**第4章 鱼类摄食行为分类模型**

本论文的主要研究内容是基于计算机视觉技术建立鱼类摄食行为模型，在第二章介绍了实验平台的搭建，完成了视频采集与数据集标注的工作；第三章的主要研究内容是对实验数据的预处理与视频数据空间变换的特征提取。本章将在前两章的基础之上提出一种视频分类方法，实现根据鱼类的摄食行为视频判断鱼类摄食欲望的算法模型。下面将从方法的提出，方法的具体内容，以及实验验证三个方面具体阐述。

4.1. 基于帧间关系的视频分类综述

本论文所提出的视频分类方法是基于视频的帧间关系进行视频分类，所以有必要回顾现阶段的基于视频帧间关系对视频行为分类的研究。总的来说，基于帧间关系的视频行为分类算法主要有两个方向：第一种直接研究视频图像，比较前后图像内容实现对视频分类；第二种研究视频画面隐含变量的相关关系进而对视频行为分类。下面分别进行介绍。

4.1.1. 基于视频浅层特征视频分类

为方便描述，我们定义一个二分类视频问题。假设存在两个类别的视频,视频中对应的图像分别为。

（1）帧间差值法

帧间差值法是对视频内容最直接的描述，通过比较前后视频帧像素值的差异可以判断视频内容的变化剧烈程度，所以基于此可以对变化程度不同的视频进行分类。比如当鱼类运动较为平静时和运动剧烈下的两个视频类别的帧间差值是不同的，通过设置不同的阈值就可以将两个类别的视频区分开来。平均帧间差值的数学定义如下：

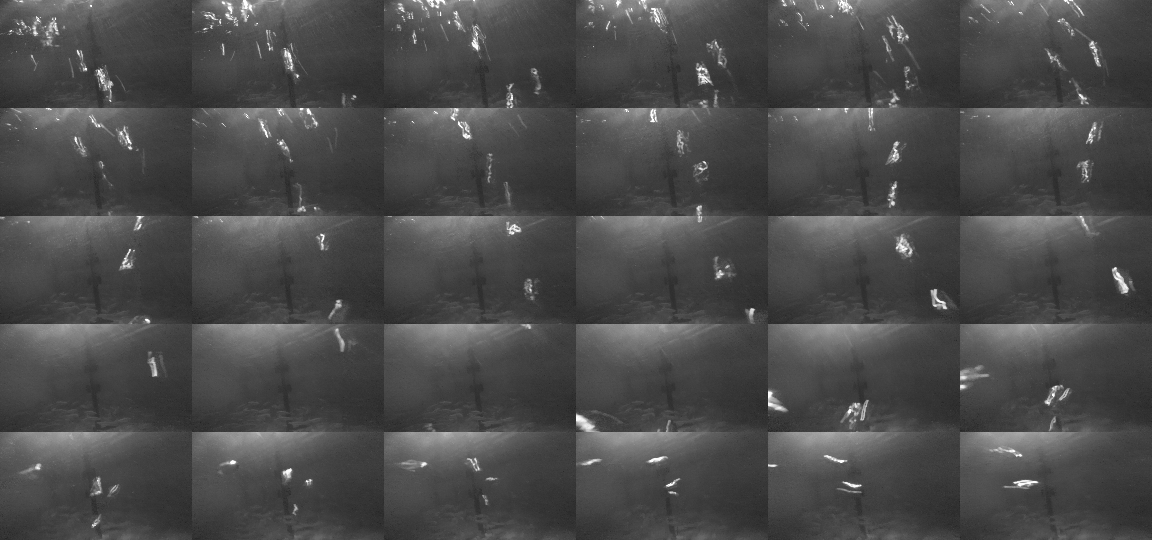
（4.1）

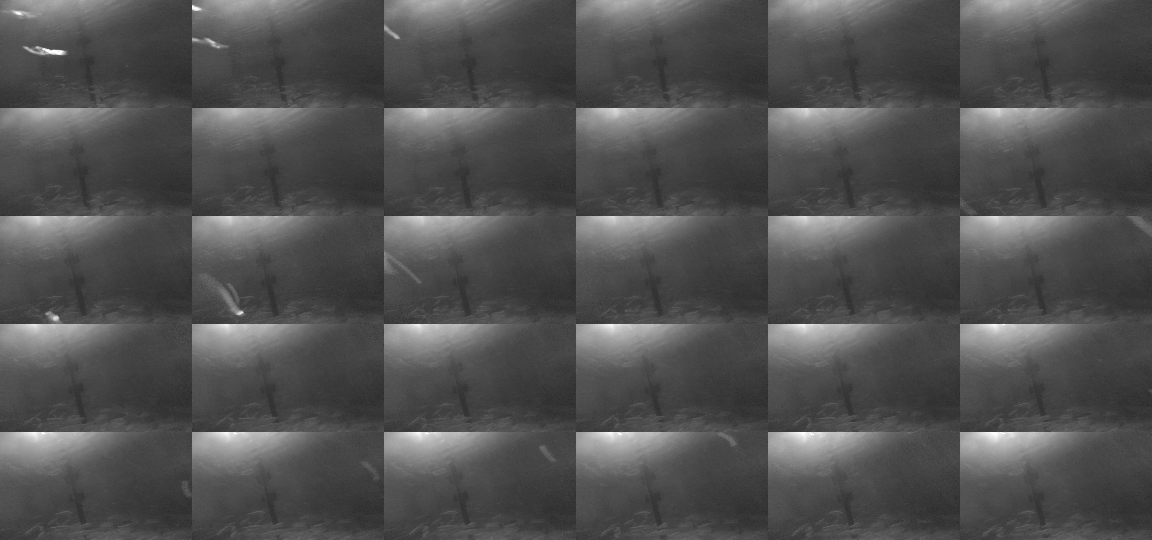
（4.2）

其中*l*为视频片段的长度，为像素值绝对值累加操作，为视频帧间差值，*x*为视频类别，为阈值。当帧间差值大于阈值时视频被分为第1类，当帧间差值小于阈值时视频被分为第2类。

