ICS 35.040

CCS L 71

团体标准

T/AI xxx.x—xxxx

|  |
| --- |
|  |

信息技术 媒体质量评价

第1部分：增强视频的编解码重建质量客观评价

Information Technology - Media Quality Assessment - Part 1: Objective video quality assessment for enhanced video compression

（征求意见稿）

（在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上）

XXXX - XX - XX实施

中关村视听产业技术创新联盟 发布

XXXX - XX - XX发布

目  录

[前言 II](#_Toc194321159)

[引言 III](#_Toc194321160)

[1 范围 1](#_Toc194321162)

[2 规范性引用文件 1](#_Toc194321163)

[3 术语和定义 1](#_Toc194321164)

[4 缩略语 1](#_Toc194321171)

[5 约定 1](#_Toc194321172)

[6 增强视频的编解码重建质量评价技术 9](#_Toc194321198)

[6.1 概述 9](#_Toc194321199)

[6.2 无参考视频客观质量评价技术方案 9](#_Toc194321200)

[6.3 全参考视频客观质量评价技术方案 17](#_Toc194321214)

[附录A （规范性） 增强视频的编解码重建质量评价技术实现 24](#_Toc194321223)

[附录B （规范性） 增强视频的编解码重建质量评价技术使用建议 25](#_Toc194321224)

前  言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件是T/AI XXX《信息技术 媒体质量评价》的第1部分。T/AI XXX已经发布了如下部分：

——第1部分：增强视频的编解码重建质量客观评价。

本文件由数字音视频编解码技术标准（AVS）工作组提出。

本文件由中关村视听产业技术创新联盟归口。

本文件起草单位：腾讯科技（深圳）有限公司、浙江大学、华为技术有限公司、天津科技大学、武汉大学、北京大学、国家广播电视总局广播电视科学研究院，北京邮电大学。

本文件主要起草人：王海强、张家琪、张立春、张滢雪、陈震中、贾川民、许晓中、刘杉、虞露、庄晓琪、郑月、赵寅、马思伟、石惠颖、雷正艺、谢嘉仪、宋莘鹏、王昊、郑蕤荻、闫石。

引  言

T/AI XXX旨在制定视频客观质量评价标准与主观质量测试规范，拟由3个部分构成。

——第1部分：增强视频的编解码重建质量客观评价。目的在于评价由视频增强及编解码过程所重建的视频质量。

——第2部分：视频质量主观评价。目的在于制定视频主观质量测试规范。

——第3部分：视频的编解码重建质量客观评价。目的在于评价视频的编解码重建质量。

本文件的发布机构提请注意，声明符合本文件时，可能涉及如下x项与视频质量评价技术相关的专利的使用。专利申请号及名称如下：

…

本文件的发布机构对上述专利的真实性、有效性和范围无任何立场。

上述专利持有人已向本文件的发布机构保证，愿意同任何申请人在合理且无歧视的条款和条件下，就专利授权许可进行谈判。上述专利持有人的声明已在本文件的发布机构备案，相关信息可以通过以下联系方式获得：

联 系 人：黄铁军（数字音视频编解码技术标准工作组秘书长）

通讯地址：北京大学理科2号楼2641室

邮政编码：100871

电子邮件：tjhuang@pku.edu.cn

电话：+8610-62756172

传真：+8610-62751638

网址：http://www.avs.org.cn

请注意除上述专利外，本文件的某些内容仍可能涉及专利。本文件的发布机构不承担识别专利的责任。

增强视频的编解码重建质量客观评价

1. 范围

本文件规定了视频客观质量评价技术，适用于评价由视频增强及编解码过程所重建的视频质量。

1. 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.1-2000 信息技术 词汇 第1部分：基本术语

GB/T 5271.34-2006 信息技术 词汇 第34部分:人工智能 神经网络

1. 术语和定义

参考视频 reference video

待编码的视频样本。



重建视频 reconstructed video

参考视频经某一视频编解码器编码后，由解码器根据位流解码得到的视频样本。



张量 tensor

一种神经网络中用来表示数据的基本数据结构。它是一种多维数组，可以表示向量、二维矩阵、高维矩阵等各种类型的数据。

1. 缩略语

下列缩略语适用于本文件。

|  |  |
| --- | --- |
| MOS | 平均意见分数（Mean Opinion Score） |

1. 约定
   1. 算术运算符

下列算术运算符适用于本文件。

表1 算术运算符定义

|  |  |
| --- | --- |
| 算术运算符 | 定义 |
|  | 加 |
|  | 减（二元运算符）或取反（一元前缀运算符） |
| 或或 | 乘 |
| ab | 幂，表示a的b次幂，也可表示上标 |
| 或a/b | 除法运算，不做取整或四舍五入 |
|  | 平方根 |
|  | 自变量取由到（含）的所有整数值时，函数的累加和 |
|  | 以2为底的对数 |
|  | 模运算，除以得到的余数，其中与都是正整数 |
|  | 二维互相关操作 |

