将有利于推动医学人工智能的发展。通过对现有的图像生成技术进行调研，扩散模型是目前生成图片像素高、质量好的新兴算法，将其与医学图像生成结合，可以帮助解决部分数据集短缺的问题。为了帮助解决这一问题，我们根据语义生成的特点并基于现有的扩散模型技术，设计了模型，可以依据语义分割图来生成大量多风格图片用于医学人工智能研究。模型参考了 SR3 的模型结构，并借鉴了空间自适应归一化来进行语义信息的保留，同时进行了加速采样的设置，提高了模型的利用效率。通过对比，发现我们的模型在图像生成方面有更良好的表现，同时在较小的数据集上的表现效果更加出色。

[关键词]：计算机视觉；扩散模型；语义图像生成

[ABSTRACT]:With the development of medical artificial intelligence, the lack of medical datasets has become an important factor constraining medical algorithm models. At present, the field of image generation is developing rapidly, and introducing its achievements into medical image generation will be beneficial for promoting the development of medical artificial intelligence. By conducting research on existing image generation technologies, diffusion model is an emerging algorithm with goodperformance for image generation. Combining it with medical image generation can help solve the problem of dataset shortage. To help solve this problem, we designe a model based on the characteristics of semantic generation and existing diffusion model techniques, which can generate a large number of multi style images for medical artificial intelligence research based on semantic segmentation images. The model refers to the model structure of SR3, and uses spatial adaptive normalization for Semantic information retention. At the same time, accelerated sampling is set to improve the utilization efficiency of the model. Through comparison, it was found that our model performs better in image generation and performs better on smaller datasets.

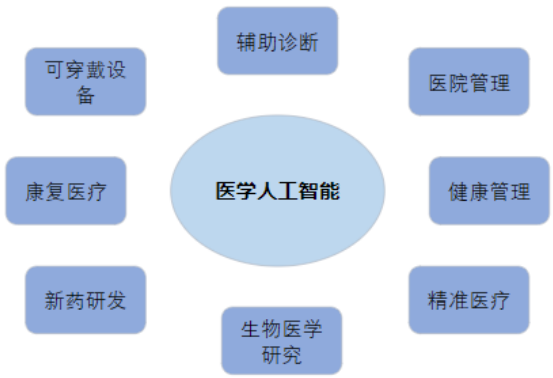
[Keywords]: Computer vision; Diffusion model; Semantic Image Generation

目录

1. 引言

1.1 研究背景与意义

近年来，国家连续发文，多次提到推广医学人工智能技术。我国人口基数大，医疗资源有限且分布不均，很多地方的医疗资源严重不足。推广医疗人工智能产业，有助于在一定程度上减轻提高就医压力，提升医疗服务水平，改善医疗发展不充分、不均衡的现状。[1]人工智能技术可以在很多方面和医学进行产业结合，比如现在很常见的健康管理平台、可穿戴医疗设备等，除此之外，辅助诊断也成为现在的研究热点。例如，目前有许多基于白内障手术数据的语义分割研究[2]，可以在术中辅助医生进行一些器材、部位等的识别工作；还有颅脑血管分割等，可以帮助医生定量分析病情，辅助医生的诊断。目前，随着众多人工智能研究成果的开发，很多医疗人工智能也开始进入大规模研究。[3]这些研究的发展无疑会给医疗工作人员带来极大的便利。

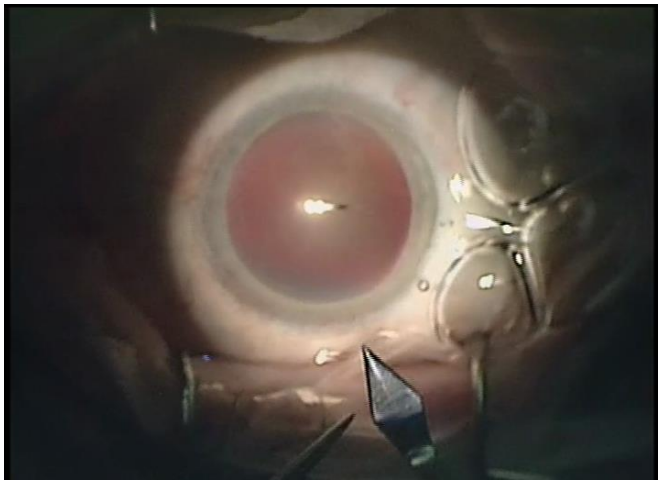


图表 1 人工智能技术与医学结合领域

然而，对于大部分医疗人工智能研究来说，它们需要大量的、有良好注释的数据集。有些模型训练甚至需要像素级的注释。以辅助诊断为例，大部分算法都是基于计算机视觉的研究，例如语义分割、图像识别等。这些算法都要以大量数据集作为基础进行训练。目前，可用于研究的医学数据集数量不足，而医学图像的注释需要大量的人力和物力。这限制了医学人工智能的进一步发展。由于医学图片数据采集的难度和标注的复杂性，目前为训练医学人工智能算法提供的数据集通常数量较少，缺乏标注，且样式有限。标注质量差或数据集较小会极大地影响算法的训练，而风格有限的数据集会导致算法的泛化性能低。随着图像生成领域的发展，目前有越来越多的医学图像生成算法的相关研究，这有利于提升医学数据集的风格多样性和数量。现有的图像生

成算法[4]大多是基于自然场景或人脸的，对于医学领域的应用场景来说，应用场景规模小，精度要求高，所以现有的方法不能直接移植使用。

因此，有必要设计一种能够生成多数量、多风格眼科图像的算法，帮助解决医学数据集不足的问题。针对于不同的医学图片，它们的类型特点有一定差异。医学图像中，常见的有超声图像、磁共振图像、手术场景图像等。其中，手术场景图像相对结构清晰，和普通图片差异性更小。而且，它应用场景也比较广泛，在手术辅助诊断中具有重要价值。因此，合成手术场景图像是可行性较高且有实际价值的一项工作。通过对手术场景图像的生成，将缓解目前有关工作数据集缺乏的问题。



