南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

本科生毕业设计（论文）

题　　目：基于扩散模型的医学图像分割

姓　　名：易辰朗

学　　号：11910713

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 教授

2023 年 5 月 10 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

基于扩散模型的医学图像分割

易辰朗

（计算机科学与工程系　指导教师：刘江）

[摘要]：去噪扩散概率模型[1]最近受到了广泛的研究关注，因为它们优于 GAN[2]等替代方法，并且目前提供了最先进的生成性能。扩散模型优越的性能使其在修复、超分辨率和语义编辑等多种应用中成为一种有吸引力的工具。白内障手术是现存受众最广的眼科手术，如此庞大的手术量给医生带来了极大的负担,同时也促进了对提升白内障手术质量的需求。语义分割是许多计算机视觉任务的关键问题，特别是包括手术工具在内的医学领域图像。然而，现存的可供分析的白内障手术的数据量较少且收集和注释非常耗时。在本工作中，我证明了扩散模型也可以应用于医学图像语义分割，特别是在标签数据稀缺的情况下。对于预训练的扩散模型，我研究了来自执行扩散模型的反向去噪网络的中间激活，证明了这些激活层有效的捕捉到了输入图像中的语义信息，并且是分割问题的优秀像素级表示。基于这些研究结果，我运用了一种简单的分割方法，在只有少量标签数据的情况下，仍能提供较好的分割效果，为白内障手术流程分析提供了支撑。

［关键词］：语义分割；扩散模型；医学图像

[Abstract]: The denoising diffusion probability models[1](DDPM) have recently received widespread research attention because they outperform alternative methods such as GAN[2] and currently provide the most advanced generation performance. The superior performance of diffusion models makes them an attractive tool in various applications such as repair, super-resolution, and semantic editing. Cataract surgery is the most widely accepted ophthalmic surgery currently available, and such a large surgical volume places a great burden on doctors, while also promoting the demand for improving the quality of cataract surgery. Semantic segmentation is a key issue in many computer vision tasks, especially in medical images, including surgical tools. However, the amount of data available for analysis on cataract surgery is limited and collection and annotation are time-consuming. In this work, I demonstrated that diffusion models can also be applied to semantic segmentation of medical images, especially in situations where label data is scarce. For the pre-trained diffusion model, I studied the intermediate activation of the reverse denoising network from the implementation diffusion model, and proved that these activation layers effectively capture the Semantic information in the input image, and are excellent pixel level

representations of segmentation problems. Based on these research results, I applied a simple segmentation method that can still provide good segmentation results even with a small amount of label data, providing support for the analysis of cataract surgery process.

[Key words]: Semantic Segmentation; Diffusion Model; Medical Image

目录

1. 引言

外科图像语义分割越来越受到医学图像处理界的关注。目标通常不是在图像中精确地定位工具，而是指示外科医生每时每刻正在使用哪些工具。使用注释工具的主要动机是为手术工作流程分析设计有效的解决方案。分析手术工作流程在报告生成、手术训练甚至实时决策支持方面都有潜在的应用。

表征学习（或表示学习）是一种将原始数据转换成为更容易被机器学习应用的数据的过程。目前，一些表征学习方法例如 BiGan[3], autoregressive models[4]已经可以很好的应用于视觉任务的表征提取。在本研究中，我将扩散模型也运用到了表征学习中，在语义分割的背景下，证明了其可以很好的提取医学图像的特征。

特别地，我研究了来自 U-Net[5]网络的中间激活层，该网络实现了 DDPM 中反向扩散过程的马尔可夫步骤。直观地说，这个网络学会了对其输入进行去噪，但是不清楚为什么中间激活层可以捕获高级视觉问题所需的语义信息。然而，实验表明，在某些扩散步骤上，这些激活确实捕捉到了这些信息，因此，可以潜在地用作下游任务的图像表示。鉴于这些观察结果，我应用了一种简单的语义分割方法，该方法利用了这些表征，即使只提供了少量标记的图像，也能成功地工作。

扩散概率模型（为了简洁起见，我们将其称为“扩散模型”）是一个参数化马尔可夫链，使用变分推理进行训练，以在有限时间后产生与数据匹配的样本。学习该链的转换以反转扩散过程，扩散过程是一个马尔可夫链，它在采样的相反方向上逐渐向数据添加噪声，直到信号被破坏。当扩散由少量高斯噪声组成时，也可以将采样链转换设置为条件高斯，从而实现特别简单的神经网络参数化。

2. 相关工作

2.1 扩散模型

扩散模型是一种无监督生成模型，其灵感源自物理学中的扩散过程。它通过模拟数据样本之间信息的传播和扩散过程来生成新的样本。扩散模型可以用于生成各种类型的数据，包括图像、音频、文本等。其核心思想是利用当前数据样本与邻近样本之间的关系，逐步生成新的样本。扩散过程是一个马尔可夫链，它在采样的相反方向上逐渐向数据添加噪声，直到信号被破坏。扩散模型通过马尔可夫链的端点来近似真实图像的分布，马尔可夫链源自简单的参数分布，通常是标

准高斯分布。每个马尔可夫步骤都由一个深度神经网络建模，该网络用于学习去噪过程，将给定噪声还原成真实图像。

扩散模型被广泛应用于图像超分，修补，编辑任务[6]；同时，扩散模型在文本生成[7]中也又很好的表现；此外，扩散模型还可以应用于文本生成图像[8]。扩散模型在生成质量和多样性方面战胜了 GAN。在本研究中，我想证明扩散模型也可以成功的应用于语义分割。

2.2 白内障手术数据集的语义分割

语义分割是一种计算机视觉技术，涉及将图像中的每个像素分配到特定的类别或类别。DNN(Deep Neural Network)架构的不断进化为语义分割带来了巨大的好处。它们通常在具有密集像素级注释的数据集上进行训练。但是，白内障手术场景数据的标注是极其耗时耗力的。CNN(如 AlexNet[9]、VGG[10]或 ResNet[11])被广泛应用于语义分割领域，它们在白内障手术的语义分割和手术任务的实时识别上具有良好表现，但这都是建立在拥有大量具有标签的数据集上的.

在存在手术场景变化的情况下，手术工具分割需要精确的工具描绘。遮挡、阴影、反射和模糊在图像中很常见。这些像差降低了图像的质量，并影响了分割过程。因此，分割手术工具的过程被认为是相当困 难的。

Tonet 等人[12]提出了通过改变设备的视觉外观来使工作更容易进行跟踪的首批尝试之一。然而，这种方法对工具消毒产生了不利影响。改变工具的外观也需要一个独特的设置。对预先存在的手术设置甚至记录的适用性受到特殊设置的限制。因此，语义分割是解决分割问题的可靠而准确的方法。它用于将类别或类标签应用于图像中的每个像素，以进行像素级标记。这有助于正确识别手术工具的各种元件，以及从周围组织中描绘器械。

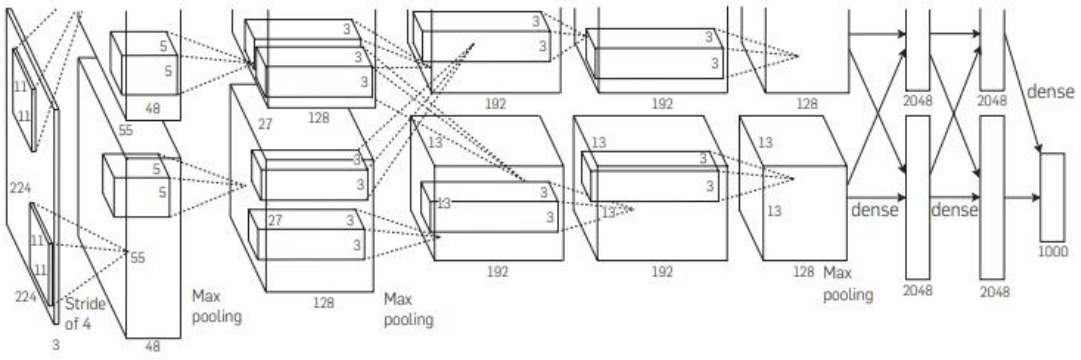

图 1 AlexNet 网络[9]