如图4.1中展示了鱼类有规律巡游和运动剧烈的视频，在图4.2中通过计算两类视频的单帧间差值绘制了帧间差值图。视频*v1*的平均帧间差值为102，视频*v2*的平均帧间差值为89，当阈值设置为96时可以成功将两类视频划分开。

对比图4.1a和图4.1b可以得出结论，a图由于鱼类的剧烈运动造成大量的空气在水中使得视频画面变化较大所以其帧间差值较大，而图b中的鱼类运动平缓水中气泡较少，视频内容的变化也较小所以帧间差值较大。同时必须注意到的是，帧间差值法的运用前提条件是两个类别的视频变化程度不同，主要依赖于像素值的变化，当视频内容像素值变化不大的情况下是无法运用帧间差值法进行视频分类的。



（a）

(b)

图4.1 (a)视频*v1* 30帧画面(b) 视频*v2* 30帧画面

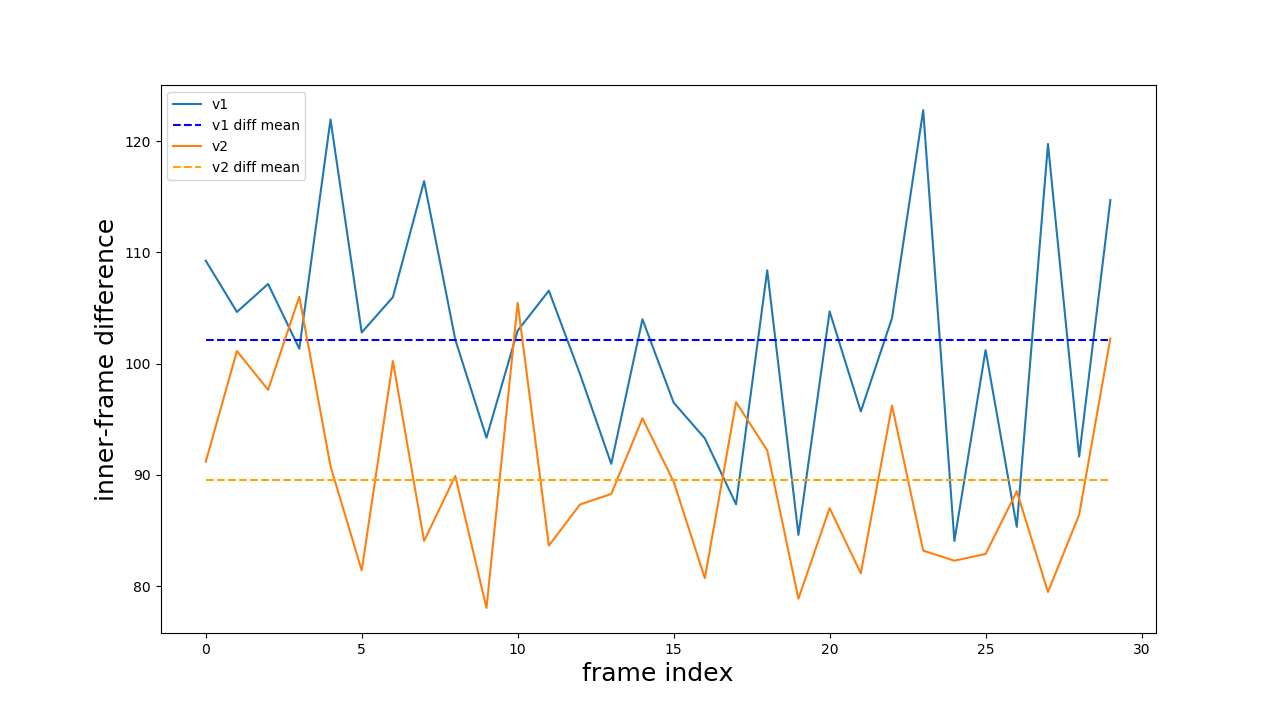


图4.2 帧间差值图

（2）基于光流的视频分类方法

光流(Optical flow or optic flow)是关于视域中的物体运动检测中的概念。用来描述相对于观察者的运动所造成的观测目标、表面或边缘的运动。

假设位于（x,y,t）的体素亮度是I（x,y,t）。该体素在两个图像帧移动距离为，则

假设移动很小，根据泰勒级数得出

其中，*H.O.T*为高阶无穷小，整理得到：

等式两边求微分：

（4.3）

其中为图像在*x*方向的梯度，为图像在*y*方向的梯度，为图像在帧间的梯度。显然公式4.3是欠约束方程，是无法求得确定解的。L-K方法假设两个相邻帧图像内容位移很小，且位移在所研究点q为中心的窗口内对所有的像素都成立。也就是说，局部图像光流（）满足：