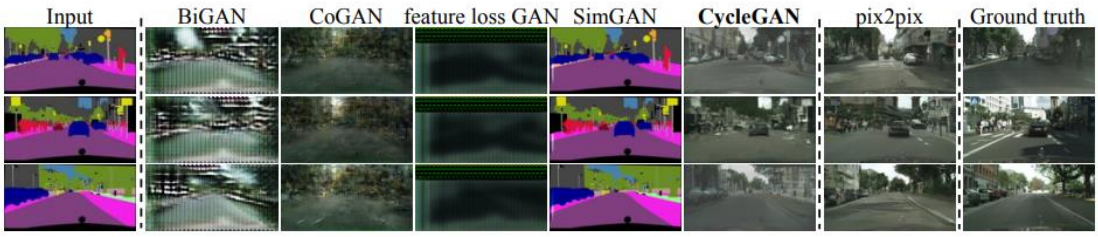
图表 2 眼科手术图像

1.2 语义图像合成

语义图像合成是指基于语义分割图来生成真实图片，是图像生成领域的一个子领域。通过语义合成图片，可以生成大量的不同风格的数据集，并且能直接提供标注好的结果。它是图像语义分割的反过程，不同的是，语义分割由真实图片分割出来的结果是一一对应的，而语义图像合成的结果只要是合理的就可以了，也就是说可以有多样的结果。

语义合成在许多方面都有重要的价值。在医学图像领域，依据语义分割图生成的图像可以直接拥有对应的语义分割标注。而在图像编辑领域，语义合成是进行编辑的基础。通过对语义分割图的简单调整，可以生成多种效果逼真的图片。

目前语义图像合成领域的经典方法有：CRN[5], pix2pixHD[6], SIMS[7], SPADE[8]。近年来，主流方法都是基于 GAN（Generative adversarial network）的。目前，有许多研究关注于这一问题，尝试生成高质量图片。随着扩散模型的发展，如今也有一些基于扩散模型的语义图像合成的研究[9]。



图表 3 语义合成图片示例[10]

1.3 手术场景图像生成任务的挑战

目前，已经有许多图像生成算法被提出，他们在街景、人脸等图像上表现出了良好的性能。在医学图像方面，也开始有研究者对这一领域进行探究。但是，在医学图像的语义生成方面，目前仍有许多问题亟待解决，具体如下：

1. 精度问题：相比于常规的图像生成问题，医学图像生成需要更高精度的图片。以白内障手术场景为例，图像中存在各类形态相似但有细微差异的手术器械，并且有部分眼部构造，这需要高精度的生成模型来生成与其符合的手术场景图像。

2. 数据问题：有标注的医学图像的数量相对于其他类别来说，数量较少且不同域的数据存在一定差异。具体体现在拍摄时的视场、图像质量等特征均等的差异，这需要模型对数据有一定泛化能力，且在小数据集也可以展示良好的性能。

3. 语义信息问题：大部分基于深度学习的方法经常将语义图直接送入神经网络进行学习。这些方法往往简单且有一定的效果，但是对于一些复杂场景，还是有其局限性存在：大部分普通神经网络中的归一化层倾向于“洗去”语义信息。

1.4 实验的创新点

在本研究中，我们提出了一种新型的基于扩散模型的医学图像生成算法。目前，基于对抗模型的图像生成模型仍是主流，但本实验采用了目前新兴的扩散模型作为基础。扩散模型是目前非常流行的 AI 图像绘画的基础，它的训练稳定而且生成图像清晰。从实验的模型整体上看，创新性有如下体现：

- 借鉴 SR3 模型[11]，修改 U-net，通过对残差块的修改和增加，提升了生成图片清晰度，使得模型相对于传统扩散模型近一步增强了清晰度。此外，通过对 U-net 输入的修改，引入了条件图像生成。
- 加入空间自适应归一化技术[8]，通过加入空间自适应归一化模块在 U-net 的解码器部位，将语义信息注射到解码器部位以用来引导图片的去噪过程。这可以增强模型的对语义信息的保留能力。

- 参考了 DDIM 模型[12]的采样加速原理，设置加速采样模块，并且通过实验挑选了合适的加速采样步数。通过该模块，可加速模型采样，增强模型的利用效率。

通过我们的实验，探索了语义生成医学图片的可能性，拓宽了扩散模型的应用领域。

2. 相关研究

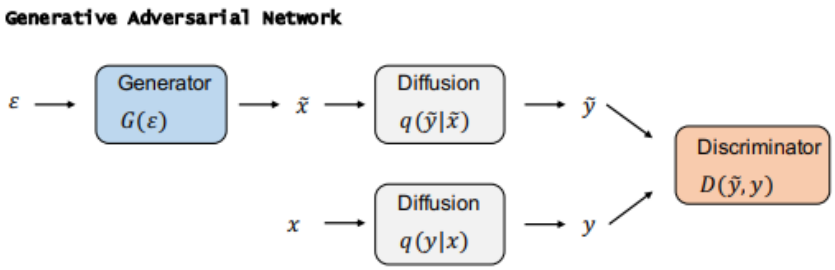
如今，许多图像生成算法研究已经取得了众多成果，主要的生成模型种类有基于生成对抗网络的生成模型，基于变分自编码器的生成模型，和基于扩散模型的生成模型。但它们在研究过程中各有其优势也有其不足之处。基于生成对抗网络的生成模型的训练过程不稳定，容易造成模式坍塌。基于变分自编码器的生成模型生成的图像相对比较模糊，而基于扩散模型的生成模型采样速度较慢。除此之外，对于语义图像合成，目前也有部分学者开始进行研究。本章中将详细介绍这些研究。

2.1 基于对抗学习的图像生成算法

生成对抗网络（GAN）是一种算法体系结构，它使用两个神经网络，一个生成网络---用来生成所需数据，一个判别网络---用来判断数据是真实的还是生成的。目前，大部分主流图像生成算法都是基于对抗学习的。[4]