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

图 2　VGG 网络[10]

2.3　基于生成模型的表征学习

随着数据量的爆炸性增长，如何从大量的未标记数据中提取有用的信息成为了一个重要的问题。无监督学习生成模型因其能够从未标记数据中自动学习特征而备受关注。本调研旨在了解无监督学习生成模型的研究现状和应用。

生成模型是一种表示学习算法，可用于生成与训练数据相似的新数据样本。生成模型背后的关键思想是学习输入数据的概率分布，然后使用该分布生成新的样本。一种流行的生成模型是生成对抗网络(GAN)，它由两个神经网络组成：一个生成器和一个鉴别器。生成器以一个随机噪声向量作为输入，生成一个数据样本，而鉴别器则试图区分生成的数据和训练集中的真实数据。生成器被训练为生

成足以欺骗鉴别器的真实数据，而鉴别器则被训练为正确识别样本是真实的还是生成的。另一种流行的生成模型是变分自动编码器(VAE[13])，它是一种神经网络，可以学习输入数据的压缩表示。VAE 学习将输入数据编码到一个低维潜在空间，然后可以用来生成与训练数据相似的新数据样本。

无监督学习生成模型可以应用于多个领域，如图像生成[14]、图像修复[15]、语音合成[16]、文本生成[17]等。在图像生成领域，无监督学习生成模型能够生成逼真的图像，可以用于电影特效、虚拟现实等方面。在图像修复领域，无监督学习生成模型可以从部分损坏的图像中恢复出完整的图像。在语音合成领域，无监督学习生成模型能够生成逼真的语音，可以用于虚拟助手、语音识别等方面。在文本生成领域，无监督学习生成模型能够生成自然语言文本，可以用于自动文摘、机器翻译等方面。

总的来说，生成模型是一种强大的表示学习算法，可用于生成新的数据样本，以及学习可用于各种下游任务的有意义的输入数据表示。在本研究中，我想利用扩散模型在生成的过程中学习到的数据表示提高语义分割的效率。

3. 方法

本节的主要内容是对我提出的方法的整体框架和实现细节进行详细说明，对如何通过扩散模型预训练模型提取表征信息、如何使用像素级分类器（MLPs）进行图像分割、以及一些实验的意义和创新点进行了阐释，并且详细的介绍了该扩散分割模型的结构和参数设置。

3.1 方法概述

图一展示了我所提出方法的整体架构和运行流程。输入为真实白内障手术场景图像，对输入的图像进行固定步长的扩散加噪得到具有噪声的图像，然后将图像传入扩散去噪模型进行噪声预测，得到具有输入图像表征信息的像素级噪声。最终将该噪声传入多层感知机，预测得到每一个像素点的类别。

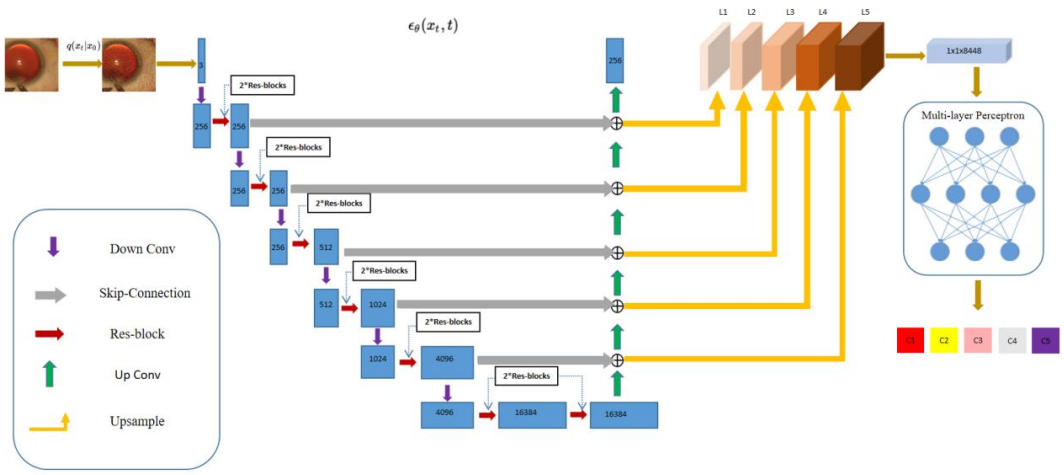

图 3　所提出扩散分割模型概述，由以下几个步骤组成：（1）输入图像 x_0 经过 $q(x_t|x_0)$ 添加高斯噪声得到 x_t。（2）从训练好的噪声预测器 $\varepsilon_\theta(x,t)$ 中提取图像的表征信息。（3）通过将特征图上采样得到与原图相同大小的特征图并将其拼接来收集像素级表示。（4）使用得到的该像素级特征向量来训练 MLPs，使其能预测每个图像的标签。

3.2 扩散模型预训练提取表征

3.2.1 扩散模型基本理论

根据扩散模型的基本理论，在白内障手术数据集上训练一个扩散模型，使其能够生成真实白内障手术场景图片。首先介绍一下扩散模型的基本框架如图二所示。

前向过程由于每个时刻 t 只与 t-1 时刻有关，所以可以看做马尔科夫过程，在马尔科夫链的前向采样过程中，也就是扩散过程中可以将数据转换为高斯分布。即扩散过程通过 T 次累积对输入数据 x_i 添加高斯噪声，如图 4 所示将这个跟马尔可夫假设相结合，于是可以对扩散过程表达成:

$$q(x_t|x_0) := N(x_t; \sqrt{\bar{\alpha}}x_0, (1-\bar{\alpha}_t)I)$$

$$\sqrt{\bar{\alpha}}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \varepsilon \subset N(0,1) \quad (1)$$

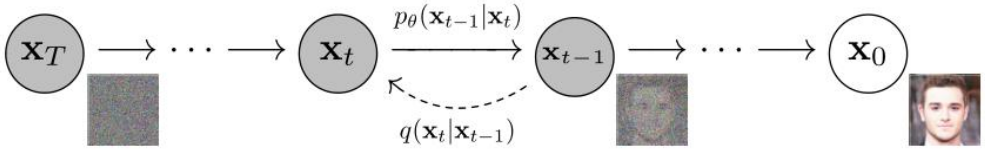

图 4　扩散过程示意图[1]

如果说扩散过程是加噪的过程，那么逆扩散过程就是去噪推断过程。如果我们能够逐步得到逆转后的分布,可以从标准高斯分布还原出样本数据的分布:

$$\begin{aligned} p_\theta(x_{T:0}) &:= p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}\mid x_t) \\ &:= p(x_T)\prod_{t=1}^{T} N(x_{t-1};\mu_\theta(xt,t),\sum\nolimits_\theta(x_t,t)) \end{aligned} \tag{2}$$

根据马尔可夫规则表示，逆扩散过程当前时间步 t 只取决于上一个时间步 t-1，所以有：

$$p_\theta(x_{t-1}\mid x_t) := N(x_{t\text{-}1};\mu_\theta(x_t,t),\sum\nolimits_\theta(x_t,t)) \tag{3}$$