记作

（4.4）

其中，

用最小二乘法求解公式4.4，

其中，n为研究像素点q附近所有的像素点的个数。

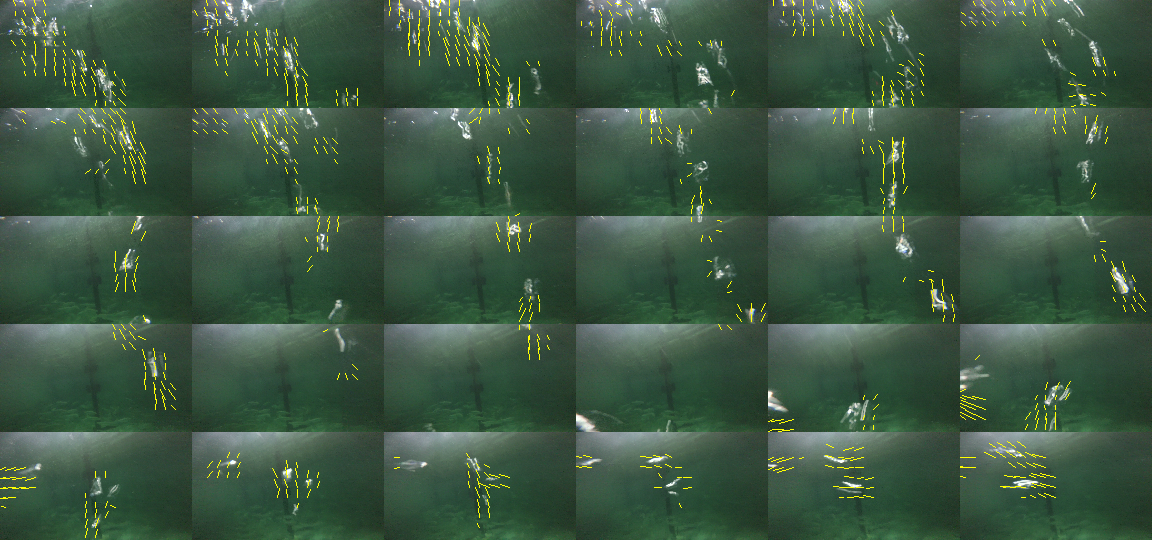
所谓基于光流法的视频分类就是根据视频光流强度的大小对视频内容进行分类。当视频画面中运动物体多，运动速度高，与环境差异大时其光流值大；反之则光流值小。通过设置合理阈值则可以将不同运动程度的视频进行分类，其数学表达如下：

（4.5）

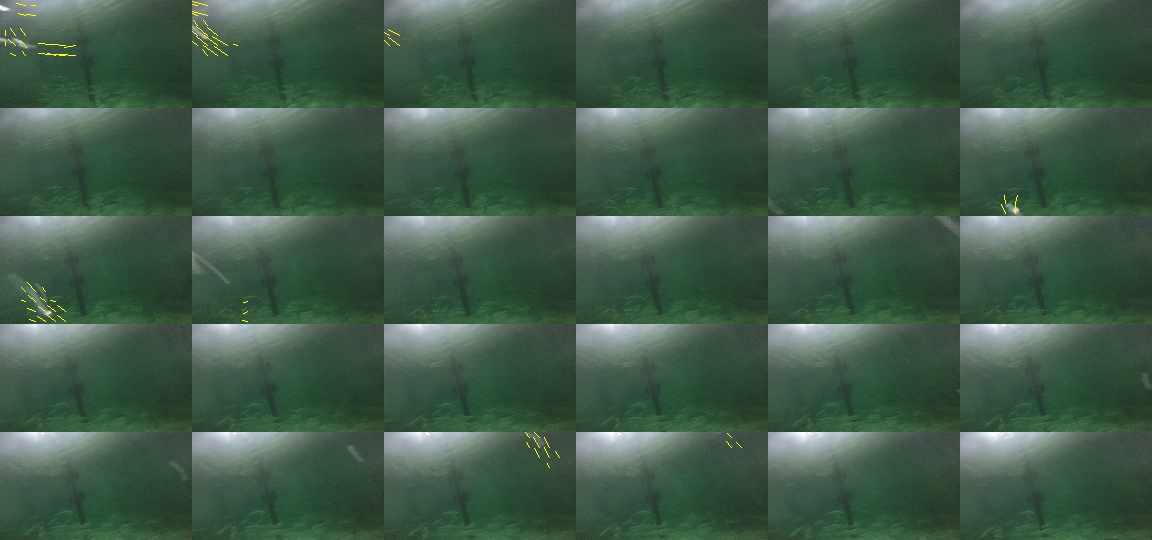
（4.6）

其中，为视频平均光流强度；*l*为视频片段长度；为按像素求和操作；为光流视频强度阈值。

计算视频片段*v1*（图4.1a）与视频片段*v2*（图4.1b）光流值，计算结果如图4.3所示。



(a)



(b)

图4.3 （a）视频*v1*光流计算结果；（b）视频*v2*光流计算结果。其中黄色的箭头代表盖像素点的光流方向与光流值。

通过计算得出图4.3a的视频平均光流强度，图4.3b的视频平均光流强度，当时可以将两类运动视频进行区别。

4.1.2. 基于深层特征的视频分类

通过视频的浅层特征可以将一些视频变换明显（帧差法）或者视频运动速度差异大（光流法）等类别的视频进行分类。但是他们存在的明显问题是，浅层特征容易受到像素值的干扰而无法准确反映视频信息。在图4.2中的单帧差值图十分容易受到像素值比较高的水泡影响，水泡区域像素的灰度值约为209而鱼群区域像素的灰度值只有70左右，即意味着水泡区域对帧间差值的贡献是鱼类对其贡献的3倍左右。而在光流法中，虽然提取了视频的运动特征可以客观的反映视频中物体的运动，但是在图4.3可以看到由于水泡的运动速度较快而鱼类相对水泡运动速度较慢所以整体上是在评估水中气泡的运动状态而非本论文想要研究的鱼类运动状态。

相比浅层特征，深层特征对视频内容特征具有更强的针对性。如前所述，浅层特征的视频分类方法实际上是根据视频中的水泡状态对视频进行分类，而这些水泡在鱼类的摄食行为研究中属于次要研究对象，甚至在有的情况下属于干扰信号，会影响到对鱼类摄食行为的判断。鱼类摄食行为视频分类的理论依据是根据鱼类在不同视频中的运动姿态，速度和聚集状态的不同实现对视频的分类。而基于深层特征的视频分类正是对三个指标的特征的提取。

1. 深度卷积神经网络

在视频分类领域中，大量的机器学习算法被提出，其中最为著名的两个方法分别是3D卷积[49][52]和双流卷积网络[50][51]。如图4.5所示，3D卷积引入一个三维卷积核，每次卷积输入多帧视频通过深度学习最后输出视频的类别。而Two-stream Convnets（双流卷积网络）结构如图4.4所示，双流卷积网络引入了两个卷积网络分别输入视频帧和图像运动信息，网络最后输出视频类别。两个方法在识别手势以及人体姿势的数据集上都取得了不错的成绩。

