研究与开发

基于深度学习的多层级恰可察觉失真预测

徐海峰[1]，王鸿奎[1]，殷海兵[1]，陈楚翘[2]
（1. 杭州电子科技大学通信工程学院，浙江 杭州 310018；
2. 杭州电子科技大学网安学院，浙江 杭州 310018）

摘　要：视觉恰可察觉失真（just noticeable distortion，JND）直接反映人眼视觉系统对视觉信号噪声的敏感程度，广泛应用于图像和视频处理领域。针对视频 JND 阈值的多层级预测问题，将其转化为用户满意率（satisfied user ratio，SUR）曲线的预测问题，并提出一种基于特征融合的 SUR 曲线预测模型。该模型主要分为关键帧选择模块、特征提取和融合模块以及 SUR 分数回归模块。在关键帧选择模块，根据视觉感知机制，提出空时域感知复杂度并以此作为视频关键帧判决指标。在特征提取和融合模块，基于密集残差块（dense residual block，RDB）提出多尺度密集残差网络实现图像特征提取和多尺度融合。实验结果表明，所提出的 SUR 曲线预测模型在 JND 阈值预测精度方面整体优于现有模型，且在运行效率上平均降低 8.1% 的时间成本。同时，该模型还可以用于预测其他层级 JND 阈值，可直接应用于视频多层级感知编码优化。

关键词：恰可察觉失真；深度学习；质量评价
中图分类号：TN919
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024015

Deep learning-based prediction of multi-level just noticeable distortion

XU Haifeng[1], WANG Hongkui[1], YIN Haibing[1], CHEN Chuqiao[2]
1. College of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
2. College of Cyberspace Security, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: Visual just noticeable distortion (JND) directly reflects the sensitivity of the human visual system to visual signal noise, and is widely used in image and video processing. Aiming at the multilevel prediction problem of video JND threshold, it was transformed into the prediction problem of satisfied user ratio (SUR) curve, and a feature fusion-based SUR curve prediction model was proposed. The model was mainly divided into key frame extraction

收稿日期：2023−10−10；**修回日期：**2024−01−12
基金项目：国家自然科学基金资助项目（No.62202134，No.62031009，No.61972123）；科技部重点研发课题资助项目（No.2023YFB4502800）；浙江省“尖兵”“领雁”研发攻关计划项目（No.2023C01149，No.2022C01068）；浙江省自然科学基金资助项目（No.LDT23F01014F01，No.LDT23F01011F01）
Foundation Items: The National Natural Science Foundation of China (No.62202134, No.62031009, No.61972123), Ministry of Science and Technology Key Research and Development Project Funding Program (No.2023YFB4502800), Zhejiang Provincial “Pioneer” and “Leading Goose” Research and Development Project (No.2023C01149, No.2022C01068), The Natural Science Foundation of Zhejiang Province (No.LDT23F01014F01, No.LDT23F01011F01)

module, feature extraction and fusion module, and SUR score regression module. In the key frame extraction module, according to the visual perception mechanism, the spatial-temporal domain perception complexity was proposed and used as the video key frame judgment index. In the feature extraction and fusion module, a multi-scale dense residual network was proposed based on dense residual block (RDB) to realize image feature extraction and multi-scale fusion. The experimental results show that the proposed SUR curve prediction model is overall better than the existing models in terms of JND prediction accuracy and reduces the time cost by 8.1% on average in terms of operational efficiency. Meanwhile, the model can also be used to predict other layers of JND thresholds, which can be directly applied to video multilevel perceptual coding optimization.

Key words: just noticeable distortion, deep learning, quality evaluation

0 引言

恰可察觉失真（just noticeable distortion，JND）直接反映人眼视觉系统（human visual system，HVS）对图像和视频中视觉噪声的敏感程度，已被广泛应用于图像和视频压缩、质量评价、图像增强等各个方面[1]。根据应用场景的不同，早期的 JND 阈值预测模型主要分为像素域预测模型[2-3]和变换域预测模型[4-6]。像素域预测模型通常考虑多种视觉掩蔽效应，采用非线性叠加模型预测每一个像素的 JND 阈值。变换域预测模型普遍采用连乘模型预测各变换分量的 JND 阈值。近年来，随着机器学习和深度学习的发展，学术界开始探索基于数据驱动的 JND 阈值预测。由于视觉实验很难确定像素级等小颗粒度的 JND 标签，因此这些数据驱动的 JND 模型普遍是图像级和视频级的。