这里，提出该方法的人的想法也很直接：我们已经拥有了整个数据集，既然无法直接求解出结果，那为何不尝试训练出一个模型来对这些噪声的条件概率进行预测呢？既然要做预测，那标签何来？这里便是将前向传播每部生成的真实噪声记录下来作为标签，前向扩散的过程除了推断外，还包含类似该数学模型所用“数据集的构建过程”。在模型做逆向扩散时，即可对前向扩散中所产生的高斯噪声进行预测，并一步一步推断，以还原最初始的样本数据。

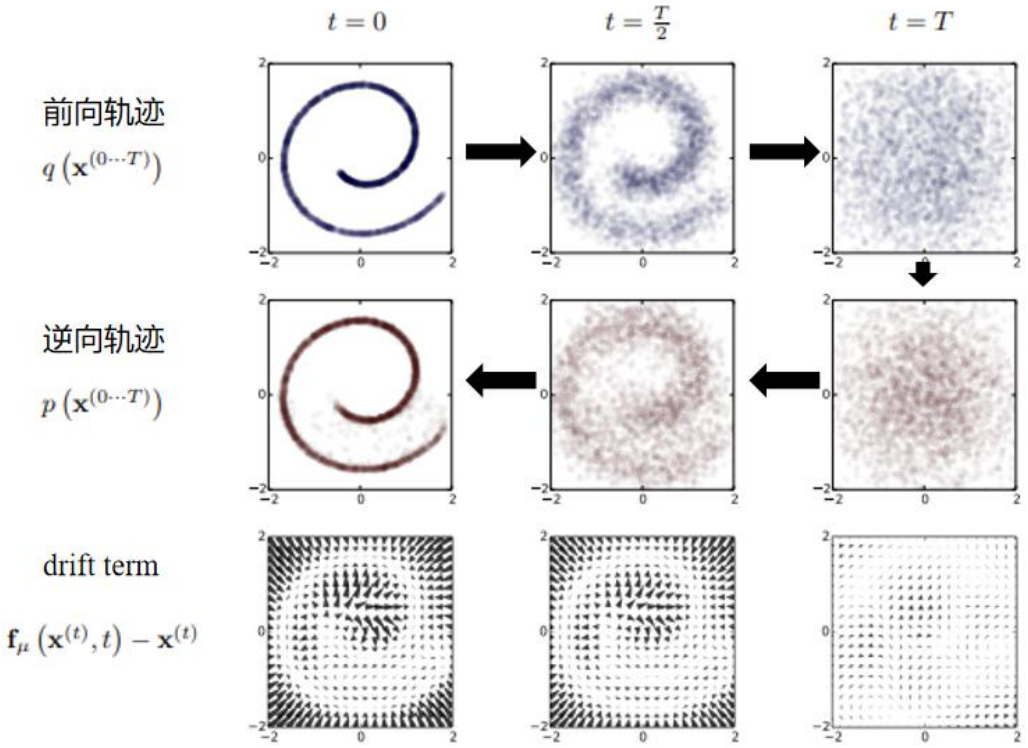

图 5　扩散模型训练过程示意图[18]

值得关注的是扩散模型中噪声预测器部分，噪声预测器生成的噪声中包含了输入图片的深层表征。

3.2.2 扩散模型细节实现

我所提出的模型中使用的扩散模型采用了 Res-UNet[19]的思想，Res-UNet 是一种结合了 ResNet 和 UNet 概念的神经网络架构。

ResNet（残差网络）是一个深度学习网络框架，它是卷积神经网络(CNN)的一种，广泛应用于图像识别和目标检测等计算机视觉任务。ResNet 的关键创新是残差连接层的使用，它允许比以前更深入的网络训练。在传统的神经网络中，每一层接收前一层的输出作为其输入。在 ResNet 中，一些层具有绕过一个或多个层的快捷连接，并将输入直接提供给后面的层。当用于反向传播的梯度在许多层中传播时变得非常小时，就会出现梯度消失的问题，但这种方法使得 ResNet 通过解决梯度消失的问题来有效地训练非常深的网络(多达数百层)。通过提供快捷连接，使梯度更容易流动，ResNets 使深度网络的训练能够在广泛的图像识别和其他计算机视觉任务上实现最先进的性能。ResNet 的网络结构示意图如图 6 所示。

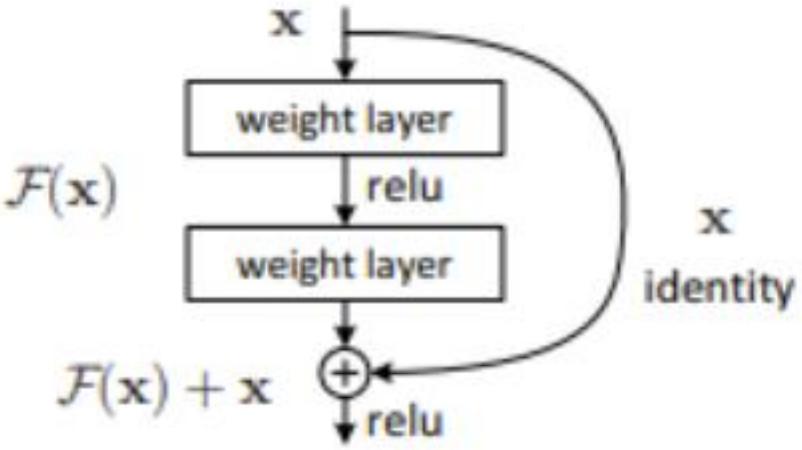

图 6　ResNet 示意图[11]

UNet 是一种卷积神经网络(CNN)架构，最初是为生物医学图像分割任务开发的，但后来被广泛用于图像分割应用。UNet 架构由一个编码器和一个解码器组成，分别由一条下采样路径和一条上采样路径连接。下采样路径由卷积层和池化层组成，它们减少了输入图像的空间维度，同时增加了特征映射的数量。该路径从输入图像中提取高级特征。上采样路径由上采样层和卷积层组成，分别增加特征映射的空间维度同时减少特征映射的数量。该路径根据编码器学习到的特征重建分割后的图像。UNet 最引人注目的一点是它使用编码器和解码器之间的跳越连接来保留低级特征，这有助于提高分割精度。UNet 在各种生物医学图像分割任务中取得了最先进的性能，包括细胞跟踪、细胞核分割和血管分割。UNet 的网络结构示意图见图 7。

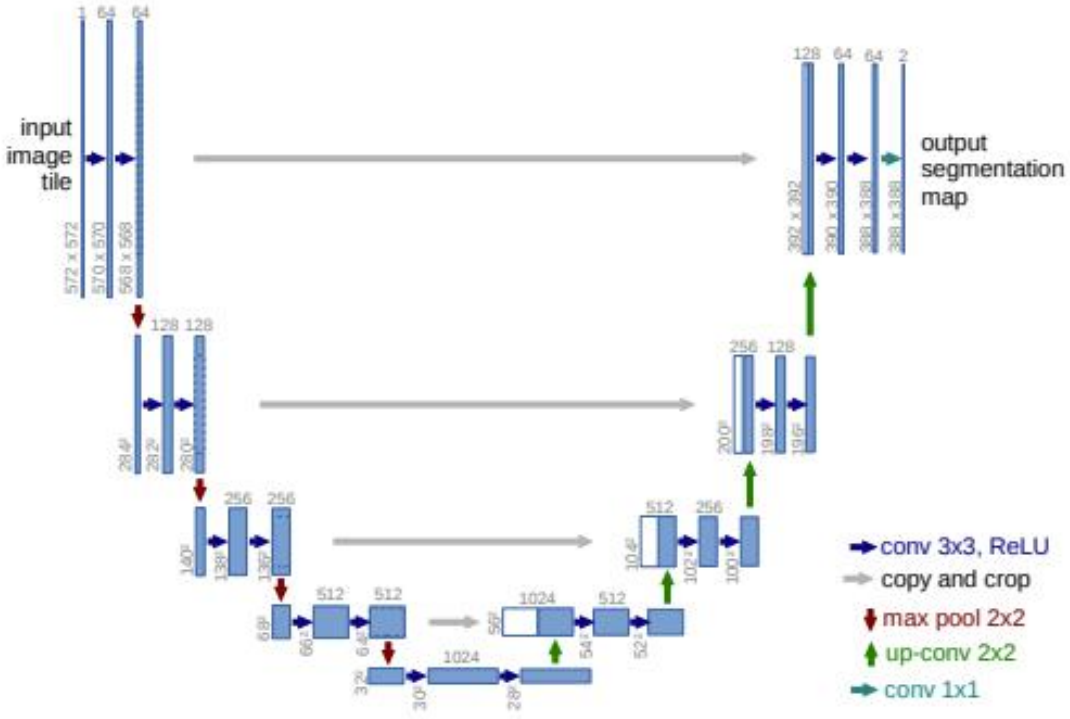

图 7　UNet 示意图[5]