原始的 GAN 网络虽然在 2014 年才首次提出，但其扩展速度非常迅速，产生了众多衍生网络，如：DCGAN[13]、SGAN[14]。它依靠一种博弈过程来完成数据的生成。生成对抗网络的博弈过程，就是使用生成网络制造的数据分布来拟合真实的数据分布和判别网络对其进行真假判断的过程。当设置为生成网络时，接收一个随机的噪声，生成图片并输出。设置为判别网络时，输入一张图片，判别器可以计算出该图为生成的或真实图片的概率。两者分别根据返回的结果反向更新网络，相互抗衡，动态变化最后达到一定的平衡。

在有条件的图像生成方面，常用的图像生成算法有 pix2pix[15]、CycleGAN[10]、SPADE 等。它们在街景、人脸等图像生成方面都有比较良好的表现。但生成对抗网络也有其弊端，它的训练相对不稳定，容易出现模式坍塌问题，可控性也相对较差。



图表 4 基于对抗学习的图像生成算法流程[16]

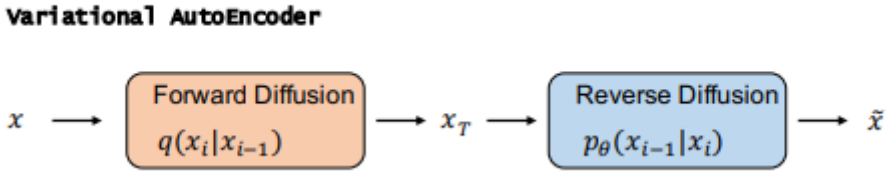
2.2 基于自动编码器的图像生成算法

自动编码器是一个简单的神经网络，其输出就是输入。它们的目标是学习如何来重构输入数据。网络的第一部分是编码器，用来输入，并将其编码在一个低维的潜在空间中。第二部分是解码器，用来接收该矢量并对其进行解码，以产生原始输入。[17]它是深度学习领域中很常用的一种无监督学习方法，它的基本思想是通过将输入数据压缩到低维表示的形式，然后将其解压缩回原始空间，通过这样来实现对数据的重构。自编码器的训练过程可以通过最小化重构误差来完成。

目前常用的自动编码器是变分自动编码器（VAE）。它对于自动编码器来说，增加了一定的随机性。它继承了传统自编码器的结构，它主要的变动是对编码的生成上。编码不再像自动编码器中是唯一映射的，而是具有某种分布，使得编码在某范围内波动时都可产生对应输出。VAE 采用了概率编码和解码的方式，并引入 KL 散度来强制潜在表示服从预先定义的高斯分布。它的优点是可以从潜在空间中采样生成新的数据，并且可以进行无监督学习。VAE 在图像生成、文本生成、图像压缩等领域都有广泛的应用。

VAE 的核心思想是通过学习数据的潜在分布来实现数据生成。具体来说，它假设原始数据是由一个潜在变量 z 和一个条件分布 $P_\theta(x|z)$生成的，其中 θ 是关于模型的参数。模型的目的是通过学习得到一个编码器 $Q_\phi(z|x)$和一个解码器 $P_\theta(x|z)$，使得从 x 到 z 的映射是可逆的，并且可以通过从潜在空间 z 中采样生成新的数据。

VAE 的结构相对简单，训练也非常快速。目前，基于 VAE 已经有非常多的衍生算法模型，但它们的缺点是目前相关的算法生成图像都相对比较模糊，清晰度较低，不适合直接用于高清图像的生成。



图表 5 基于变分自动编码器的图像生成算法流程[16]

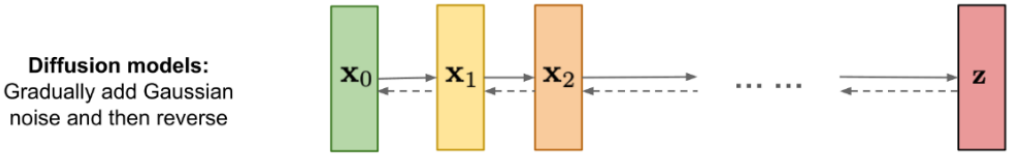
2.3 基于扩散模型的图像生成算法

扩散模型是一类生成模型，其灵感来自非平衡热力学。和变分自动编码器（VAE），生成对抗网络（GAN）等生成网络不同的是，扩散模型在前向阶段对图像逐步施加噪声，直至图像被破坏变成完全的高斯噪声，然后在逆向阶段学习从高斯噪声还原为原始图像的过程。[18]

扩散模型早在 2015 年的 Deep Unsupervised Learning using Nonequilibrium Thermodynamic 文章[19]中提出，但直到 2020 年的 DDPM[18]的提出，扩散模型的应用价值才逐渐显现。它的基本原理是来自物理学的非平衡热力学。在物理学中，气体分子从高浓度区域扩散到低浓度区域，这与由于噪声的干扰而导致的信息丢失是相似的。所以扩散模型通过引入噪声，然后尝试通过去噪来生成图像。在一段时间内通过多次迭代，模型每次在给定一些噪声输入的情况下学习生成新图像。具体来说，这是基于马尔科夫链，如图所示。从 x0 到 x_T 的过程是扩散的前向过程，图像慢慢扩散至随机高斯噪声，其中每一步的随机过程为 $q(x_t|x_{t-1})$，这个过程由我们自己定义，是已知的。从 x_T 到 x_0 的过程是扩散的逆向过程，图像慢慢从高斯噪声逆熵为正常图像，其中每一步的随机过程为 $q(x_{t-1}|x_t)$，该过程是未知的过程。扩散模型要做的，就是定义一个随机过程为 $p_\theta(x_{t-1}|x_t)$的逆向过程，并优化参数 θ，使得该过程尽可能接近真实逆向过程 $q(x_t|x_{t-1})$，从而使得我们最终能通过一个随机高斯噪声 x_T 生成一个正常的图 x_0。