但是这两类方法均需要在大型的数据集上进行长时间的训练，除此之外，3D卷积会引入大量的参数并且使得网络训练困难，而双流卷积网络则需要提前计算视频运动信息（光流）使得计算量变的十分巨大。所以基于深度学习的视频行为分类实际上是将图像帧内容通过卷积网络将视频信息转换到卷积后的特征空间，然后通过神经网络的非线性分类方法进行视频分类。3D卷积网络是直接对时间和空间信息进行卷积得到最后分类所需的特征空间，而双流卷积网络是通过光流获取视频的运动信息单帧视频内容获取视频静态信息，是两种不同的特征转换方法。相比之下，3D卷积网络特征提取更为全面，但是其局限性在于仅限于短时间的视频分类动作，因为囿于现阶段的计算能力和内存数量是无法将3D卷积的空间维度无限延长的。双流卷积网络则可以实现对上长时间的视频行为分类。

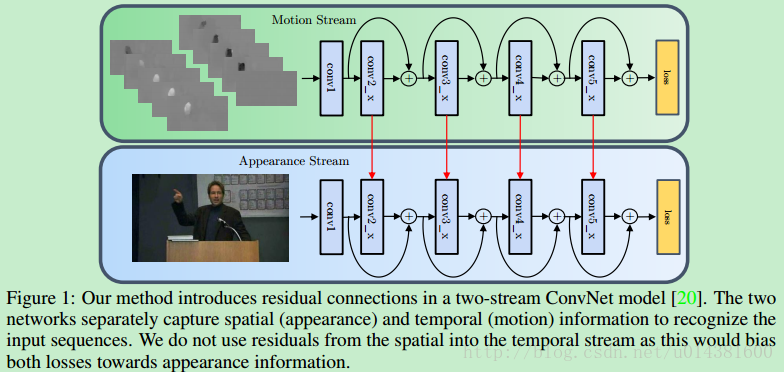


图4.4 双流卷积网络结构示意图

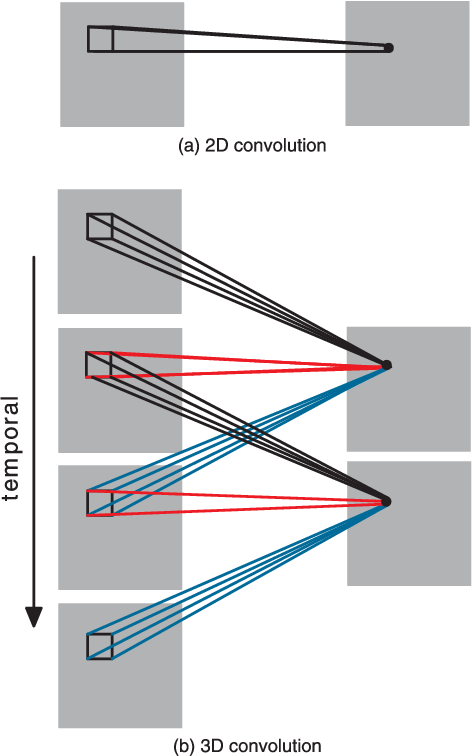


图4.5 3D卷积结构示意图

1. 动态混合纹理

与深度卷积神经网络不同，动态混合纹理建立了更加清晰可解释的数学模型描述视频内容，动态混合纹理的基本内容介绍如下。

首先，动态混合纹理假设视频的内容是通过多种具有二阶稳态的图像纹理组成，这些纹理的不断变化构成了视频的主要内容。视频中某一种纹理的变化的形成可以描述为，在t时刻存在一种隐含空间状态，则在该时刻的纹理状态可以通过一个发射矩阵C和观测误差线性变换得到，这个纹理被定义为观测值。而下一时刻(t+1)的观测值同样可以通过得到，是由通过一个状态转移矩阵A和状态转移误差通过线性变换得到。其数学模型描述如下：

在混合动态纹理中，视频的观测内容是通过所有不同的纹理通过概率的方式进行组合而得到，所以纹理的数量需要提前确定，即在动态混合纹理中纹理数量是一个超参数。混合动态纹理的结构示意图如图4.6所示。其中z为动态纹理的数量，不同的纹理拥有不同的模型参数，通过图中的z选择对应的动态纹理参数。则完整的混合动态纹理可以描述为如下形式：

其中，为观测结果出现的概率，为第*j*个动态纹理在某个概率分布（一般可以假设为伯努利分布）下的分布。为第j个动态纹理在观测结果中的概率。,,,为各个纹理对于的模型参数。最后通过极大似然估计和EM算法可以求得模型参数。



图4.6 混合动态纹理结构示意图

将混合动态纹理运用于视频分类的主要步骤分为两步：一、训练模型，针对不同类别的视频数据分别训练不同的混合动态纹理模型。二、通过输入视频t时刻的观测结果到模型可以求得进而得到，最后实现对的预测，比较模型输出的与目标视频中的观测结果，判断该视频的类别为模型预测值和实际值最相近的类别。

基于混合动态纹理的视频分类的优点在于存在清晰的数学模型，具有可解释性，但是其缺点在于更加适合对二阶稳态的纹理性极强的视频建模，当建模对象为非二阶稳态时，模型描述能力非常有限，无法准确描述真实的数据。如图4.7所示，混合动态纹理生成的观测结果与真实观测结果基本一致。而当训练视频内容是非二阶稳态时，模型生成效果较差。在图4.8中鱼类水下运动视频被用来训练混合动态纹理，但是可以看到，模型只能实现对背景部分的重建，而鱼类的运动情况无法通过混合动态纹理实现重建，所以运用混合动态纹理对非二阶稳态的视频数据进行分类时存在较大的局限性。