* 1. 逻辑运算符

下列逻辑运算符适合于本文件。

表2 逻辑运算符定义

|  |  |
| --- | --- |
| 逻辑运算符 | 定义 |
| || | 逻辑或 |
| && | 逻辑与 |
| ! | 逻辑非 |

* 1. 关系运算符

下列关系运算符适用于本文件。

表3 关系运算符定义

|  |  |
| --- | --- |
| 关系运算符 | 定义 |
|  | 大于 |
| 或 | 大于或等于 |
|  | 小于 |
| 或 | 小于或等于 |
|  | 等于 |
|  | 不等于 |

* 1. 位运算符

下列位运算符适用于本文件。

表4 位运算符定义

|  |  |
| --- | --- |
| 位运算符 | 定义 |
| & | 与 |
| | | 或 |
| ~ | 取反 |
|  | 将a以2的补码整数表示的形式向右移b位。仅当b取正数时定义此运算。向右移至最高有效位时，其值与a移位运算前的最高有效位相等 |
|  | 将a以2的补码整数表示的形式向左移b位。仅当b取正数时定义此运算。向左移至最低有效位时，其值等于0 |

* 1. 条件运算符

下列条件运算符适用于本文件。

表5 条件运算符定义

|  |  |
| --- | --- |
| 条件运算符 | 定义 |
|  | 如果为真，则表达式为；如果为假，则表达式为 |

* 1. 赋值

下列赋值运算适用于本文件。

表6 赋值运算定义

|  |  |
| --- | --- |
| 赋值 | 定义 |
| = | 赋值运算符 |
|  | 取从至（含）的值，其中、和是整数 |
| ++ | 自加，++相当于=+1,当用于数组下标时，在自加运算前先求变量值 |
| -- | 自减，--相当于=-1,当用于数组下标时，在自减运算前先求变量值 |
| += | 自加指定值，例如，+=3相当于=+3，+=(-3)相当于=+(-3) |
| -= | 自减指定值，例如，-=3相当于=+(-3)，-=(-3)相当于=-(-3) |

* 1. 数学函数

以下数学函数运算符适用于本文件。

|  |  |
| --- | --- |
| ceil(x)或 | 向上取整，大于或等于x的最小整数 |
| floor(x)或 | 向下取整，小于或等于x的最大整数 |
| round(x) | 四舍五入取整 |
| clip(i,j,x) | 截断， |
| min(x,y) | 取最小值, |
| max(x,y) | 取最大值, |
| abs(x), |x| | 取绝对值， |
| sign(x) | 符号函数， |
|  | p范数， |

* 1. 张量运算
     1. 矩阵转置 matrix transpose

矩阵转置表示为Transpose(),

1. 输入二维数组input，维度为(h, w)；
2. 输出二维数组output，维度为(w, h)。

计算方法如下：

*output*[*j*,*i*] = *input*[*i*,*j*]; 0≤*i*<*h*, 0≤*j*<*w*.

* + 1. 矩阵相乘 matrix multiplication

矩阵相乘表示为，

1. 输入二维数组input1，维度为(h, L)；
2. 输入二维数组input2，维度为(L, w)；
3. 输出二维数组output，维度为(h, w)。

计算方法如下：

0≤*i*<*h*, 0≤*j*<*w*.

* + 1. 对应位置元素相加 per-element addition of tensors

对应位置元素相乘表示为，

1. 输入张量input1，维度为(C, h, w)；
2. 输入张量input2，维度为(C, h, w)；
3. 输出张量output，维度为(C, h, w)。

计算方法如下：

*output*[*c*,*i*,*j*]=*input*1[*c*,*i*,*j*]+*input*2[*c*,*i*,*j*]; 0≤*c*<*C*, 0≤*i*<*h*, 0≤*j*<*w*.

* + 1. 对应位置元素相乘 per-element multiplication of tensors

对应位置元素相乘表示为，

1. 输入张量input1，维度为(C, h, w)；
2. 输入张量input2，维度为(C, h, w)；
3. 输出张量output，维度为(C, h, w)。

计算方法如下：

*output*[*c*,*i*,*j*]=*input*1[*c*,*i*,*j*]⋅*input*2[*c*,*i*,*j*]; 0≤*c*<*C*, 0≤*i*<*h*, 0≤*j*<*w*.

* + 1. 变形reshape

变形表示为Reshape(C,h,w)，输入张量为input[Cin, hin, win]，输出张量为output[C, h, w]。输入张量和输出张量具有相同的元素数量，即Cin\*hin\*win = C\*h\*w。

计算方法如下：

对于*c*=0..*Cin*-1, *i*=0..*hin*-1和*j*=0..*win*-1,

*index* = *c⋅hin ⋅win* + *i⋅win* + *j*，

*output[floor(index/(h⋅ w)), floor(index%(h⋅w)/w, index%w)*=*input[c,i,j].*

* + 1. 填充 padding

填充表示为Padd(pver, phor)，输入张量为input[Cin, hin, win]，输出张量为output[Cout, hout, wout]，其中*hout= hin*+*2*\**pver，wout= win*+*2*\**phor*。

计算方法如下：

*output*[*c,i*,*j*] = ((*pver*≤*i*<*hin*+*pver*) && (*phor*≤*j*<*win*+*phor*))?*input*[*c,i*,*j*]:0, 0≤*c*<*C*, 0≤*i*<*hout*, 0≤*j*<*wout*.