扩散模型训练过程稳定且生成的图像清晰，是目前新兴的图像生成技术。但它的缺点是采样速度较慢，图像生成效率低。目前，有关于采样加速的相关研究，如 DDIM、dpm-solver[20]等，它们在采样加速上都有一定的良好表现。其中 DDIM 是影响力比较大的加速采样模型，其主要原理是放弃采样的马尔科夫链，设其为一个固定过程，从而利用公式计算去噪过程，可以省去依次计算的过程。简单说就是从一个噪声大的图像，直接跳步恢复到一个噪声小的图像。在街景、人脸等图片合成中，它都表现了良好的效果。

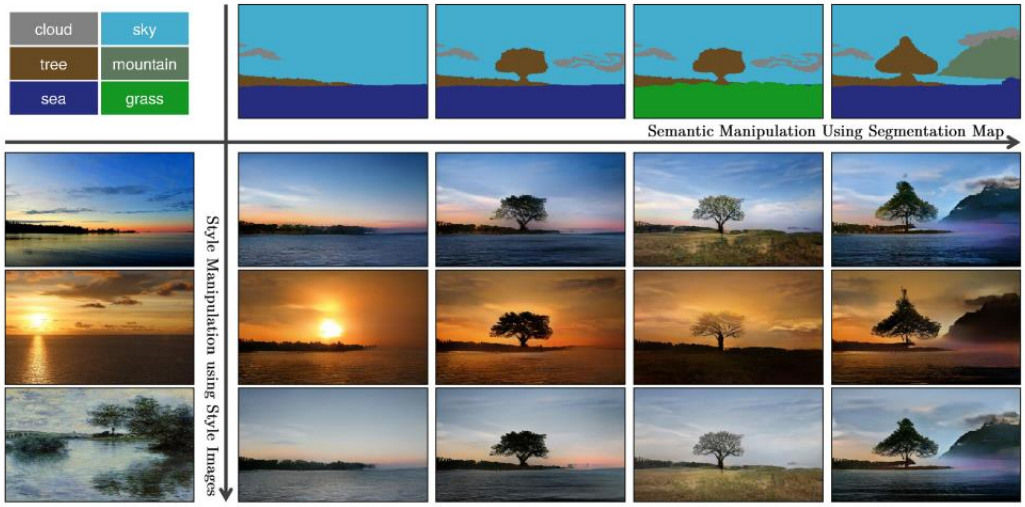
扩散模型本身是相对简单的一个架构，但对其改进，可以有许多不同的用途，最近大火的 AI 图像合成就是在此基础上进行的加工改造。对扩散模型进行近一步融合架构，有巨大的研究前景。



图表 6 基于扩散模型的图像生成算法流程

2.4 语义图像合成

语义图像合成是依据语义信息来合成图片，可以看作是语义分割的逆过程。不过语义合成具有更多的多样性。该问题有各种方面的广泛应用，例如图像编辑、交互式绘画和内容生成。

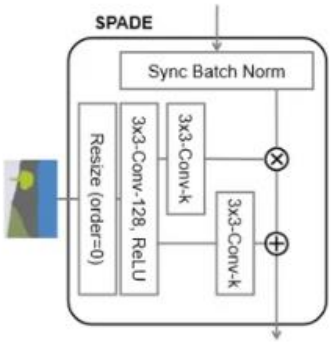


图表 7 语义合成示例[8]

目前的关于语义合成的方法大多是基于对抗学习的。在大多数方法中，都是将语义信息作为条件与其它信息共同输入到网络中，进行网络的训练。例如 Pix2PixHD[6] 利用了多尺度生成器从语义标签图生成高分辨率图像。但现在，也有一些将语义信息注射到网络中来引导语义合成的。例如，SPADE 提出了空间自适应归一化，以更好地将语义布局嵌入到生成器中。CLADE[21]通过提出的一个新的类自适应归一化层。SCGAN[22]引入了动态加权网络来加强语义相关性、结构和细节合成。

SPADE 是语义合成中比较重要的一个模型。在 SPADE 中，它提到了空间自适应归一化层，将语义信息注射到网络中，用来引导图像合成。其核心是其 SPADE 模块。在该模块中，主要思想就是利用语义图信息来指导特征图进行归一化。不仅可以保留

住归一化的功能，还可以更好地保留住初始的语义信息。在相关的语义合成模型中，SPADE 表现良好。最近，也有语义合成在扩散模型上的尝试[9]。



图表 8 SPADE 模块[8]

2.5 本章小结

针对于不同的图像生成技术进行分析，可以发现 GAN 和扩散模型是比较适合于语义图像合成这一问题的。但扩散模型在稳定性和清晰度上有更好的表现，因此本实验将采用基于扩散模型的图像生成技术来完成从语义分割图到手术场景图的生成。针对于扩散模型采样较慢的缺点，可依据 DDIM 的采样原理对其进行加速。同时针对于语义合成这一条件，调查发现 SPADE 模型的归一化模块有较好表现，因此计划采用空间自适应归一化来控制图像的生成。

3. 方法

3.1 基于扩散模型的主体模型

扩散模型在图像生成领域发挥了越来越重要的作用。2020 年的 DDPM 的提出，扩散模型的应用价值逐渐显现。目前的大多数扩散模型都是在 DDPM 上进行的改进。

DDPM 的工作原理是对已有的数据连续加入高斯噪声让数据变为纯噪声，再反转该过程，即去噪，来完成图片的生成。模型的训练也就集中在加噪和去噪这两过程中。训练目的是通过学习加噪过程中对噪声的估计来完成去噪过程中对图片的重建。具体过程如下：

1）前向过程

前向过程即加噪过程，又称扩散过程。设原始图片为 x_0，逐步加噪声变为 x_t 用公式表示，为

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} \quad (1)$$

其中，$\{\alpha_t\}_{t=1}^{T}$ 是关于噪声的参数，通常数值较小，$\epsilon_{t-1} \sim N(0,1)$ 是不断加入的高斯噪声。在实际应用中，还有参数 β，α_t=1-β_t。β 是逐渐变大的。