事实上，作为图像和视频信号接收器，HVS 无法感知信号中像素的细微失真，可感知到的失真下限就是所谓的恰可察觉失真，即 JND 阈值。视觉实验进一步表明，对于一系列不同压缩程度的图像或视频，HVS 仅能察觉出有限的质量层级，每一层级的失真上限就是该层级的 JND 阈值[7-9]。压缩视频的传统和感知 RD 函数如图 1 所示，传统率失真（rate distortion，RD）曲线反映了视频压缩码率与客观失真的非线性关系。事实上，HVS 并不能察觉出一定码率范围内压缩视频中的细微差异。每一个质量层级的恰可察觉失真，对应该层级的码率下限，同时也对应该层级可取量化参数（quantization parameter，QP）的上限。该码率下限就是该质量层级的视频压缩极限，对应的量化参数可定义为该质量层级的恰可感知量化参数（just noticeable quantization parameter，JNQP）。

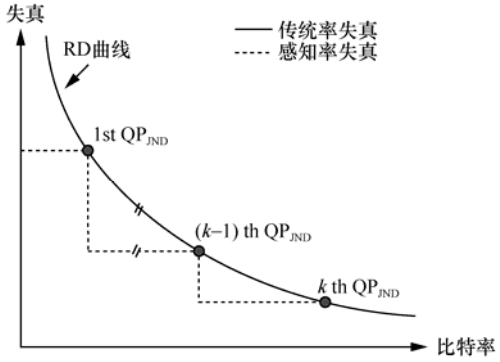


图 1 压缩视频的传统和感知 RD 函数

随着图像和视频 JND 数据集的陆续发布，学术界近年来开始探索多层级 JND 阈值预测方法。例如，Huang 等[10]使用支持向量机来预测图像中第一质量层级的 JND 阈值；Zhang 等[11]采用深度学习方法来预测视频的 JND 阈值；文献[12]将 JND 阈值预测问题转化为用户满意率（satisfied user ratio，SUR）曲线预测问题。本文针对视频信号的多层级 JND 阈值预测问题，借鉴文献[12]将 JND 阈值预测问题转化为 SUR 曲线预测问题并提出基于特征融合的多层级 SUR 预测网络。该网络包含关键帧选择（key frame selection，KFS）模块、特征提取和融合模块以及 SUR 分数回归

模块。在关键帧选择模块，本文在深入分析空域和时域感知特征基础上，提取视频若干帧为 JND 阈值预测感知关键帧。在特征提取和融合模块，本文设计了特征提取网络分别提取原始视频和失真视频特征，尝试了两种特征融合策略探索适合本任务的最佳特征融合策略。在 SUR 分数回归模块，本文采用通道注意力机制在降低冗余特征影响情况下增强有效特征，并使用 2 个双层卷积网络实现最终的 SUR 预测回归。

1　数据驱动的 JND 阈值预测

为了精确预测图像和视频的恰可察觉失真，近年来学术界相继发布了一系列 JND 数据集并探索了基于数据驱动的 JND 阈值预测方法。

如图 1 所示，QP_{JND} 表示以 QP 为参数的不同感知层级 JND 阈值，HVS 对于一系列压缩视频仅能识别出有限的质量层级，质量层级临界 QP 点在一定程度上反映 JND 阈值。因此，数据驱动的 JND 阈值预测问题可以转化为 JNQP 的预测问题。对于同一视频源，测试人员采用视频编码器以不同 QP 将其压缩为不同质量的失真源视频。然后，将原始视频作为对比对象，和其他 QP 的压缩视频同时放置于播放器左右两侧，测试人员对比两边视频，当测试人员察觉两边视频不一致时就可确定第一质量层级的 JNQP；而后将该压缩视频作为新的对比对象，重复上述视觉实验，以确定第二质量层级的 JNQP；依次类推，确定所有质量层级的 JNQP，构建不同测试人员的多层级 JNQP 数据集。基于上述数据集构建方法，学术界近年发布一系列面向图像和视频的 JND 数据集，由于文献[10,13,16]中的数据库并未命名，在本文中称其为 Huang2017、Lin2015 和 Shen2020，压缩图像和视频数据库见表 1。其中压缩图片数据集 6 个[7, 13, 14-17]，包括立体图片[14]和全景图片[15]；压缩视频数据集 4 个[8-10, 13]。

基于上述 JND 数据集，学术界采用统计建模、支持向量回归以及深度学习等方法提出一系列 JND 阈值预测方法[18-19]。对于相同失真的图像和视频，不同测试人员的主观感受并不完全一致，这就导致 JNQP 样本数据近似服从高斯分布。主观测试中 JND 样本的直方图、高斯分布函数（PDF）和 SUR 曲线示例如图 2 所示，本文提供了 VideoSet[9]中抽取的单个视频的第一质量层级 JNQP 样本分布。假设一个视频的 JNQP 服从样本均值为 μ 和方差为 σ 的高斯分布，即：