* + 1. 中心裁剪 center crop

中心裁剪表示为CenterCrop(h, w)，输入张量为input[Cin, hin, win]，输出张量为output[C, h, w]，其中hhin，wwin。

计算方法如下：

1. 依据输入张量和输出张量的维度，计算裁剪的边界。

*top* = *round*((*hin*-*h*)/2),

*left* = *round*((w*in*-*w*)/2).

1. 按照边界对输入张量进行裁剪，

*output*[*c,i*,*j* ] = *input*[*c,i+top*,*j+left* ], 0≤*c*<*C*, 0≤*i*<*h*, 0≤*j*<*w*.

* + 1. 卷积层 convolutional layer

二维卷积表示为Conv(kverkhor, s, Cin, Cout, pver, phor)，输入张量为input[Cin, hin, win]，输出张量为output[Cout, hout, wout]。其中kver和khor表示卷积核的高和宽，s表示卷积步长，Cin表示输入张量的通道数，Cout表示输出张量的通道数，pver表示在高度方向对输入张量的填充量，填充的缺省值为0，phor表示在宽度方向对输入张量的填充量，填充的缺省值为0。

输出张量的维度计算方法为：

当且时，对于，，，

式中：

——权重张量，可学习参数；

偏置张量，可学习参数。

* + 1. 可变形卷积层 deformable convolution layer

二维可变形卷积表示为DeformConv(kverkhor, s, Cin, Cout, pver, phor)，输入张量为input[Cin, hin, win ]，输出张量为output[Cout, hout, wout ]。其中kver和khor表示卷积核的高和宽，s表示卷积步长，Cin表示输入张量的通道数，Cout表示输出张量的通道数，pver表示在高度方向对输入张量的填充量，phor表示在宽度方向对输入张量的填充量。

输出张量的维度计算方法为：

,

当且时，对于，

式中：

；

；

——位置偏移张量，可学习参数；

——权重张量，可学习参数；

——偏置张量，可学习参数。

* + 1. 归一化层 normalization layer
       1. 批量归一化batch normalization

批量归一化表示为BatchNorm( )，计算方法如下：

，

式中：

，——维度为C的可学习张量；

——输入张量在通道维度的均值；

——输入张量在通道维度的方差。

* + - 1. 层归一化层layer normalization

层归一化表示为LayerNorm( )，计算方法如下：

,

式中：

，——维度为[C, h, w]的可学习张量；

，——输入张量的均值和方差。

* + - 1. 均方根层归一化层root mean square layer normalization

均方根层归一化表示为RMSNorm( )，计算方法如下：

,，

式中是维度为[C, h, w]的可学习张量。

* + 1. 池化层 pooling layer
       1. 最大值池化 max pooling

最大值池化表示为MaxPooling(kverkhor, s)，输入张量为input[C, hin, win]，输出张量为output[C, hout, wout]。其中kver和khor表示池化窗口的高和宽，s表示池化窗口的滑动步长。计算方法如下：

对于，，，

.

* + - 1. 平均值池化 average pooling

平均值池化表示为AvgPooling(kverkhor , s)，输入张量为input[C, hin, win ]，输出张量为output[C, hout, wout ]。其中kver和khor表示池化窗口的高和宽，s表示池化窗口的滑动步长。计算方法如下：

对于，，，

.

* + 1. 全连接层 fully connected layer

全接层表示为FC(Nin, Nout)，输入张量为input[Nin]，输出张量为output[Nout]。其中Nin表示输入张量的特征数量，Nout表示输出张量的特征数量。计算方法如下：

对于，，

，

式中：

——权重张量，可学习参数；

——偏置张量，可学习参数。

* + 1. 激活层 activation layer
       1. 线性整流单元 rectified linear unit

线性整流单元表示为ReLU( )，输入张量为input[C, h, w]，输出张量为output[C, h, w]。计算方法如下：

对于，，，

* + - 1. SiLU整流单元 sigmoid linear unit

SiLU整流单元表示为SiLU( )，输入张量为input[C, h, w]，输出张量为output[C, h, w]。计算方法如下：

对于，，，

.

* + 1. 残差块 residual block

残差块表示为ResBlock(C)，其结构如图1所示，包括2个卷积层Conv、2个归一化层BatchNorm和2个激活函数ReLU。输入的特征向量经卷积层Conv、归一化层BatchNorm、激活函数ReLU、卷积层Conv、归一化层BatchNorm依次处理后，再将所得特征向量与输入的特征向量相加，经过激活函数ReLU得到残差块的输出特征向量。

ReLU( )

ReLU( )

Conv(3x3,1,C, C,1,1)

Conv(3x3,1,C,C,1,1)

BatchNorm( )

BatchNorm( )