整个扩散过程是一个马尔科夫链，可表示为：

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \quad (2)$$

由（1）推导可得：

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon \quad q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\overline{\alpha}_t}\mathbf{x}_0, (1-\overline{\alpha}_t)\mathbf{I}) \quad (3)$$

其中，$\overline{\alpha_t} = \prod_{i}^{t} \alpha_i$，$\epsilon \sim N(0,1)$。

2）反向过程

反向过程即去噪过程。该过程使得数据从 x_t 恢复到 x_0。可以通过知道每一步的 $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 来逐步去噪生成原数据。

公式表达为：

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \quad (4)$$

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (5)$$

加上条件 x_0 后，计算可得：

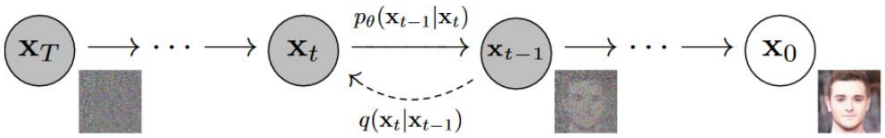
$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}) \quad (6)$$

通过近一步优化计算可得：

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0), \sigma_t^2 \mathbf{I}) \, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}) \quad (7)$$

通过对公式（7）的运用，可以逐步去噪获得原图像。

本实验的架构就是基于扩散模型的。扩散模型生成图片稳定且清晰度高，是目前图像生成领域的重要组成部分。



图表 9 扩散模型加噪去噪过程[18]

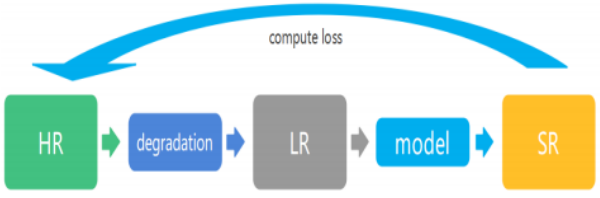
3.2 基于 SR3 超分扩散模型的模型改进

本实验借鉴了 SR3（Image Super-Resolution via Iterative Refinement，一种超分扩散模型）来实现从语义分割图到手术场景图的生成。SR3 是一种超分辨率扩散模型，它以低分辨率图像作为输入，然后从纯噪声中构建相应的高分辨率图像。它对扩散模型的改进主要基于其 U-net 模块和噪声的生成方式。

首先，在 U-net 的改进方面，SR3 将原先 DDPM 中的残差块替换为 BigGAN 模型[23]的残差块，并将连接缩放为$\frac{1}{\sqrt{2}}$。并且将残差块的数量从 2 个增加到 3 个，残差块数量的增加有利于分辨率的提升。其次，在输入方面，SR3 将低分辨率图片三次插值上采样到高分辨图片的分辨率后，与随机采样的噪声图（x_t）进行拼接，将此作为条件输入到 U-net 中进行噪声的预测。得到的噪声预测可以对噪声图进行去噪，然后再继续与上采用的低分辨率图进行拼接，输入到 U-net 中预测噪声，循环该步骤来进行去噪。

在噪声生成方面，原扩散模型（DDPM）中和噪声生成有关的参数α_t是基于采样的 t 得到的，每步都需要重新采样得到。而 SR3 则将其改为初次采样 t 后，设置好相应的初始值，然后其余的α_t经过α_t和α_{t-1}之间均匀选取的，同时不再直接输入给 U-net 直接的数值 t，而是输入参数α_t。

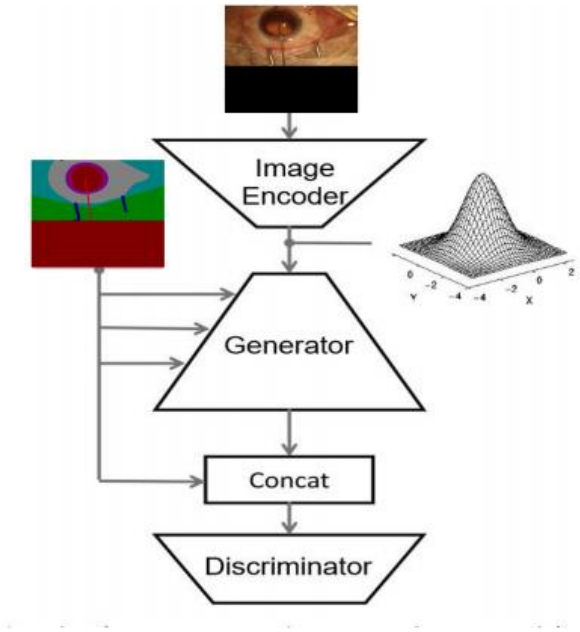
在本实验中，借鉴 SR3 有条件的生成的处理方法，将语义分割图作为条件输入 U-net 中进行噪声的预测。同时，学习 SR3 中提升分辨率的方法，增加残差块数量，修改残差块类型，使得模型的能生成图片分辨率更高。



图表 10 SR3 整体架构

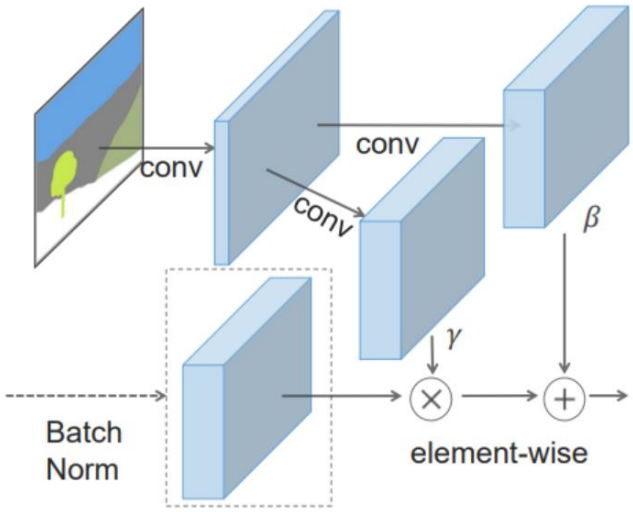
3.3 基于 SPADE 的归一化处理

在语义图片合成中，很重要的一部分是对于语义信息的保存。在 SPADE 模型[8]中，它将语义信息提取处理，控制了图片的合成。模型的核心要素在于其归一化的设置。它设置了一个条件归一化层，然后通过空间自适应和学习转换等一系列步骤，来调节输入语义布局的激活，这一能使得语义信息更好的网络中得以传播。