图4.7 混合动态纹理模型重建水面（左侧真实观测值，右侧模型生成观测值）



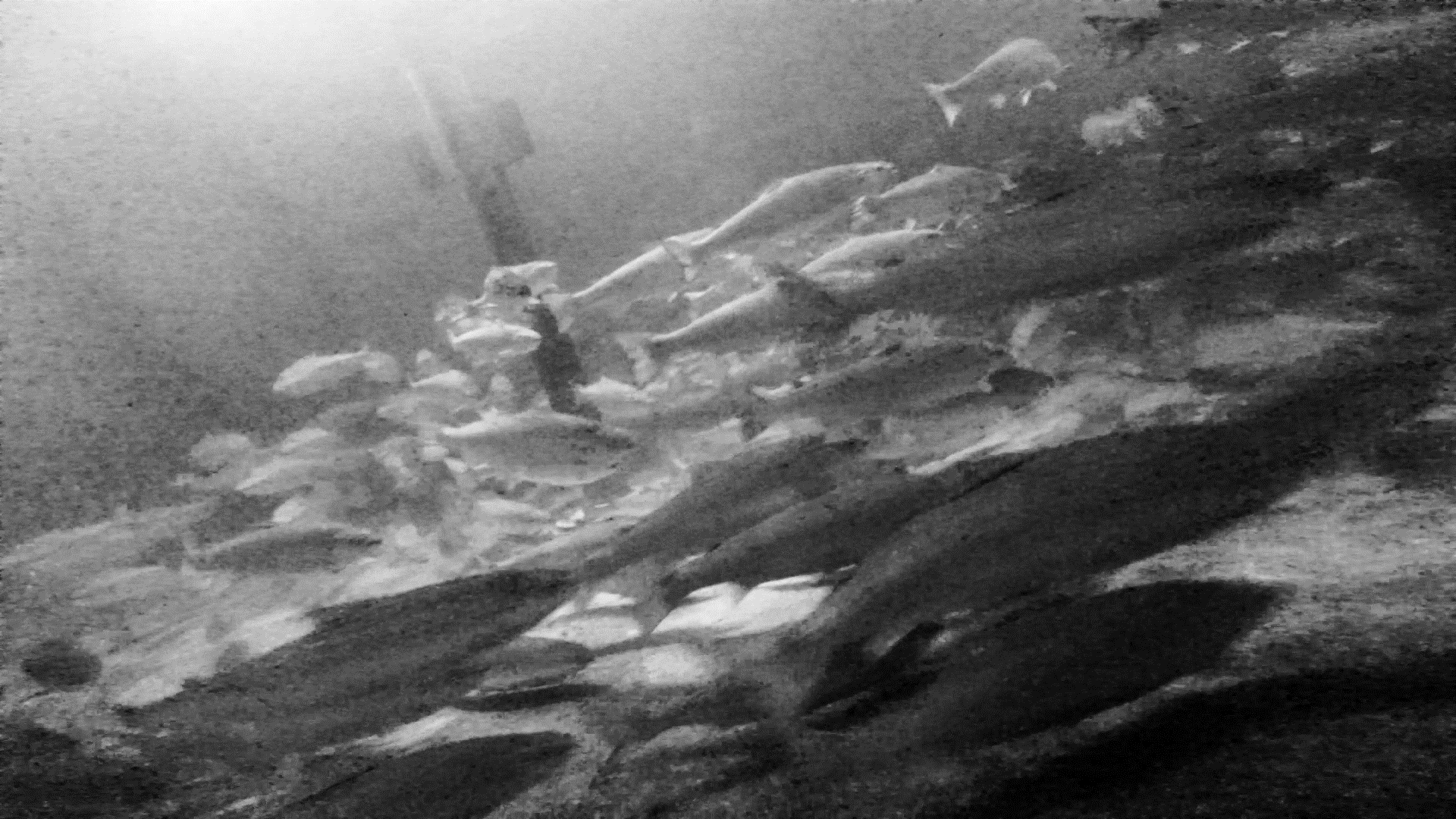
图4.7 混合动态纹理模型水下鱼类运动视频（上侧真实观测值，下侧模型生成观测值）

4.2.帧间关系贝叶斯估计网络视频分类

为准确描述帧间关系贝叶斯估计网络（IRBEN）视频分类方法，这里定义相关符号与表达。在t时刻的视频帧被记作,则已知当前帧为,帧之后的视频帧为的概率被记作。定义一个先验概率为标准多元高斯分布的概率空间z,则根据变分自动编码器是在已知图像空间下的采样值时在多元高斯空间下的概率分布。

通过概率方法实现视频分类主要有两种类别：1.基于单帧图像的视频分类，其算法步骤为首先选择视频中具有代表性的视频帧,针对不同类别的视频建立对应的概率模型（,…,,其中K为对应的视频类别），然后计算在所有概率模型下的概率值,视频类别为概率值最大的对于模型类别；2.基于多帧图像的帧间关系视频类别判断，其算法步骤为建立所有视频类别的后验概率模型,…),将目标视频相应帧输入到所有类别的后验概率模型中，目标视频的类别为概率值最大的类别。