而本研究所运用的 Res-Unet 模型正是将这两者进行很好的结合。如图 8 所示，该网络延续了 UNet 的编码器和解码器结构,与传统 UNet 网络不同的是，Res-Unet 网络在上采样和下采样过程中加入了 Res-blocks（残差块）。编码器和解码器都由 6 个 layers 组成，每个 layer 都包括两个残差块和一个上采样块/下采样块，总计 18 个编码块和 18 个解码块。

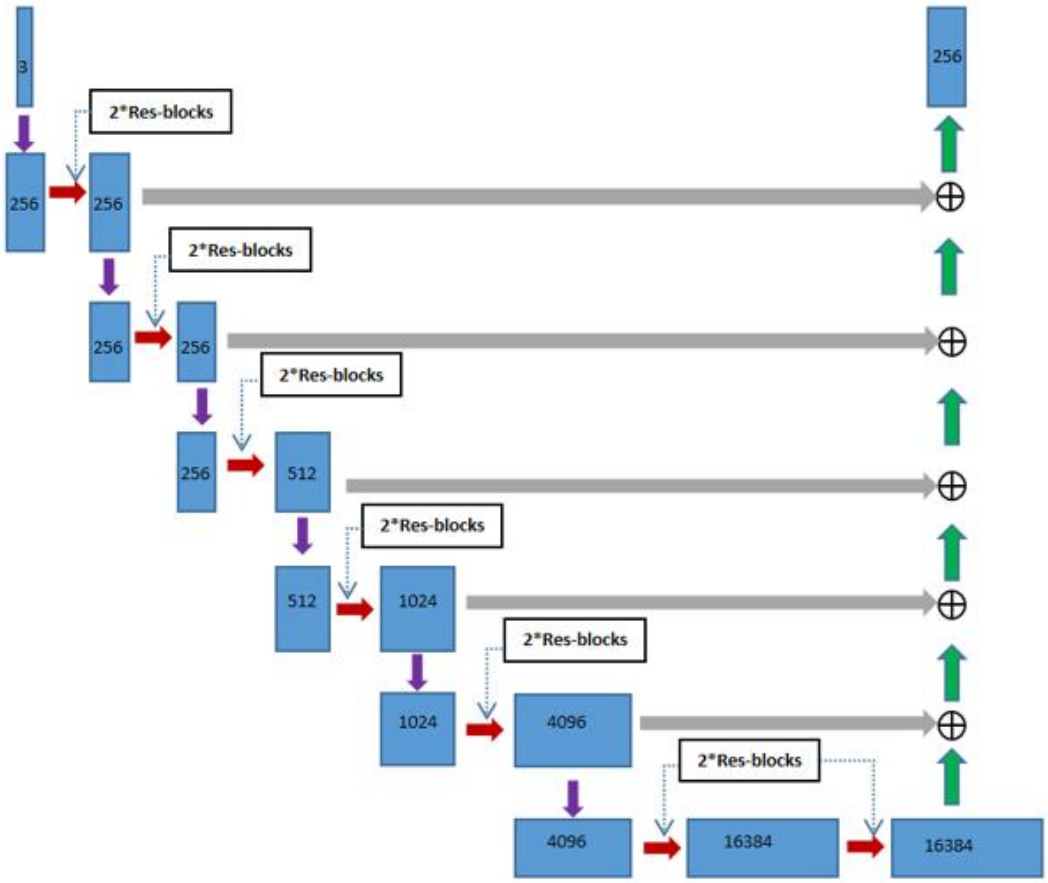

图 8　Res-UNet 示意图

3.2.3 特征提取

对于一个输入的图片 $x_0 \in RH \times W \times 3$,给定扩散时间步长 t，通过（2）等式对图像加噪，我们可以得到被破坏了的图片 x_t。之后具有噪声的 x_t 作为输入传入传入 Res-UNet 噪声预测网络，并从该网络的中间激活层中得到多组激活张量，并将这些激活张量上采样到 H×W 作为 x_0 的像素级激活表示。这些激活张量包含了输入图片 x_0 的表征信息

3.3 像素级分类器

该像素级分类器的本质是 MLPs(Multi-layer Perceptrons),用于对输入表征的每个像素点进行预测。经过噪声预测器得到的像素级表征被传入该 MLPs，训练该 MLPs,使得其尽可能对每一个像素点进行正确的预测。

MLPs 的中文全称是多层感知器，网络结构如图 9 所示。它是一种通常用于机器学习的人工神经网络(ANN)。MLPS 由多个线形层组成，每一层都连接到下

一层。输入层接收输入数据，输出层产生最终输出。中间层被称为隐藏层，负责处理输入数据。MLPS 用于各种任务，包括分类、回归和预测。它们对于输入和输出之间存在非线性关系的问题特别有用。MLP 使用一种称为反向传播的监督学习算法进行训练，该算法调整网络的权重和偏差，以最大限度地减少预测输出与实际输出之间的误差。总的来说，MLPS 是一个强大而灵活的机器学习工具，它们已经成功地应用于广泛的应用，包括图像和语音识别、自然语言处理和金融预测。在本实验中使用 MLPS 来对具有输入图像表征信息的噪声的每一个像素点进行预测。

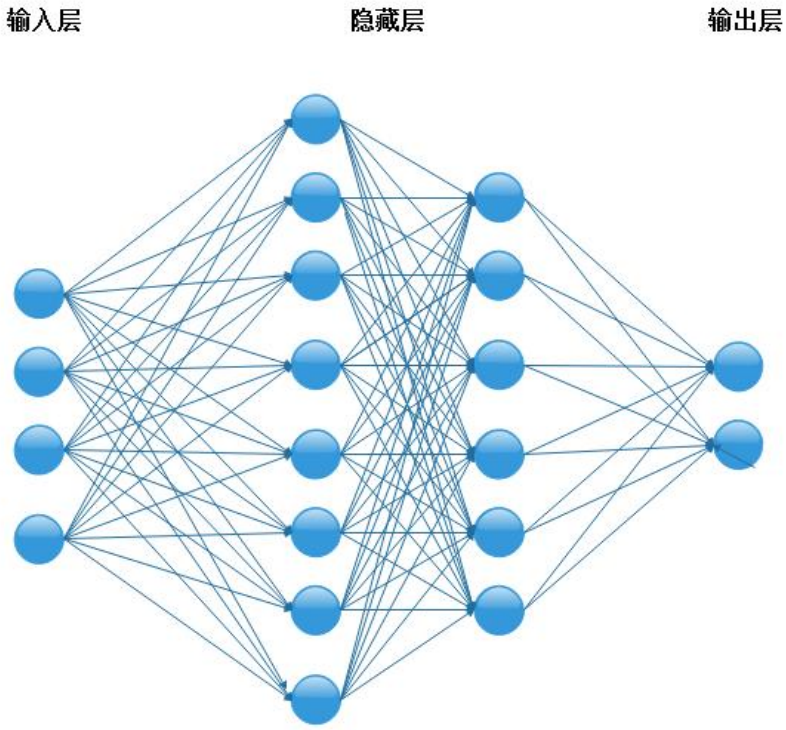

图 9　MLPS 示意图

4 实验

4.1 数据集

4.1.1 数据集简介

在本研究中使用了一个真实白内障手术视频 CaDIS[20]数据集，是一个用于图像分割的数据集。它涉及 4738 张注释图像，这些图像是从 2015 年 1 月 22 日至 2015 年 9 月 10 日在布雷斯特大学医院进行的白内障手术的 50 段视频中