图表 11 SPADE 整体架构

SPADE 归一化设置的原理就是利用卷积学习语义分割图的 γ 和 β，然后将其作为归一化的参数，与图片的均值和方差进入归一化的相关计算当中。这样可以保证语义信息能更好的被保留下来。



图表 12 SPADE 归一化层设置[8]

在本实验中，为了让语义分割图的信息能更好的运用于图像的合成过程，实验借鉴了 SPADE 的归一化模块，将其运用在 U-net 结构上采样的归一化过程中，选取了特定的 Resblock 加入了空间自适应归一化模块，从而提升模型对于语义信息的利用。

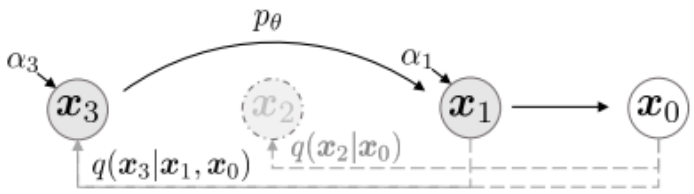
3.4 基于 DDIM 的加速采样

扩散模型有一个缺点，就是它的采样过程相对缓慢。[18]因为相比于 GAN 等其他深度学习模型，扩散模型每生成一张图片需要更多的采样步数。每一步采样就是一次推理过程。扩散模型有两个过程，前向加噪过程和逆向去噪过程。DDIM[12]是对逆向去噪过程进行优化改进从而加速采样的。DDPM 中，前向与逆向过程的步数都是相同的，但其实逆向的步数是可以进行缩减的，这样就可以大大提升采样速度，而且不影响图片的采样质量。

DDIM 中，它不再限制扩散过程必须是一个马尔科夫链，这使得采样过程不必满足逐步扩散这一限制，同时生成样本的过程也变成了确定的，不再逐步增加随机噪声。它的主要思路可以概括为在满足原有的扩散模型的逆向推理的条件下，找到新的用 x_t 和 x_0 表达 x_{t-1} 的一种方式，并且这种方式可以简化计算。

原有的扩散模型的去噪过程是马尔科夫链分布，如果想要简化过程，就需要找到一个非马尔科夫链分布的去噪过程。在 DDPM 中，损失函数并没有出现 $p(x_t|x_{t-1})$，更需要的是 $p(x_t|x_0)$，因此，DDIM 尝试从 $p(x_t|x_0)$ 出发，得到可以符合要求的用 x_t 和 x_0 表达 x_{t-1} 的一种方式。

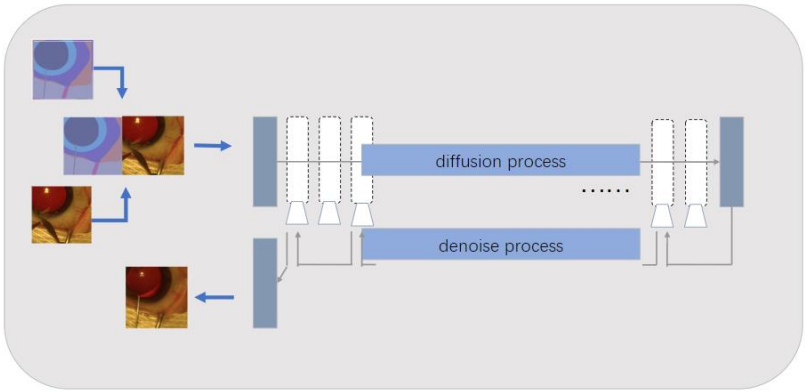
首先定义 $p(x_t|\ x_0)=N(\sqrt{\bar{\alpha}_t}x_0,1-\bar{\alpha}_t)$，再依旧假定 $p(x_{t-1}|\ x_t,x_0)$ 为高斯分布，可得 $p(x_{t-1}|\ x_t,x_0)=N\left(\sqrt{\bar{\alpha}_{t-1}}x_0+\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\frac{x_t-\sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}},\sigma_t^2\right)$。若令 $\sigma_t^2=0$，则该过程变成一个确定过程，这就成为了一个确定的隐变量模型，即 DDIM。



图表 13 DDIM 的加速过程[12]

DDIM 由于在后续采样过程中不再随机增加噪声，所以它的结果与一开始加入的随机噪声有关。另外，由于每步去噪过程的影响程度不同，在 DDIM 模型中，β 的设置由线性改变换成 cosine 的改变方式会有更好的表现。同时在实践中发现，由 DDPM 训练的模型可以不需要修改，直接利用 DDIM 模型进行采样。因此利用 DDIM 的原理可以较好的对采样进行加速。与其他模型对比时，DDIM 也有较好的表现[20]。

在本实验中，采样速度的降低会影响实验的效率，影响模型的可用性。通过引入 DDIM 中的加速机制，大大缩短了图片生成所需的时间，增强了模型的效率。



图表 14 模型整体架构

4. 实验

4.1 数据集情况

本实验利用 CaDIS 数据集[24]。CaDIS 数据集是一个以白内障手术视频数据集分割而成的手术图片数据集，可用于分割模型训练，它从 25 个视频中采样而来，每组视频的风格都有一定差异，其共有图片 4023 张（训练集 3490 张，测试集 533 张）每