Input

Output

图 1 残差块结构图

* + 1. 下采样残差块 downsample residual block

下采样残差块表示为DownResBlock(Cin, Cout)，其结构如图2所示，其包括3个卷积层Conv、3个归一化层BatchNorm和2个激活函数ReLU。输入的特征向量经主路径中卷积层Conv、归一化层BatchNorm、激活函数ReLU、卷积层Conv、归一化层BatchNorm依次处理得到主路径的特征向量，同时输入的特征向量经副路径中的卷积层Conv和归一化层BatchNorm处理后得到副路径的特征向量，然后将两路径所得特征向量相加后再经过一个激活函数ReLU，最终得到下采样残差块的输出特征向量。

ReLU( )

ReLU( )

Conv(3x3,2,Cin,Cout,1,1)

Conv(3x3,1, Cout, Cout,1,1)

BatchNorm( )

BatchNorm( )

Input

Output

Conv(1x1,2, Cin,Cout,0,0)

BatchNorm( )

图 2 下采样残差块结构图

* + 1. 级联 concatenation

两个张量级联表示为Concat(C1, C2)，该过程是将两个具有相同空间大小的张量按照指定的顺序在通道维度进行拼接。

1. 输入张量input1，维度为(C1, h, w)；
2. 输入张量input2，维度为(C2, h, w)；
3. 输出张量output，维度为(C1+C2, h, w)。

计算方法如下：

*output*[*c*,*i*,*j*]=(*c*<*C*1)?*input*1[*c*,*i*,*j*]:*input*2[*c*,*i*,*j*], 0≤*c*<*C*1+ *C*2, 0≤*i*<*h*, 0≤*j*<*w*.

1. 增强视频的编解码重建质量评价技术
   1. 概述

本文件规定两个视频客观质量评价技术方案，包括一个无参考视频客观质量评价技术方案和一个全参考视频客观质量评价技术方案。其中，无参考视频客观质量评价技术方案适用于对单个视频作为输入时进行质量评价的场景。全参考视频客观质量评价技术方案适用于在参考视频和重建视频同时作为输入时进行质量评价的场景。

* 1. 无参考视频客观质量评价技术方案
     1. 技术框架

无参考视频客观质量评价技术方案系统框架构见图3，包括视频预处理模块、文本特征提取模块、视觉特征提取模块、视觉特征降维模块、多模态质量预测模块和质量融合模块，各个模块详细阐述如下。

Graphical user interface, text, application, chat or text message

Description automatically generated

图 3 无参考视频客观质量评价技术方案系统架构图

* + 1. 视频预处理模块

视频预处理模块对输入的视频帧进行色彩空间转换处理，输出为RGB格式的视频，维度为[N, 3, W, H]。 其中N为视频的帧数，3为通道数量, W为视频水平方向像素数量，H为视频垂直方向像素数量。再对每个视频帧进行填充、上采样、下采样、裁剪等处理，输出的视频帧维度为[N, 3, 448, 448]。

* 输入：待评估的视频
* 输出：预处理后RGB格式视频，维度为[N, 3, 448, 448]。
  + 1. 视觉特征提取模块
       1. 概述

Text

Description automatically generated

图 4 视觉特征提取模块结构图

该模块包含一个视频嵌入模块、一个第一类Transformer模块和一个层归一化层。其中第一类Transformer模块由一系列级联的数量可配置的第一类Transformer层组成，其结构见图4。预处理后的视频帧经视觉特征提取模块处理后得到第三视觉特征。

* + - 1. 视频嵌入模块

该模块把预处理后的视频帧嵌入为适合神经网络处理的特征。该模块包含一个卷积层，一个级联层和一个层归一化层，其结构见图5。

Graphical user interface, text, application, chat or text message

Description automatically generated

图 5 视频嵌入模块结构图

其中，卷积层的卷积核大小为14x14，步长为14，可学习参数中不包含偏置项。输入通道数为3，输出通道数量为1024。预处理后的每个视频帧经过卷积层处理后，得到的特征维度为1x1024x32x32。其中，1对应所处理的单个视频帧，本技术方案进行每次神经网络推理所处理的视频帧数量为3，因此输出的特征维度为3x1024x32x32。

进一步地，将该特征同一通道的特征进行维度变换处理，得到维度为3x1024x1024的特征。该特征级联一个可学习的维度为3x1x1024的类别编码特征，经扩充后的特征维度为3x1025x1024。

之后，经级联得到的特征与一个可学习的维度为3x1025x1024的位置编码特征相加。相加后的特征经过一个层归一化层处理，得到第一视觉特征。

* 输入：预处理的视频帧，维度为[3,3,448,448]
* 输出：第一视觉特征VisualFeature1，维度为[3,1025,1024]
  + - 1. 第一类Transformer模块

该模块对第一视觉特征提取深度视觉特征。该模块包含一系列级联的第一类Transformer层。第一类Transformer层包含一个自注意力层和一个残差层。其结构见图6。

Graphical user interface, text, application

Description automatically generated

图 6 第一类Transformer层结构

其中，自注意力层使用三个全连接层对输入特征进行处理，得到Q、K、V三个矩阵。之后，将Q与K的转置矩阵做矩阵乘法得到自注意力矩阵，该矩阵经过Softmax层处理后与V做矩阵的乘法，经过全连接后与该层的输入特征相加，即得到自注意力层的输出。