基于单帧图像的视频分类，主要的应用场景是针对完全不同的视频内容进行分类，比如运动中的行人和运动中的汽车这类视频的分类。当视频分类的目标是是行走的行人和奔跑的行人时，这时采用单帧图像的视频分类是具有较大的挑战的，因为在整个视频片段中是难以选取出能代表视频片段的关键视频帧。所以相比之下基于多帧图像的视频分类具有更为广阔的应用前景。在图 4.8中展示了两个视频类别的视频帧截图，假设视频两个类别分别为鱼类有序运动和无序运动，其对应的概率模型分别为和，显然存在和的概率值关系。而对于基于单帧的视频分类方法，当被不幸的选择成为了关键帧则完全无法实现视频的分类。



a时刻视频帧（视频1与视频2相同）

时刻视频1的视频帧

时刻视频2的视频帧

图4.8 不同视频类别的视频帧展示

本论文所提出的IRBEN视频分类方法是一种基于多帧的帧间关系视频分类方法，其主要内容是采用条件概率完成视频分类。在多帧视频分类方法中其核心任务是完成对于不同视频类别的概率模型建立，及求取 其中为视频类别序号。但是在实际情况下与为图像矩阵其维度都十分高，直接建立这样的概率模型在计算上是无法完成的，受启发混合动态纹理，本论文提出帧间关系贝叶斯估计（IRBEN）视频分类方法假设存在某个维度较低的隐藏空间控制的生成，而变分自动编码器的解码器正是实现从一个多元高斯分布生成图像，而隐藏空间的建模则可以通过VAE中的编码器实现。而在变分自动编码器中隐藏空间的维度是根据数据要求设定的超参数，而且隐藏空间的维度一般都远小于视频帧的维度，则在隐藏空间中可以实现对视频帧间关系的建模。IRBEN空间关系转换以及隐藏空间帧间关系模型的建立如图4.9所示，隐藏空间下的变量用z表示并且和关系将通过一个全连接网络（）拟合。

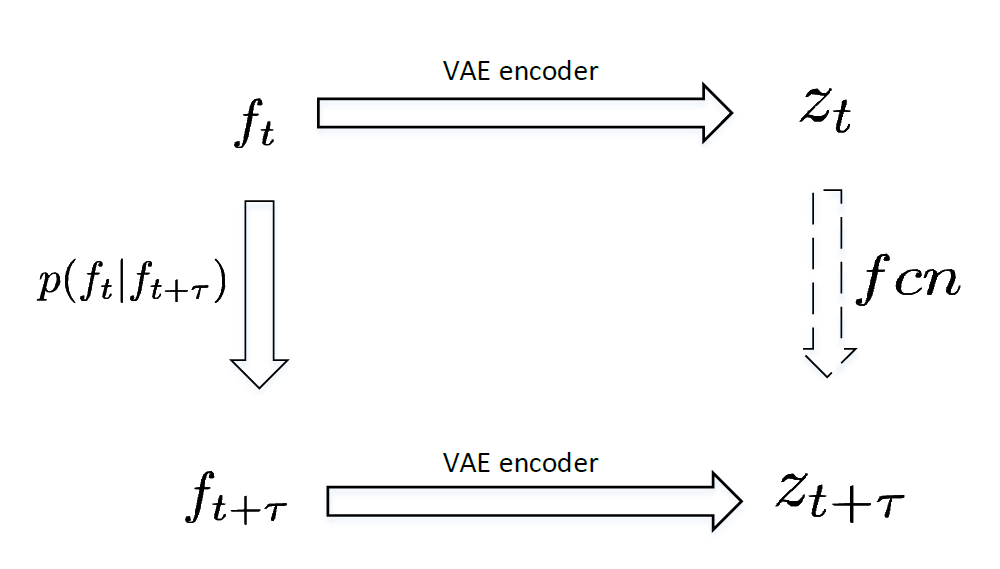


图4.9 视频空间变换以及隐藏空间下的帧间关系模型建立说明

IRBEN实际上是一个两阶段（Two-Stage）的视频分类算法，其结构如图4.10所示。第一步通过VAE变分自动编码器对图像进行空间映射，第二步完成在条件概率下隐藏空间的生成。模型训练过程也是分两步进行：1.VAE网络训练，在3.2章节中详细介绍了VAE模型的基本原理以及训练方法，这里不再赘述；2.隐藏空间全连接网络训练，下面详细介绍全连接网络的训练。



图4.10 IRBEN算法结构示意图

在隐藏空间下的其结构为多元高斯分布，即其变量为两组向量分别表示均值向量（）和方差向量（）。隐藏空间网络的目的是拟合从到的条件概率关系，即其训练的输入数据应该为对应视频类别下前后视频帧编码后的多元高斯分布，其结构被定义为两个独立结构但是具有相同损失函数的全连接神经网络。其结构如图4.11所示，其输入层为的均值向量和方差向量，其输入层为的均值向量和方差。

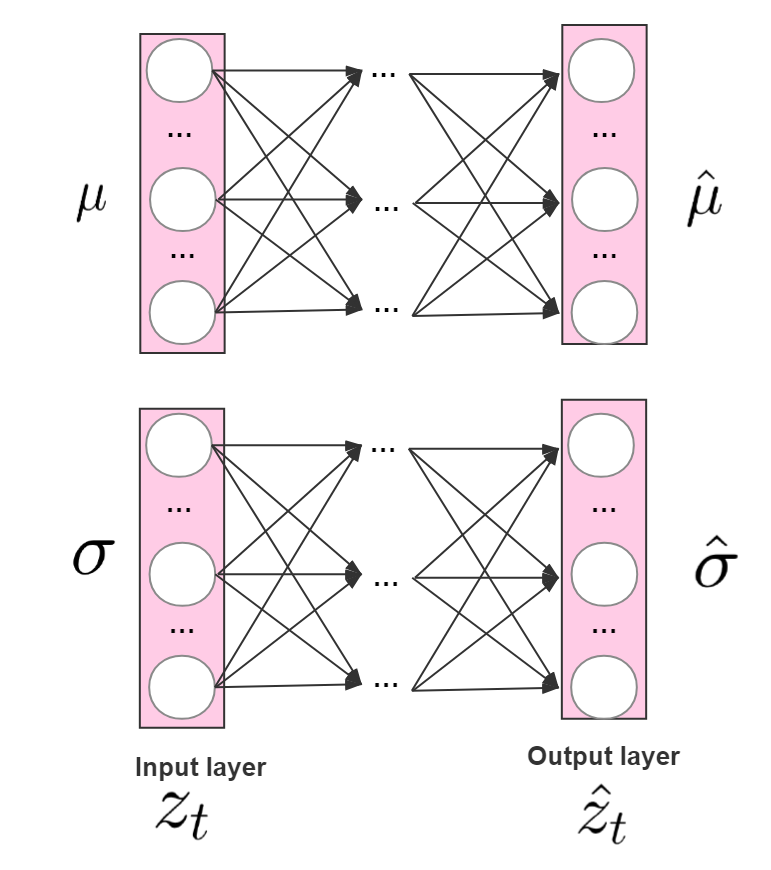


图4.11 隐藏空间全连接网络结构

全连接网络的训练损失函数被定义为两个多元高斯的散度距离，及网络的目的是希望尽可能的是其输入结果与真实的距离相近，多元高斯的KL散度距离计算参考附录1，其结果如下，