个图像中的每个像素都用 36 个已识别类别中的相应仪器或解剖类别进行标记。

图表 15 数据集图片示例

4.2 实验细节

4.2.1 软硬件环境

实验的服务器的软硬件平台配置如下：操作系统为 Windows 10，CPU 为 Intel i9-10900，内存大小为 64GB，GPU 为 NVIDIA GeForce RTX 3060（显存大小 12GB），代码使用 Pytorch 框架。

4.2.2 对比方法介绍

实验选取了 pix2pix，cycleGAN，SPADE 作为对比方法进行实验。它们都是目前常见的可以根据语义分割图生成图像的模型。

pix2pix 是首个提出用 GAN 来解决图像生成问题的通用框架的模型，并证明了其有效性。它可以输入成对的数据（语义分割图和真实图片）进入模型中进行训练，从而生成图片。CyclgeGAN 引入了循环一致损失，让输入不再局限于成对的照片。SPADE 引入空间自适应归一化设置，让语义信息得到更好的保留。

4.2.3 实验任务介绍

实验需对比不同方法生成的图片的质量。实验包括了两部分

第一，是尝试不同的采样步数，查看生成的图片的质量。此部分利用 CADIS 训练集中的 533 张图片作为模型的训练集，利用 CADIS 测试集作为模型的测试集。

第二，对比在训练集上进行训练时，各方法的图像生成能力。

4.2.4 评估指标

实验选取了 PSNR，SSIM，FID 作为评价指标。

PSNR 是峰值信噪比，数值在 0-50db 之间，越大表示两张图片相似度越高，是判断两个图片的相似度的。其计算公式是：

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[I(i,j)-K(i,j)\right]^2 \quad (1)$$

$$PSNR = 10 \cdot log_{10}(\frac{MAX_I^2}{MSE}) \tag{2}$$

SSIM 是基于两个输入之间的三个值进行比较：亮度（l）、对比度（c）和结构（s）。

设两个输入分别为 x，y，其计算公式为：

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{3}$$

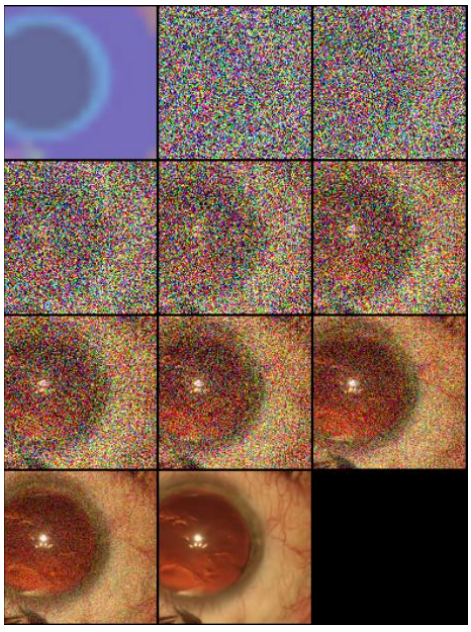
$$SSIM(x,y) = [l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma}] \tag{4}$$

FID 是基于生成数据和真实数据在特征层次的距离来进行对比，它需要神经网络来提取图片的抽象特征。其公式如下：

$$FID = \| \mu_r - \mu_g \|^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{5}$$

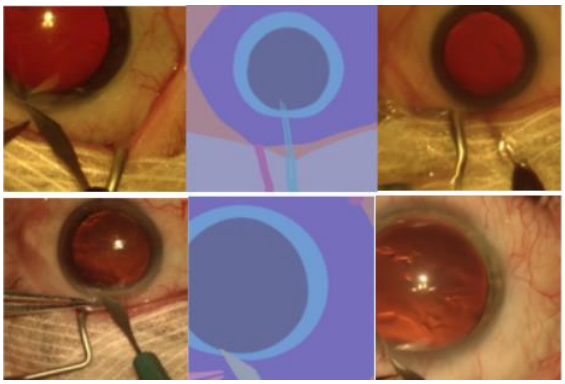
4.3 实验结果

通过本实验模型，可以高效的生成较为清晰明确的手术场景图。生成过程如下图所示。



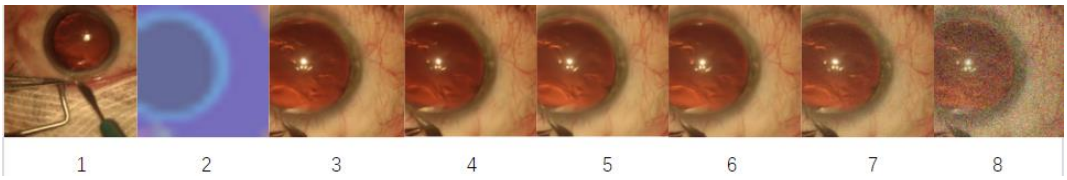
图表 16 采样生成图片过程

选取了训练集中风格不同的几组数据（数据集大小约为 500 左右）进行训练，模型可以较好的学习到图片的风格。



图表 17 生成的图片示例（从左到右为风格图，语义分割图，生成图）

在加速采样实验中，共利用模型进行了 6 组实验，采样步数分别设置为 2000，80，40，20，10，4。取训练集中的 533 张作为采样实验的训练集进行了实验。生成图像结果可见下图，可知在采样步数为 40 时，图片质量仍然保持较为良好的水平。



图表 18 不同采样步数下生成的图片（第 1 张为风格图片，第 2 张图片为语义分割图，第 3-8 张图为采样步数在 2000，80，40，20，10，4 下生成的图片）

对其用 PSNR 和 SSIM 指标进行数据对比，结果如下：

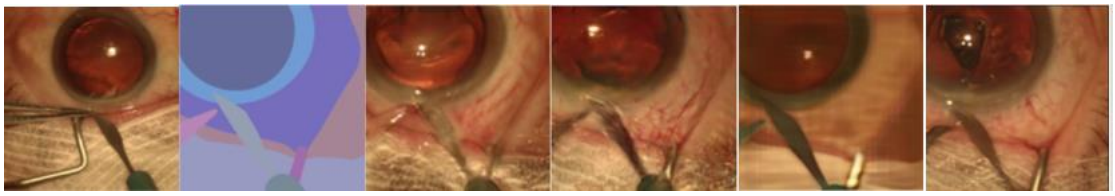
表 1　不同采样步数的评价指标表

采样步数	PSNR↑	SSIM↑
2000	13.722	0.281
80	13.726	0.278
40	13.834	0.282
20	14.046	0.278
10	14.219	0.229
4	13.901	0.093