采样得到的。我使用了 CaDIS 数据集中 4738 图片训练 Diffusion 模型，之后从中挑选了 20 张带标签的图片作为训练集、100 张图片作为测试集。

4.1.2 数据清理

一个好的数据集是实验成功的基础。CaDIS 数据集比较脏，有许多杂乱数据和错误数据，为了能得到更好的实验结果，我对该数据集通过以下步骤进行了数据清理。1、根据白内障手术彩色图谱[21]删除所有错误的数据。2、使用 Coco annotator 对数据集中未注释的部分进行注释。3、根据白内障手术彩色图谱合并一些类别。4、删除没有任何标注的图片。5、重新标注标签错误的图片。6、调整图片大小以适应网络结构。

4.1.3 标签分类

由于该 CaDIS 数据集手术器械类别较多（31 类）且每类手术器械数量分布不均匀，因此想要对每一类手术器械都得到很好的分割结果是相当困难的。为了简化问题的同时也能达到预期效果，我将所有手术器械合并为了一类，与瞳孔、虹膜、角膜、皮肤组成该数据集的五类标签如表 1 所示。

表 1　Cadis 数据集标签表

名称	标签	颜色
瞳孔	L1	红色
皮肤	L2	黄色
虹膜	L3	粉色
角膜	L4	灰色
手术器械	L5	紫色

4.2　主实验细节

4.2.1 软硬换件设置

我所使用的服务器软硬件平台配置如下：硬件方面，CPU 为 Intel i9-10900，内存大小为 64GB，GPU 为 NVIDIA GeForce RTX 3060（显存大小为 12GB）。软件方面，操作系统为 Windows10，编译器为 PyCharm，使用 Pytorch 机器学习框架（1.10.0+cu113）。

4.2.2 扩散模型预训练

我使用的扩散模型的基本框架采用了上文中提到过的 Res-UNet 的思想。我选用的了 CaDIS 数据集中的 4738 张图片来训练该 Diffusion 模型，我选择的 Batch size 为 1，learning rate 为 1e-4, 图像大小为 256*256，编码器和解码器分别由六个 layer 组成，每个 layer 包含两个残差块和一个上采样块/下采样块，扩散总步数为 4000 步。在训练完成之后，我用训练好的模型生成了一些该域下的图片，我挑选了几张生成图片与真实图片对比结果展示如下：

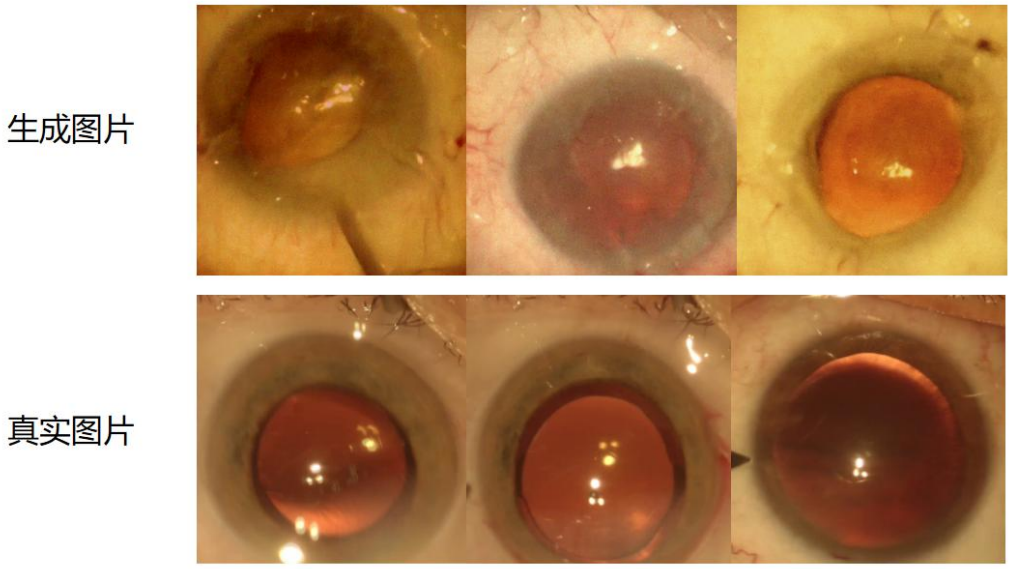

图 10　Diffusion 模型生成图片与真实图片对比

可以看到，生成的图片很好的学习到了真实图片的信息，无论是角膜，虹膜，还是瞳孔，甚至是手术器械都能达到较好的生成效果，这也证明了我所运用的 diffusion 模型是有效的，为之后的特征提取分析以及分割做了铺垫。

4. 2. 3 图像表征分析

上文提到 Res-UNet 噪声预测器的中间激活层可以捕获语义信息，但是并不是所有的中间激活层都能有效的捕捉到输入图像的语义信息，盲目的将所有中间激活层合并起来作为输入传入 MLPS 会导致网络参数过多，且有效信息不能很好地突出。同时扩散步长 t 同样也决定了最后分割结果的好坏，t 过小和过大都会导致噪声预测器无法很好的提取到输入图像的语义信息。因此，为了探究中间激活层中哪些层较好地捕捉到了输入图像地语义信息且选择多大的扩散步长 t 比较合适，我在给定不同扩散步长 t（20、150、300、500）对 18 个 Res-UNet 的解码器块中的第 2、5、8、11、14 块所得到的表征信息进行单独预测像素级语义标签。

MLPS 在 20 张从 CaDIS 数据集挑出的图像上进行训练，在 100 张图片上进行测试，使用 MIoU 作为评价指标，我得到了以下的结果如图 11 所示。