之后，自注意力层的输出经过一个残差层处理得到第一类Transformer层输出，该残差层对输入的特征依次进行层归一化层、全连接层、激活函数ReLU和另外一个全连接层进行处理。最后一层全连接层的输出与自注意力层的输出相加后即得到第一类Transformer层的输出。

* 输入：第一视觉特征VisualFeature1，维度为[3,1025,1024]
* 输出：第二视觉特征VisualFeature2，维度为[3,1025,1024]
  + - 1. 层归一化

预处理的视频帧经过视频嵌入模块和第一类Transformer处理后，再经过一层归一化层处理，即得到视频帧的视觉特征。该归一化层为LayerNorm层。

* 输入：第二视觉特征VisualFeature2，维度为[3,1025,1024]
* 输出：第三视觉特征VisualFeature3，维度为[3,1025,1024]
  + 1. 视觉特征降维模块
       1. 概述

该模块对第三视觉特征进行降维，包含一系列级联的数量可配置的第二类Transformer层、全连接层和级联层，其结构见图7。

Graphical user interface

Description automatically generated

图 7 视觉特征降维模块结构

该模块的输入包含第三视觉特征VisualFeature3和两个可学习的tokens。其中，可学习的tokens1与第三视觉特征作为第一个第二类Transformer层的输入。第二个和之后的第二类Transformer层的输入包含第三视觉特征和上一个第二类Transformer层的输出。最后一个第二类Transformer层的输出经过一个全连接层处理后，再与可学习的tokens2进行级联得到视觉降维模块的输出。

* + - 1. 第二类Transformer模块

该模块包含一系列级联的第二类Transformer层。第二类Transformer层包含一个互注意力层和一个残差层。其结构见图8。该模块的输入包含两部分，第一部分为第三视觉特征VisualFeature3，第二部分为可学习的tokens1或者上层第二类Transformer层的输出。

* 输入：第三视觉特征向量，维度为[3, 64,1024]
* 输出：第四视觉特征向量，维度为[3, 64,1024]

A screenshot of a phone

Description automatically generated with medium confidence

图 8 第二类Transformer层结构

* + - 1. 全连接层与级联

最后一个第二类Transformer层的输出维度为3x64x1024，经过一个全连接层处理后，维度变为3x64x4096。再与一个可学习的维度为3x1x4096的tokens2进行级联处理，得到维度为3x65x4096的特征，进一步将该特征维度转换为1x195x4096，作为视觉特征降维模块的输出。

* 输入：第四视觉特征，维度为[1,1025,1024]
* 输出：第五视觉特征，维度为[1,195,4096]
  + 1. 文本特征提取模块
       1. 概述

该模块对输入的文本指令进行特征提取，包含一个分词器模块和一个文本嵌入模块。其结构见图9。输入文本指令包含提问文本提示指令，如How would you rate the quality of this video? 和回答文本提示指令，如The quality of this video is \_\_\_\_，其中下划线部分用于提示语言预测模块进行回答来预测输入对应的质量文本描述。

A picture containing text

Description automatically generated

图 9 文本嵌入模块结构

* + - 1. 分词器模块

分词器模块对输入的文本指令进行分词处理，把文本句子分成若干个分词。再通过查询分词表得到每个分词的索引。

* 输入：输入文本指令，包括提问文本提示指令和回答文本提示指令
* 输出：分词后的提示指令文本，维度为1\*M，其中M取决于具体的输入文本指令
  + - 1. 文本嵌入模块

对每个分词所对应的索引，通过嵌入向量矩阵得到每个分词所对应的文本特征向量。本提案中所使用的分词库大小为32000，每个分词特征维度为4096。

* 输入：分词后的提示指令文本，维度为1\*M
* 输出：文本嵌入特征，维度为1\*M\*4096
  + 1. 多模态质量预测模块
       1. 概述

该模块用于预测输入视频的质量描述。它包含一个第三类Transformer模块和一个全连接层。该模块的输入为级联后的文本嵌入特征和第五视觉特征。其结构见图10。

Text

Description automatically generated

图 10 多模态质量预测模块结构

* + - 1. 第三类Transformer模块

该模块包含一系列可配置的第三类Transformer层。其中，第三类Transformer层在每个自注意力层对输入的不同模态的特征分别进行处理，因此具有多模态内容的处理和融合能力。具体地，自注意力层中K、V矩阵对第三视觉特征向量和文本嵌入特征向量进行处理，Q同时包含第三视觉特征向量和文本嵌入特征。其结构见图11。

Graphical user interface, application

Description automatically generated

图 11 多模态质量预测模块结构

* + - 1. 全连接层

该层用于预测输入视频质量的文字描述和对应的可能性大小。具体地，该层的输出对所有可能的分词，分别预测每个分词用于描述输入视频质量的可能性。

* 输入：级联后的第三视觉特征向量和文本嵌入特征向量，维度为[1, (195+M) ,4096]
* 输出：预测的质量文字描述概率向量，维度为[1,1,32000]
  + 1. 质量融合模块
       1. 概述