其中与为第j维的均值与方差，相同地与为的均值与方差；J为隐藏空间下多元高斯的维度。

为使得IRBEN方法能够考虑多帧视频的相关关系进行视频分类，本文根据公式4.7提出平均散度训练损失函数，其定义如下，

其中，为单个训练视频样本的长度。为最终训练损失函数。

将UVDASSB中的摄食和未摄食视频数据样本分别进行训练，其网络的参数设置如表4.1所示，得到和全连接网络用于鱼类的摄食行为分类。

完成所有类别的视频的全连接网络训练后即可进行鱼类摄食行为分类，算法流程如图4.12,其算法流程可描述为如下步骤：

1. 将目标视频的与输入到VAE中得到隐含空间的多元高斯分布,。
2. 将输入到完成训练的和得到隐藏空间预测值,。
3. 通过,与计算得到散度距离差并与进行比较最后得到分类结果。

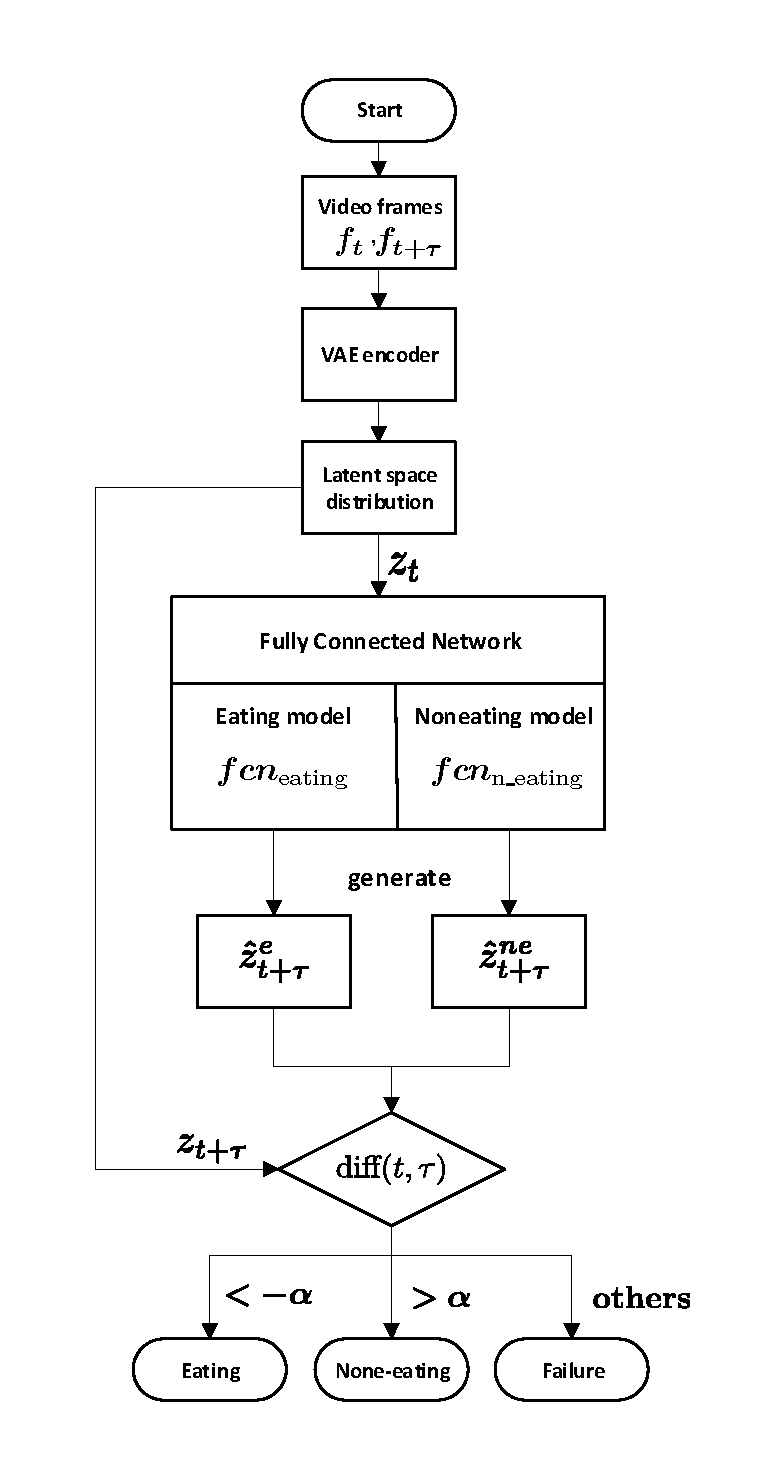


图4.12 IRBEN视频分类算法流程图。

图4.12流程图中定义了一个散度距离差的指标，其计算公式如下，

与训练时相同，为了使的视频的分类是基于一段视频而不是两帧视频，定义平均散度距离如下，

(4.9)