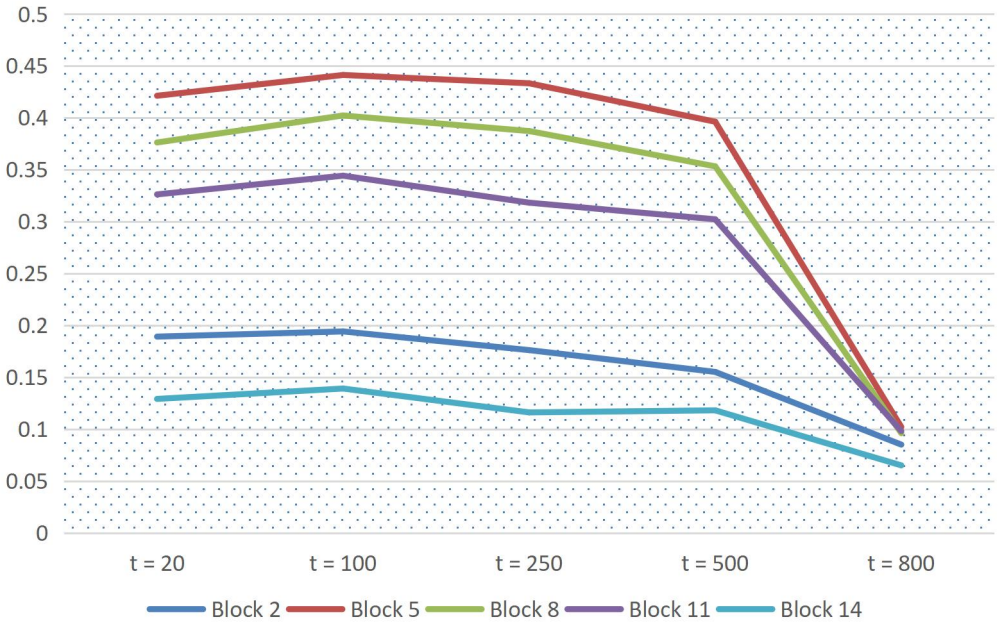

图 11　MIoU 对比图

根据该实验结果，我发现当解码块选择 5-11 块且当扩散步长处于 20-250 之间时，能得到较好的分割结果，这也从侧面说明该条件下中间激活层较好的捕捉到了输入图像的语义特征。最终我选择了 UNet 解码块中的 5，6，7，8，10 块和扩散步长 20，50，250，总共得到了 15（3×5）个解码块的激活层，并将这些激活层的信息拼接起来，最终得到的像素级特征中每个像素点的长度是 8448。

4.2.4 图像分割

经过图像表征分析，我得到了具有输入图像语义信息的像素级表征，每个像素的长度是 8448。我将每个像素点的向量作为我的像素级分类器的输入，经过该像素级分类器对每一个像素点进行预测，得到像素级的分割结果。我选择的 batch size 是 1，learning rate 是 1e-3，损失函数为交叉熵损失，优化器为 ADAM。MLPs 的主要结构由三个线形层组成，分别为（256，128）、（128，32）、（32，5），选择 relu 作为激活函数，使用 BatchNorm1d 来进行归一化。

4.3 对比方法介绍

在本实验中选择了三个基于无监督学习的表征学习方法来进行对比，分别是

DataSetgan[22], MAE[23]和 SwAV[24]。

4.3.1 DataSetgan 方法介绍

DatasetGAN 使用 NVIDIA 的 StyleGAN[25]技术生成真实感图像, 如图 9 所示。人类注释器对图像中的对象部分进行详细的标签，然后对该数据进行解释器训练，以从样式的潜在空间生成特征标签。结果是一个系统，它可以生成无限数量的图像和注释，然后可以作为任何计算机视觉（CV）系统的训练数据集。

NVIDIA 对 DatasetGAN 的理解是，作为生成器输入的潜在空间必须包含有关生成图像的语义信息，因此可以用于为图像创建注释映射。研究小组通过首先生成几个图像并保存与之相关的潜在向量，为他们的系统创建了一个训练数据集。对生成的图像进行人工标注，并将潜在向量与这些标注配对进行训练。然后利用该数据集训练一组多层感知器（MLP）分类器作为风格解释器。分类器输入由 GAN 产生的特征向量组成，生成每个像素，输出为每个像素的标签；例如，当 GAN 生成人脸图像时，解释器输出指示人脸部分的标签，例如脸颊、鼻子或耳朵。

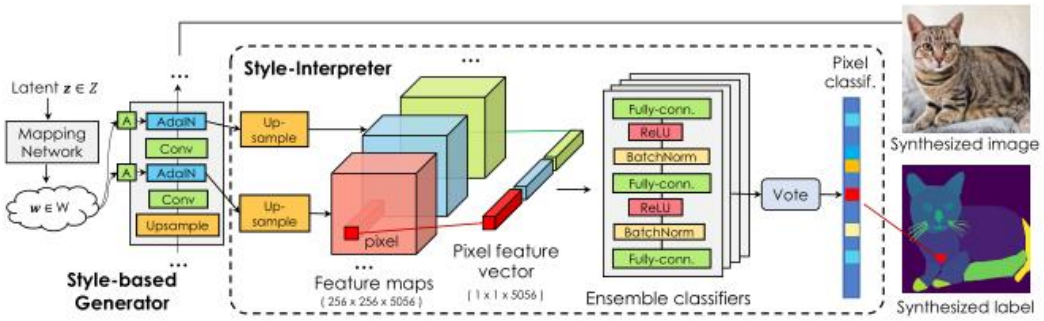

图 12　DatasetGAN 结构图[13], 特征图从 StyleGAN 上采样到最高分辨率，以构建合成图像上所有像素的逐像素特征向量。然后训练 MLP 分类器的集合，用于将像素的特征向量中的语义知识解释为其部分标签。

我使用了该网络结构中 Generator 作为特征提取器，后续分割使用的 MLPs 与主实验中使用的 MLPs 一样。

4.3.2 MAE 方法介绍

MAE(Masked Autoencoders)是用于 CV(Computer Version)的自监督学习方法如图 13 所示，优点是扩展性强的（scalable），方法简单。在 MAE 方法中会随机 mask 输入图片的部分 patches，然后重构这些缺失的像素。MAE 基于两个核

心设计：（1）不对称的（asymmetric）编码解码结构，编码器仅仅对可见的 patches 进行编码，不对 mask tokens 进行任何处理，解码器将编码器的输出（latent representation）和 mask tokens 作为输入，重构 image；（2）使用较高的 mask 比例（如 75%）。MAE 展现了很强的迁移性能，在 ImageNet-1K 上取得了 best accuracy（87.8%），且因为方法简单，可扩展性极强（scalable）在 NLP 领域自监督学习方法使用十分广泛，但是在 CV 领域，大多数预训练还是采用监督方式，因此 MAE 最大的贡献就是证明了自监督预训练同样可以在 CV 领域获得和监督预训练一样，甚至更好的效果，自监督也许可以像统治 NLP 一样统治 CV 领域在 NLP 领域自监督学习方法使用十分广泛，但是在 CV 领域，大多数预训练还是采用监督方式，因此 MAE 最大的贡献就是证明了自监督预训练同样可以在 CV 领域获得和监督预训练一样，甚至更好的效果，自监督也许可以像统治 NLP 一样统治 CV 领域。

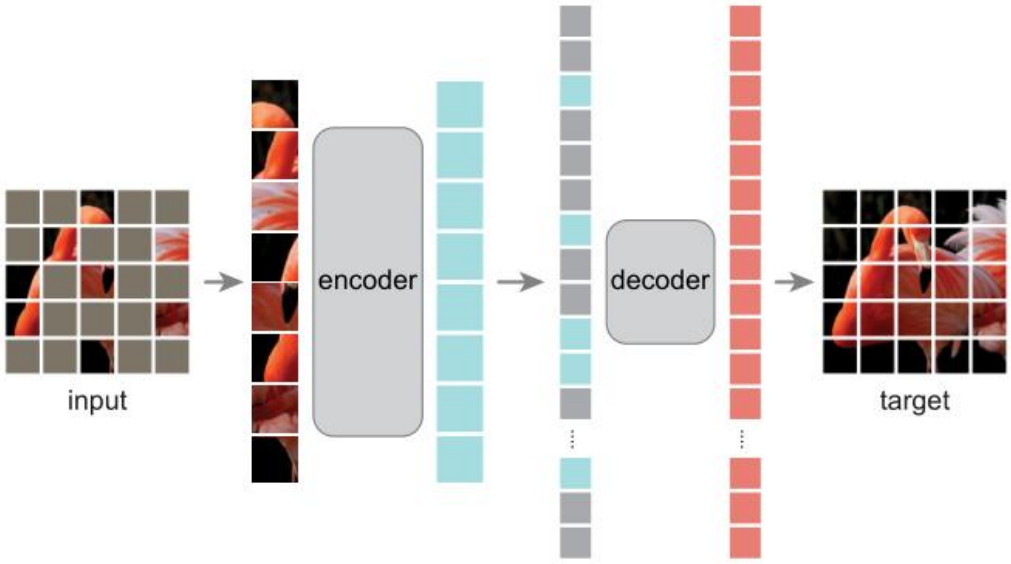

图 13　MAE 结构图，在预训练期间，图像补丁的大的随机子集（例如，75%）被屏蔽掉。编码器应用于可见补丁的小个子集。在编码器之后引入掩码令牌，并且由以像素重建原始图像的小型解码器处理完整的编码补丁和掩码令牌集。在预训练之后，解码器被丢弃，编码器被应用于未损坏的图像（完整的补丁集）以用于识别任务。