该模块用于预测视频的质量分数，包括质量转换模块用于将预测的质量文字描述概率向量转换为质量分数，和质量融合模块用于融合视频局部质量分数。

* + - 1. 质量转换模块

本方案采用以下7个等级用于文字描述视频的感知质量，["excellent", "high", "good", "acceptable", "fair", "poor", "bad"]，并记它们所对应的质量分数向量为[5.155906242605395,4.79629981305803,4.586142074626978,4.322160456775861,3.8911754755800505,3.1810649702958242,2.106344504594364]。

对于每一次质量预测，首先从质量文字描述概率向量中查询上述7个等级文字所对应的概率，进一步和所对应的质量分数向量进行加权平均，得到输出即为该次质量预测的MOS分数。

* 输入：预测的质量文字描述概率向量，维度为[1,1,32000]
* 输出：预测的局部质量分数，维度为1
  + - 1. 质量融合模块

经一次推理所得到的预测值为该次推理所处理的视频帧所对应的局部质量分数。对输入视频所对应的所有局部质量分数取平均操作即得到待评估视频的质量分数。

* 1. 全参考视频客观质量评价技术方案
     1. 技术框架

全参考视频客观质量评价技术方案系统框架构见图12，包括视频预处理模块、空域视觉特征提取模块、时域视觉特征提取模块、空域质量预测模块、时域质量预测模块和质量融合模块。各个模块详细阐述如下。

时域质量

预测模块

空域视觉特征提取模块

视频预处理模块

视频预处理模块

参考视频

重建视频

时域视觉特征提取模块

时域视觉特征提取模块

空域质量

预测模块

质量融合

模块

视频质量分数

图 12 全参考视频客观质量评价技术方案系统架构图

* + 1. 视频预处理模块

该模块对输入的视频进行色彩空间转换处理，输出一个RGB格式视频作为空域视觉特征提取模块的输入，其维度为[N, 3, W, H]，以及一个仅包含亮度分量的灰度视频作为时域视觉特征提取模块的输入，其维度为[N, 1, W, H]。其中W为输入视频水平方向像素数量，H为输入视频垂直方向像素数量，N为输入视频的帧数。

* 输入：重建视频或参考视频
* 输出1：预处理后的RGB格式视频，维度为[N, 3, W, H]
* 输出2：预处理后的灰度格式视频，维度为[N, 1, W, H]
  + 1. 空域视觉特征提取模块
       1. 概述

该模块包含两个基于卷积神经网络的嵌入模块，每个嵌入模块后接四个卷积模块，相同尺度的卷积模块输出特征经过特征融合模块融合，并输入至重建视频帧特征提取路径中的下一个卷积模块，其结构见图13。RGB格式参考视频帧和RGB格式重建视频帧经该模块处理后得到空域视觉特征向量。

卷积模块一

嵌入模块

RGB格式参考视频帧

卷积模块三

卷积模块二

卷积模块四

卷积模块一

嵌入模块

RGB格式重建视频帧

卷积模块三

卷积模块二

卷积模块四

空域视觉特征向量

特征融合模块1

特征融合模块2

特征融合模块3

特征融合模块4

图 13 空域视觉特征提取模块流程图

* + - 1. 嵌入模块

嵌入模块包含一个卷积层、一个归一化层、一个激活函数和一个池化层。结构如图14所示。

ReLU( )

Conv(7x7,2,3,64,3,3)

MaxPooling(3x3,2)

BatchNorm( )

VisualFeature1

Frame

图 14 嵌入模块结构图

卷积层的卷积核大小为7x7，步长为2x2，填充大小为3x3，输入通道数为3，输出通道数为64。归一化层为BatchNorm，激活层为ReLU，池化层为MaxPooling，滑动窗口大小为3x3，步长为2x2。RGB格式的参考视频帧和重建视频帧经过该模块处理后，得到各自对应的第一视觉特征。

* 输入：预处理后的RGB格式视频帧
* 输出：第一视觉特征VisualFeature1，维度为[64, , ]
  + - 1. 卷积模块一

该模块对第一视觉特征向量进行处理，提取深度视觉特征。卷积模块一的结构如图15所示，由2个级联的残差块组成。

ResBlock(64)

ResBlock(64)

VisualFeature1

VisualFeature2

图 15 卷积模块一结构图

每个残差块中的2个卷积层的卷积核大小均为3x3，步长均为1x1，填充大小均为1x1，输入输出通道数量均为64，归一化层为BatchNorm，激活层为ReLU。

输入的第一视觉特征向量经2个级联的残差块处理后得到第二视觉特征。

对于参考视频特征提取分支：

* 输入：参考视频第一视觉特征VisualFeature1，维度为[64, , ]
* 输出：参考视频第二视觉特征VisualFeature2，维度为[64, , ]

对于重建视频特征提取分支：

* 输入：重建视频第一视觉特征VisualFeature1，维度为[64, , ]
* 输出：重建视频第二视觉特征VisualFeature2，维度为[64, , ]
  + - 1. 卷积模块二