理论上，当时即可判断，该段视频为鱼类非摄食类别，但是为了提高算法的分类能力，本文设置了最小置信散度距离。带有最小置信散度距离的分类规则如下：



注意，其中为超参数，更加系统需要进行设置，且。

4.3.基于IRBEN方法的鱼类摄食行为分类实验

本节将详细介绍运用IRBEN方法对鱼类摄食行为进行分类的算法。实验目的是检验IRBEN方法在UVDASBB下对鱼类摄食行为进行分类的表现。视频有两个类别，分别是：摄食（eating）和未摄食（noneating）。IRBEN训练过程具有两个阶段，即VAE模型训练和模型训练。关于VAE模型训练的实验已经在3.2和3.3中进行了描述，所以本章只介绍模型的训练与IRBEN在UVDASSB上的实验结果。

4.3.1. 模型训练

首先设置神经元参数，在本实验中采用6层深度学习网络，其参数设置如表4.1所示。其他超参数设置参见表4.2。其训练的损失函数图如图4.13所示。其中蓝色的线代表的全连接网络在经过VAE编码后的隐藏空间下的损失函数图，相应的另外一条橙色的线指的是的损失函数图。可以看到两个类别的均在迭代2000次后收敛。

表4.1 全连接网络参数设置表



表4.2 其他超参数设置表

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 50 | 150 |

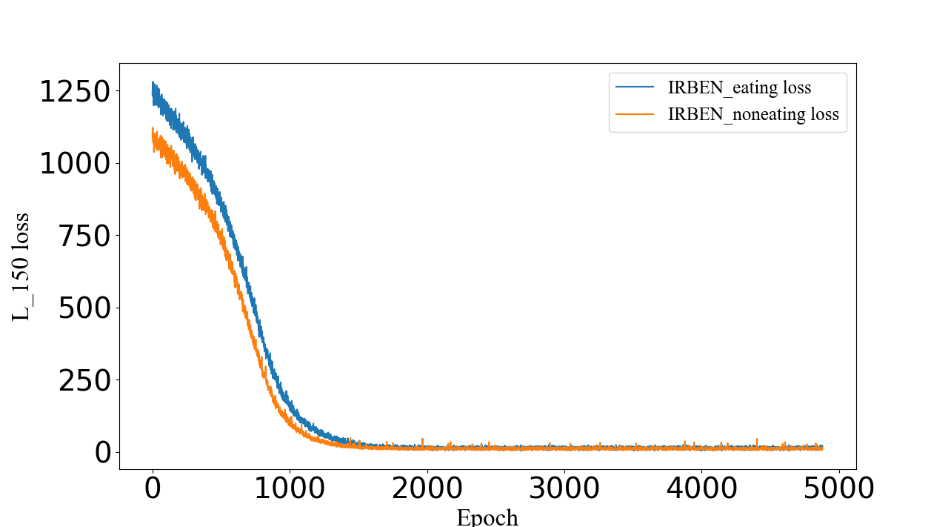
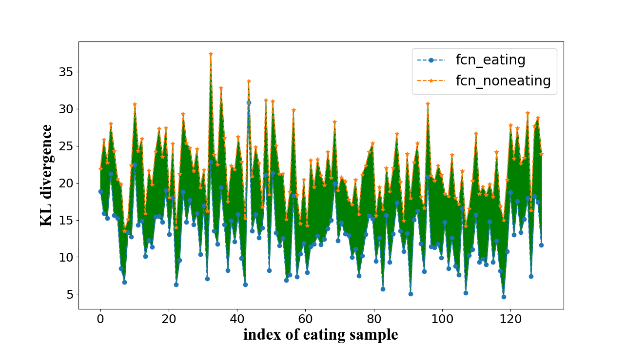
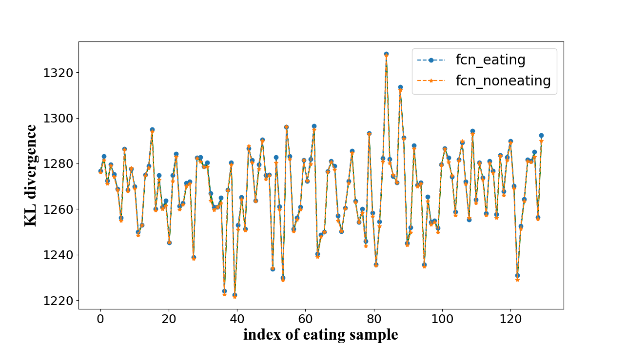


图4.13 全连接网络训练损失函数图

4.3.2. 模型训练评估

模型目标是预测视频下一帧在隐藏空间中的概率分布参数，为评估训练后的模型的预测能力，将136个摄食视频片段分别输入到训练过的模型中和随机初始化参数的模型中，然后分别计算其对应的平均置信度散度距离差，得到实验结果如图4.14所示，其中横轴为鱼类摄食样本序号，纵轴为平均散度距离差，图中的两条线分别代表样本在模型和下的实验结果。4.14a为未经训练的模型与训练后的模型即4.14b相比有两个显著的特点：1.纵轴数量级大，其原因在于由于随机初始化的参数预测得到的隐藏空间的概率分布也是随机的所以得到的散度距离必然会很大，最后导致平均散度距离差也较大；2.摄食样本在模型和模型下的平均散度距离差相近，其原因在于由于和都是随机初始化所以可以认为这两个模型实际上是一个模型所以得到的平均散度距离差必然也是相近的，而在图4.16b中可以看到两条曲线之间存在一个显著的“间隙”，形成间隙的原因正是因为视频类别不同导致，而且当视频类别差异越大形成的“间隙”会越大。



（a） 随机初始化参数的模型 （b）完成训练后的模型

图4.14 模型评估实验图

4.3.3. 不同值下在UVDASSB的分类结果

完成训练后将UVDASSB测试集数据输入到模型中，通过设置不同的最小置信距离并根据2.3.4节的评估标准进行评价得到的实验结果如表4.3所示。可以看到当时IRBEN在各项指标中都取得了较好的成绩并且优于同类方法（CNN方法[45]）。

表4.3 IRBEN不同在UVDASSB测试实验结果



4.4.本章小结