我使用了该网络结构中 encoder 作为特征提取器，后续分割使用的 MLPs 与主实验中使用的 MLPs 一样。

4.3.3 SwAV 方法介绍

SwAV（Swapping Assignments between Views）利用了对比学习方法的优点，而不需要计算成对的比较，如图 11 所示。具体来说，SwAV 在对数据进行聚类的同时，强化同一图像不同视图的聚类分配之间的一致性，而不是像对比学习那样直接比较特征。简单地说，我们使用“换位”预测机制，从一个视图的表示中预测另一个视图的 code。我们的方法可以进行大批量和小批量训练，并且可以扩展到无限量的数据上。与之前的对比方法相比，我们的方法更加高效，因为它不需要大的内存或特殊的动量网络。此外，我们还提出了一种新的数据增强策略 multi-crop，该策略混合使用多个分辨率的视图来代替两个全分辨率的视图，而不增加内存或计算需求。

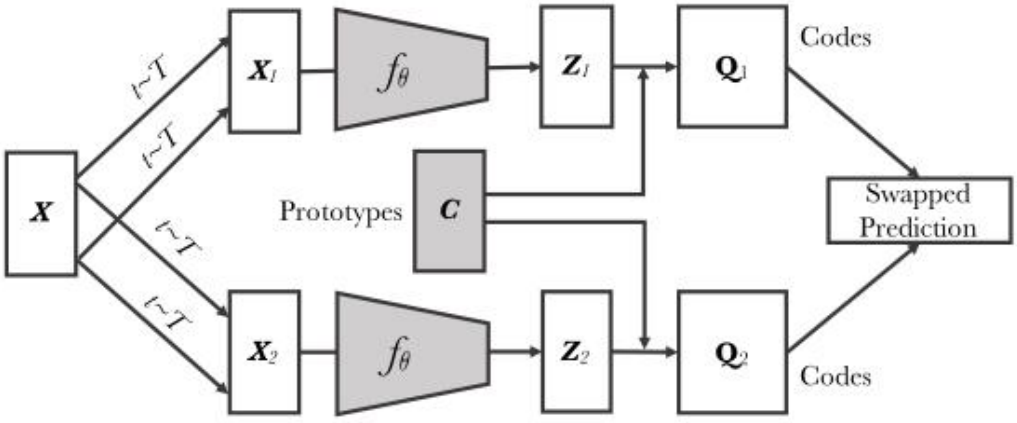

图 14　SwAV 结构图，在 SwAV 中，首先通过将特征分配给原型向量来获得“代码”。然后，我们解决了一个“交换”预测问题，其中使用另一个视图来预测从一个数据增强视图获得的代码。因此，SwAV 不直接比较图像特征。原型向量与 ConvNet 参数一起通过反向传播进行学习。

与前两组对比实验类似，我使用了f_θ作为特征提取器，后续分割使用的 MLPs 与主实验中使用的 MLPs 一样。

4.4 评价指标

我们的实验着眼于模型的分割精度，因此我使用了 MIoU 作为提出模型的评价指标。

在介绍 MIoU 之前，我们首先要知道什么是混淆矩阵，混淆矩阵的结构是一个方阵，每一行代表真实的类别，每一列代表预测的类别。在二分类问题中，混淆矩阵的四个元素通常被称为真正例（True Positive，TP）、假正例（False Positive，

FP）、真反例（True Negative，TN）和假反例（False Negative，FN）。这些元素可以用来计算许多不同的分类性能指标，例如准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 分数（F1-Score）等如表 2 所示。

其中，TP 表示模型正确地将正样本预测为正样本的数量，FP 表示模型错误地将负样本预测为正样本的数量，FN 表示模型错误地将正样本预测为负样本的数量，TN 表示模型正确地将负样本预测为负样本的数量。

表 2　混淆矩阵，纵列为真实值，即分割任务中的 Ground Truth；横行为模型的预测值。

GT / Pred	Positive	Negative
Positive	TP	FP
Negative	FN	TN

在图像分割任务中，我们希望将一张输入图像中的每一个像素分配到其对应的类别中，因此我们需要比较模型预测的分割结果和真实的分割结果之间的差异。MIoU 是一种常用的评估指标，它通过计算预测分割结果与真实分割结果的交集和并集的比值来评估算法的性能。具体来说，MIoU 的计算公式如公式 4 所示：

$$MIoU = (1/n)*\sum(TP\_i / (TP\_i + FP\_i + FN\_i)) \quad (4)$$

其中，n 是类别数，TP_i、FP_i 和 FN_i 分别是第 i 个类别的真实正样本数、假正样本数和假负样本数。这里的 TP_i 表示模型正确预测为第 i 个类别的像素数，FP_i 表示模型错误地将其它类别的像素预测为第 i 个类别的像素数，FN_i 表示模型错误地将第 i 个类别的像素预测为其它类别的像素数。

MIoU 的取值范围在 0 到 1 之间，值越大表示模型的性能越好。它可以对每个类别的性能进行单独评估，也可以对所有类别的性能进行综合评估。MIoU 在语义分割、实例分割等图像分割任务中广泛应用。

4.5 实验结果

本文针对稀缺标签下的语义分割问题进行了深入的研究，通过设计了四组实验并分析实验结果，得出了如下结论。

4.5.1 MIoU 结果分析

表 3　CaDIS 数据集五组实验分割 MIoU 结果

名称	MIoU
扩散分割模型	0.512
MAE	0.477
DatasetGAN	0.521
SwAV	0.418
UNet	0.164

可以看到当只用传统 UNet 分割网络在只有 20 张带数据标签的数据集上训练得到的 MIou 分割结果很差，而当使用基于无监督学习的表征学习方法所得到的 MIoU 分割结果相较于传统 UNet 网络有较大提升，且我所提出的扩散分割模型在 CaDIS 数据集上的分割表现优于大部分对比方法。这也证明了基于无监督学习的对比学习和生成学习都具有很好的提取特征的能力，且扩散模型在生成图像过程中相较于其他无监督方法有更好的提取特征的能力。