该模块对第二视觉特征向量进行处理，进一步提取深度视觉特征。卷积模块二的结构如图16所示，由1个残差块和1个下采样残差块构成，其中残差块级联在下采样残差块之后。

DownResBlock(64,128)

ResBlock(128)

VisualFeature2

VisualFeature3

图 16 卷积模块二结构图

下采样残差块中主路径的第一层卷积层的卷积核大小为3x3，步长为2x2，填充大小为1x1，输入通道数量为64，输出通道数量为128；第二层卷积层的卷积核大小为3x3，步长为1x1，填充大小为1x1，输入输出通道数量均为128；副路径中卷积层的卷积核大小为1x1，步长为2x2，输入通道数量为64，输出通道数量为128。

残差块中2层卷积层的卷积核大小均为3x3，步长均为1x1，填充大小均为1x1，输入输出通道数都为128。

输入的第二视觉特征向量依次经过下采样残差块和残差块处理后得到第三视觉特征。

对于参考视频特征提取分支：

* 输入：参考视频第二视觉特征VisualFeature2，维度为[64, , ]
* 输出：参考视频第三视觉特征VisualFeature3，维度为[128, , ]

对于重建视频特征提取分支：

* 输入：第二视觉融合特征FusionFeature2，维度为[64, , ]
* 输出：重建视频第三视觉特征VisualFeature3，维度为[128, , ]
  + - 1. 卷积模块三

该模块对第三视觉特征向量进行处理，更进一步提取深度视觉特征。卷积模块三的结构如图17所示，由1个残差块和1个下采样残差块构成，其中残差块级联在下采样残差块之后。

DownResBlock(128,256)

ResBlock(256)

VisualFeature3

VisualFeature4

图17 卷积模块三结构图

下采样残差块中主路径的第一层卷积层的卷积核大小为3x3，步长为2x2，填充大小为1x1，输入通道数为128，输出通道数为256；第二层卷积层的卷积核大小为3x3，步长为1x1，填充大小为1x1，输入输出通道数均为256；副路径卷积层的卷积核大小为1x1，步长为2x2，输入通道数为128，输出通道数为256。

残差块中的2层卷积层的卷积核大小均为3x3，步长均为1x1，填充大小为1x1，输入输出通道数均为256。

对于参考视频特征提取分支：

* 输入：参考视频第三视觉特征VisualFeature3，维度为[128, , ]
* 输出：参考视频第四视觉特征VisualFeature4，维度为[256, , ]

对于重建视频特征提取分支：

* 输入：第三视觉融合特征FusionFeature3，维度为[128, , ]
* 输出：重建视频第四视觉特征VisualFeature4，维度为[256, , ]
  + - 1. 卷积模块四

该模块对第四视觉特征向量进行处理，再更进一步提取深度视觉特征。卷积模块四的结构如图18所示，由1个残差块和1个下采样残差块构成，其中残差块级联在下采样残差块之后。

DownResBlock(256,512)

ResBlock(512)

VisualFeature4

VisualFeature5

图18 卷积模块四结构图

下采样残差块中主路径的第一层卷积层的卷积核大小为3x3，步长为2x2，填充大小为1x1，输入通道数为256，输出通道数为512；第二层卷积层的卷积核大小为3x3，步长为1x1，填充大小为1x1，输入输出通道数均为512；副路径卷积层的卷积核大小为1x1，步长为2x2，输入通道数为256，输出通道数为512。

残差块中的2层卷积层的卷积核大小均为3x3，步长均为1x1，填充大小均为1x1，输入输出通道数均为512。

对于参考视频特征提取分支：

* 输入：参考视频第四视觉特征VisualFeature4，维度为[256, , ]
* 输出：参考视频第五视觉特征VisualFeature5，维度为[512, , ]

对于重建视频特征提取分支：

* 输入：第四视觉融合特征FusionFeature4，维度为[256, , ]
* 输出：重建视频第五视觉特征VisualFeature5，维度为[512, , ]
  + - 1. 特征融合模块
         1. 概述

该模块将重建视频空域视觉特征提取路径中的卷积模块输出的特征与参考视频空域视觉特征提取路径中对应的卷积模块输出特征进行融合，特征融合模块1结构如图19所示，特征融合模块2、3、4结构如图20所示，均包含1层卷积层、1层可变形卷积层和1个激活层，激活层为ReLU。

FusionFeature

DeformConv(3x3,1,3\*Cin,Cout,1,1)

Conv(3x3,1,Cin,18,1,1)

ReLU( )

Concat(Cin,Cin,Cin)

ReferenceVisualFeature

ReconstructedVisualFeature

offset

图19 特征融合模块1结构图

DeformConv(3x3,1,2\*Cin,Cout,1,1)

Conv(3x3,1,Cin,18,1,1)

ReLU( )

Concat(Cin,Cin)

FusionFeature

ReferenceVisualFeature

ReconstructedVisualFeature

offset

图 20 特征融合模块2-4结构图

卷积层中的卷积核大小为3x3，步长为1x1，填充大小为1x1，输出通道数为18；可变形卷积层中的卷积核大小为3x3，步长为1x1，填充大小为1x1，特征融合模块1、特征融合模块2，特征融合模块3和特征融合模块4的输出通道数量分别为64、128、256和1。