通过对数据的对比和图片的观察可得，采样步数降低至 40 次时，图片的质量仍然与原先的图片质量相差不大，后续采用采样步数为 40 的设置进行实验。

确定好采样步数后，又扩大训练集数量，进行了与其它对比方法的对比。

对比试验测评结果：



图表 19 ClycleGAN,Pix2pix,SPADE 和我们的方法的结果对比

对其进行 FID, PSNR, SSIM 指标的评估：

表 2　评价指标结果对比表

方法	FID↓	PSNR↑	SSIM↑
CyleGAN	136.3	9.516	0.544
Pix2pix	131.3	17.263	**0.611**
SPADE	335.6	8.324	0.245
Our method	**123.1**	**18.572**	0.458

由于图片生成具有一定多样性，所以评估图像生成质量需要数据和图片的共同比对，通过实验数据和图像可以看出，我们的实验模型有相对更好的表现。

4.4 本章总结

通过实验比对可以发现，我们基于扩散模型改进的模型在语义图像生成上有相对良好的表现性能，并且在较小的训练集上也能有比较良好的效果。实验目前也有一定不足，在图片评估方面，目前的指标难以直接评估语义信息的留存程度，需要近一步的分割实验才能有更明确的数据指标。后续可在此方面进行近一步探索。

5. 结论

医学图像生成对于医学算法的研究有着重要的帮助作用。由于医学图像精度高、结构特殊，这一类任务通常认为是一项相对复杂的任务。在本次研究中，我们提出了利用扩散模型进行医学手术场景图像生成的算法。我们结合了 SR3 模型中，关于提升分辨率的 U-net 模块，增强了模型生成图片的分辨率；引入空间自适应归一化模块，增强语义信息对图像生成的引导；参考 DDIM 的原理，加快了模型的采样效率。我

们的研究可以帮助许多其它医学人工智能模型提供数据支持。同时语义合成在扩散模型的探索也可以帮助图像编辑领域的发展，探索了医学图像生成的更多可能性。

除此之外，我们的实验也仍有需要改进的部分。对比方法相对较少，可以增加相关语义合成算法的对比试验；数据集类别较少，可以寻找其它手术场景图片进行测试；评估指标可以增加，可以利用生成图片进行语义分割算法的训练，对比其效果。在未来的工作中，这些方面可以加以改进。

参考文献

[1] 袁天蔚, 薛淮, 杨靖, et al. 从战略规划与科技布局看国内外人工智能医学应用的发展现状 [J]. 生命科学, 2022, 34(08): 974-82.

[2] Yang Y, Soatto S. Fda: Fourier domain adaptation for semantic segmentation; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2020 [C].

[3] Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow [J]. Frontiers in medicine, 2020, 7: 27.

[4] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview [J]. IEEE signal processing magazine, 2018, 35(1): 53-65.

[5] Chen Q, Koltun V. Photographic image synthesis with cascaded refinement networks; proceedings of the Proceedings of the IEEE international conference on computer vision, F, 2017 [C].

[6] Wang T-C, Liu M-Y, Zhu J-Y, et al. High-resolution image synthesis and semantic manipulation with conditional gans; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018 [C].

[7] Qi X, Chen Q, Jia J, et al. Semi-parametric image synthesis; proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, F, 2018 [C].

[8] Park T, Liu M-Y, Wang T-C, et al. Semantic image synthesis with spatially-adaptive normalization; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2019 [C].

[9] Wang W, Bao J, Zhou W, et al. Semantic image synthesis via diffusion models [J]. arXiv preprint arXiv:220700050, 2022.

[10] Zhu J-Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks; proceedings of the Proceedings of the IEEE international conference on computer vision, F, 2017 [C].

[11] Saharia C, Ho J, Chan W, et al. Image super-resolution via iterative refinement [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[12] Song J, Meng C, Ermon S. Denoising diffusion implicit models [J]. arXiv preprint arXiv:201002502, 2020.

[13] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [J]. arXiv preprint arXiv:151106434, 2015.

[14] Huang X, Li Y, Poursaeed O, et al. Stacked generative adversarial networks; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017 [C].

[15] Isola P, Zhu J-Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017 [C].

[16] Yang L, Zhang Z, Song Y, et al. Diffusion models: A comprehensive survey of methods and applications [J]. arXiv preprint arXiv:220900796, 2022.

[17] Bowman S R, Vilnis L, Vinyals O, et al. Generating sentences from a continuous space [J]. arXiv preprint arXiv:151106349, 2015.

[18] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-51.

[19] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using

nonequilibrium thermodynamics; proceedings of the International Conference on Machine Learning, F, 2015 [C]. PMLR.

[20] Lu C, Zhou Y, Bao F, et al. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps [J]. arXiv preprint arXiv:220600927, 2022.

[21] Tang H, Qi X, Sun G, et al. Edge Guided GANs with Contrastive Learning for Semantic Image Synthesis [J]. arXiv preprint arXiv:200313898, 2020.

[22] Yu F, Koltun V, Funkhouser T. Dilated residual networks; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017 [C].

[23] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis [J]. arXiv preprint arXiv:180911096, 2018.

[24] Grammatikopoulou M, Flouty E, Kadkhodamohammadi A, et al. Cadis: Cataract dataset for image segmentation [J]. arXiv preprint arXiv:190611586, 2019.

致谢

感谢课题组的老师给予的指导，让我对计算机视觉方面有了系统的了解，带我了解科研的过程。感谢课题组的同学给予的帮助，无论在创新实践还是组会中，同学们的交流都让我受益匪浅。