4.5.2 可视化结果分析

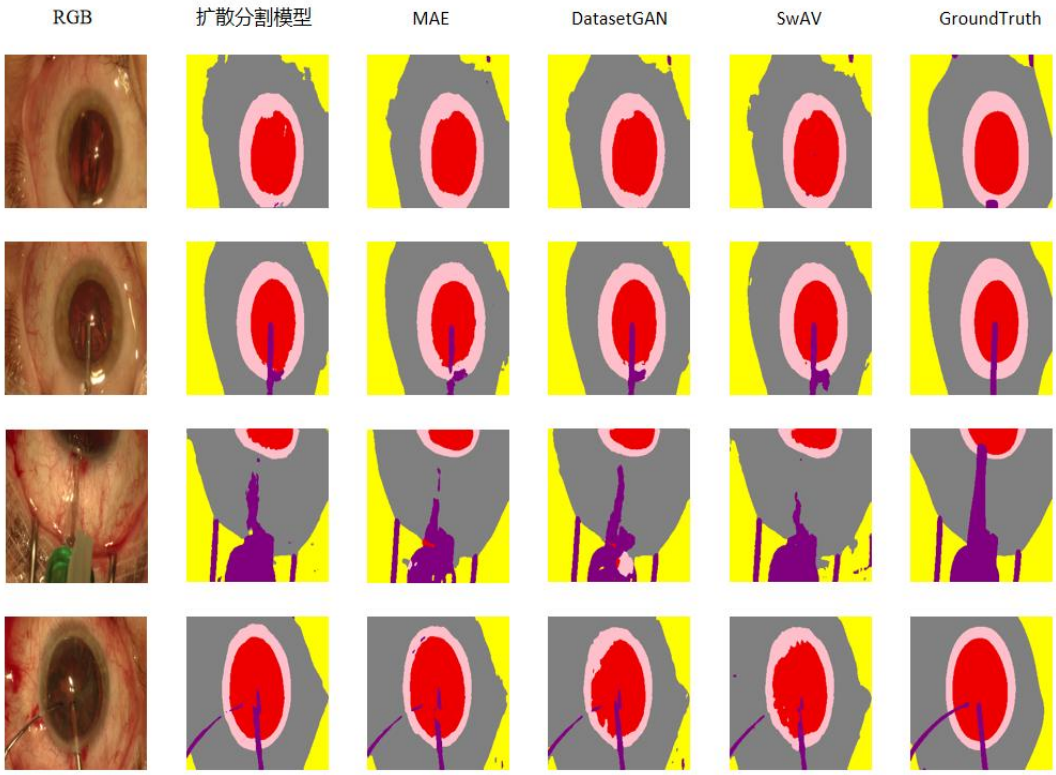

图 15　CaDIS 数据测试集分割结果可视化图

图 15 为 CaDIS 数据测试集的分割结果可视化图，我从中挑选了四组具有代表性的图片进行展示，可以看出在瞳孔和虹膜以及角膜的分割效果上，扩散分割模型具有更好的平滑度。在手术器械的分割效果上，扩散模型的优势则更为明显，对手术器材的分割完整的和平滑程度都要优于大部分对比方法。

5. 结论

白内障是一种常见的眼病，需要通过手术治疗。随着医学技术的发展，白内障手术越来越普及。然而，手术过程中医生需要根据视野中的图像来进行操作，因此对于图像的准确分割具有重要意义。然而由于眼球本身形态复杂，手术器械繁多，可用数据集较少且收集带标签的图像非常耗时耗力，该任务一直被认为是医学图像处理任务中较为困难的一种。在本研究中，我提出了一种基于扩散模型的白内障手术数据集分割算法，以扩散模型作为无监督预训练提取特征，使用 MLPs 进行图像分割，在只具有很少带标签图片的数据集上仍能表现出较好的分割效果，证明了设计思路的有效性。

当然我目前的工作还存在着以下几个不足：

1. 数据分布不平衡，有些手术器械的数据量很大，有些手术器械的数据量很小，影响特征提取和最后的分割效果。
2. 手术器械类别太多，仅仅挑选 20 张图片很难覆盖到所有手术器械的特征。
3. 对比方法较少，应该多添加一些无监督学习（包括对比学习和生 成学习）的对比方法。
4. 我使用的 Res-UNet 网络还具有很大的改进空间，例如加入注意力机制，增加 Res-block，使用更多的中间激活层等等一些增加性能的操作。

在后续的工作中，我将尝试从上述角度来改进我的方法，争取在少标签的白内障手术数据集分割中达到更好的效果。

参考文献

[1] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

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

[3] Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning[J]. arXiv preprint arXiv:1605.09782, 2016.

[4] Chen M, Radford A, Child R, et al. Generative pretraining from pixels[C]//International conference on machine learning. PMLR, 2020: 1691-1703.

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[6] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10684-10695.

[7] Gong S, Li M, Feng J, et al. Diffuseq: Sequence to sequence text generation with diffusion models[J]. arXiv preprint arXiv:2210.08933, 2022.

[8] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. Advances in Neural Information Processing Systems, 2022, 35: 36479-36494.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and

pattern recognition. 2016: 770-778.

[12] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International Conference on Machine Learning. PMLR, 2015: 2256-2265.

[13] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[14] 陈佛计,朱枫,吴清潇,郝颖明,王恩德,崔芸阁. 生成对抗网络及其在图像生成中的应用研究综述[J]. 计算机学报,2021,044(002):347-369

[15] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: Feature learning by inpainting[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2536-2544.

[16] Wang Y, Stanton D, Zhang Y, et al. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis[C]//International Conference on Machine Learning. PMLR, 2018: 5180-5189.

[17] Li J, Tang T, Zhao W X, et al. Pretrained language models for text generation: A survey[J]. arXiv preprint arXiv:2105.10311, 2021.

[18] Tonet O, Ramesh T U, Megali G, et al. Tracking endoscopic instruments without localizer: image analysis-based approach[J]. Stud Health Technol Inform, 2006, 119: 544-549.

[19] Diakogiannis F I, Waldner F, Caccetta P, et al. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 162: 94-114.[11]He K, Zhang X, RenS, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[20] Quellec G, Lamard M, Cochener B, et al. Real-time segmentation and recognition of surgical tasks in cataract surgery videos[J]. IEEE transactions on medical imaging, 2014, 33(12): 2352-2360.

[21] 王新俊. 《白内障显微手术彩色图谱》一书出版[J]. 临床眼科杂志, 2009, 17(5):1.

[22] Zhang Y, Ling H, Gao J, et al. Datasetgan: Efficient labeled data factory

with minimal human effort[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10145-10155.

[23] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16000-16009.

[24] Caron M, Misra I, Mairal J, et al. Unsupervised learning of visual features by contrasting cluster assignments[J]. Advances in neural information processing systems, 2020, 33: 9912-9924.

[25] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4401-4410.

致谢

在我即将毕业的时刻，我想用这篇致谢来表达我最真诚的感激之情。在这四年的大学生涯中，有太多的人和事给予了我无私的支持和帮助，让我能够完成自己的学业，实现自己的梦想。在此，我要向每一位曾经给予我帮助的人表示由衷的感谢！

首先，我要向我的课题组的导师刘江老师和指导我的李衡博士，刘浩峰学长表达我的感激之情。感谢你们在学术上的指导和启发，让我学会了如何去发现问题、解决问题。感谢你们对我学习、生活、职业规划等方面的关心和帮助，让我在大学的时光中不断成长，成为一个更好的自己。

其次，我要感谢我的父母和家人。感谢您们在我远离家乡、面临学业压力、独自面对困难时，始终给予我的精神支持、物质保障和鼓励。感谢您们一直在背后默默支持我，让我能够安心学习、积极成长。

此外，我还要感谢我的同学和朋友们。感谢你们在我遇到困难时给予的帮助和支持，让我感受到了友谊的温暖。感谢你们一起度过的美好时光，让我的大学生涯充满了欢声笑语和回忆。

最后，我要感谢学校和各位老师们的培养和教育。感谢学校为我提供了优秀的学习环境和各种资源，让我得以在这里接受全面的知识教育。感谢各位老师的教育和教诲，让我在思想上得到了极大的拓展和提高。

虽然我无法一一列举每一个帮助过我的人和事，但是我会一直铭记在心。再次感谢所有支持和帮助我的人，是你们让我能够走到今天，也是你们让我的未来更加充满信心和希望！

最后，祝福大家身体健康、工作顺利！