特征融合模块1中输入的参考视频帧特征通过卷积层处理，得到可变形卷积的偏移量，同时输入的重建视频帧特征、相应的参考视频帧特征以及它们之间的特征差在通道维度上进行级联，级联后的特征与偏移量同时输入到可变形卷积中进行处理，最后再通过一个激活层输出融合特征。特征融合模块2、3、4中输入的参考视频帧特征通过卷积层处理，得到可变形卷积的偏移量，同时输入的重建视频帧特征和相应的参考视频帧特征在通道维度上进行级联，级联后的特征与偏移量同时输入到可变形卷积中进行处理，最后再通过一个激活层输出融合特征。

* + - * 1. 特征融合模块1
* 输入1：参考视频帧的第二视觉特征VisualFeature2，维度为[64, , ]
* 输入2：重建视频帧的第二视觉特征VisualFeature2，维度为[64, , ]
* 输出：第二视觉融合特征FusionFeature2，维度为[64, , ]
  + - * 1. 特征融合模块2
* 输入1：参考视频帧的第三视觉特征VisualFeature3，维度为[128, , ]
* 输入2：重建视频帧的第二视觉特征VisualFeature3，维度为[128, , ]
* 输出：第三视觉融合特征FusionFeature3，维度为[128, , ]
  + - * 1. 特征融合模块3
* 输入1：参考视频帧的第四视觉特征VisualFeature4，维度为[256, , ]
* 输入2：重建视频帧的第四视觉特征VisualFeature4，维度为[256, , ]
* 输出：第四视觉融合特征FusionFeature4，维度为[256, , ]
  + - * 1. 特征融合模块4
* 输入1：参考视频帧的第五视觉特征VisualFeature5，维度为[512, , ]
* 输入2：重建视频帧的第五视觉特征VisualFeature5，维度为[512, , ]
* 输出：第五视觉融合特征FusionFeature5，维度为[1, , ]
  + 1. 空域质量预测模块

该模块对第五视觉融合特征进行空间池化，采用的池化方法是全局平均。

* 输入：第五视觉融合特征，维度为[N, 1, , ]
* 输出：帧级空域质量分数，维度为[N, 1]
  + 1. 时域视觉特征提取模块

该模块对重建灰度视频和参考灰度视频以相同的方式进行独立处理。该模块主要包括视频相邻帧之间的像素差绝对值计算和局部差值平均池化两个步骤，其中局部差值平均池化的滑动窗口大小为12x12，步长为12x12。该模块不包含可学习参数。

首先对灰度视频相邻帧逐像素计算像素差的绝对值，得到灰度视频的绝对帧差图。其中，，，W为灰度视频水平方向像素数量，H为视频垂直方向像素数量，N为灰度视频的帧数。

然后对绝对帧差图在空间维度进行局部平均池化操作，获得平均池化帧差作为视频时域特征，其中，，。池化方法如公式（1）所示。

（1）

* 输入：灰度视频，维度为[N, 1, W, H]
* 输出：视频时域特征，维度为[(N-1), 1, , ]
  + 1. 时域质量预测模块

该模块包括视频时域特征失真计算和非线性变换。首先计算重建视频时域质量特征相对于参考视频时域质量特征之间的绝对差异，如公式（2）所示，其中，，。

（2）

然后对上述时域质量特征差作公式（3）所示的对数非线性变换。

（3）

* 输入：参考视频时域特征和重建视频时域特征，维度均为[(N-1), 1, , ]
* 输出：重建视频帧时域质量分数，维度为[N-1, 1]
  + 1. 质量融合模块

该模块包括视频帧级时空质量融合和帧级质量整合。

视频帧级时空质量融合过程中，按照公式（4）逐帧对帧级空域质量分数和相应帧的时域质量分数求取均方根，得到帧级质量分数。其中。

（4）

视频帧级质量整合过程中，首先根据时域质量分数搜索出时域质量增幅最大的4个视频帧，选取每一个的前40帧作为质量关键帧。最终的视频质量由所有质量关键帧的质量分数平均得到。

* 输入：重建视频帧级空域质量分数和时域质量分数，维度分别为[N, 1]和[N-1, 1]
* 输出：重建视频的质量预测分数

1. （规范性附录）  
   增强视频的编解码重建质量评价技术实现

客观视频质量评价技术所需的结构描述性信息和网络权值参数已对AVS会员单位开源。参考软件代码地址为：https://github.com/VisualMCL/AVSVQA。

1. （规范性附录）  
   增强视频的编解码重建质量评价技术使用建议

本文件描述了两个视频客观质量评价技术，包括一个无参考视频客观质量评价技术方案和一个全参考视频客观质量评价技术方案。

对于只有单个视频作为输入，评估其人眼视觉感知质量的使用场景，本文件建议使用无参考视频客观质量评价技术方案。

对于同时有重建视频和参考视频作为输入，评估重建视频的人眼视觉感知质量的使用场景，本文件建议使用全参考视频客观质量评价技术方案。