表 1　压缩图像和视频数据库

数据库	类型	原始样本数	分辨率/像素	QF/QP/VBR	编码器	压缩样本数	测试者数量	发布时间
Lin2015[13]	图片	5	1 920×1 080	QF:1～100	JPEG	5×100	20	2015 年
	视频	5	1 920×1 080	QP:1～51/VBR	x264/x265	5×51×2×2	20	2015 年
MCL-JCI[7]	图片	50	1 920×1 080	QF:1～100	JPEG	50×100	20	2015 年
MCL-JCV[8]	视频	30	1 920×1 080	QF:1～100	x264	30×51	50	2016 年
Huang2017[10]	视频	40	1 920×1 080	QP:1～51	HM16.0-RA	40×51	30	2017 年
VideoSet[9]	视频	220×4=880	1 920×1 080 1 280×720 960×540 640×360	QP:1～51	H.264/AVC	880×51	30+	2017 年
SIAT-JSSI[14]	立体对称图片	12	1 920×1 080	QF:1～300	JPEG2000	7020	50	2019 年
SIAT-JASI[14]	立体非对称图片	12	1 920×1 080	QP:1～51	HM16.7-AI	7020	50	2019 年
JND-Pano[15]	全景图片	40	5 000×5 000	QF:1～100	JPEG	4040	24	2018 年
Shen2020[16]	图片	202	1 920×1 080	QP:13～51	VTM5.0-AI	39×202	20	2020 年
KonJND-1k[17]	图片	1 008	640×480	QF:1～100	JPEG/BPG	77 112	42	2022 年

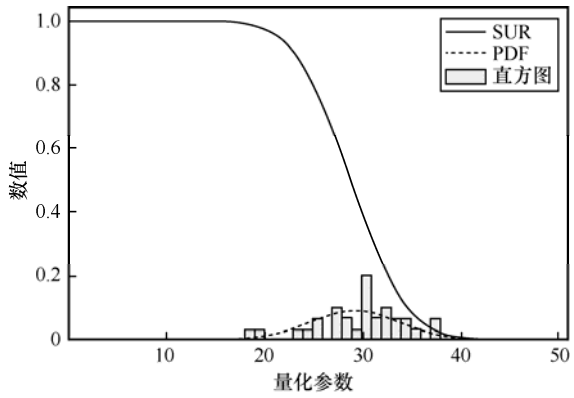


图 2　主观测试中 JND 样本的直方图、PDF 和 SUR 曲线示例

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

其中，x 表示 JNQP 样本值，其互补累积函数可表示为：

$$\overline{S_J}=(d_j|\mu,\sigma^2)=1-\int_{-\infty}^{d_j}\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \tag{2}$$

其中，d_j 表示每个压缩视频的测试样本值，j=1,2,⋯,51 表示 51 个 QP。该函数就是 SUR 曲线。如图 2 所示，SUR 曲线反映了观众对不同失真视频的满意比例。例如，当 SUR=0.75 时，意味着给定的失真级别能够使 75%的观众满意。

SUR 曲线反映了所有测试人员对失真视频整体的感知情况，相比于 JNQP 预测更能反映用户对失真视频整体的满意程度。因此，部分学者将 JND 阈值预测问题转化为 SUR 曲线的预测问题。根据文献[14]，本文以 7 为间隔对 51 个 QP 点对应的 SUR 值进行下采样预测，从而降低整体的预测复杂度。

2　基于特征融合的 SUR 曲线预测

本文所提出的 SUR 曲线预测模型由 3 个主要模块组成：KFS 模块、特征提取和融合模块和 SUR 分数回归模块。SUR 曲线预测模型的框架如图 3 所示，其中，卷积统一表示为二维卷积，除了 SUR 分数回归子网络中 Q 网络和 W 网络的激活函数分别为 Sigmoid、ReLU，其余激活函数均为 LeakyReLU。

2.1　关键帧选择模块

视觉掩蔽效应表明视频中的复杂运动与复杂纹理严重影响 HVS 对视频细节的内容感知[3]。本文分别采用空域感知复杂度和时域感知复杂度来表示 HVS 对视频内容的敏感程度[20]。其中，视频信号帧级空域感知复杂度被定义为视频帧经 Sobel 滤波后的帧级标准差。空域感知复杂度越高表明图像内容纹理越复杂，HVS 更难察觉视频质量的差异性变化。时域感知复杂度被定义为帧间残差的标准差，时域感知复杂度越高，表明视频运动掩蔽效应越强，HVS 对失真感知越弱。空时域感知复杂度分别表示为：

$$Y_{\mathrm{S}}(f_n)=\sigma_{\mathrm{s}}\left[T_{\mathrm{s}}(f_n)\right] \tag{3}$$

$$Y_{\mathrm{T}}(f_n)=\sigma_{\mathrm{s}}(f_n-f_{n-1}) \tag{4}$$

其中，f_n 表示视频帧，n 为帧索引，T_{s} 表示 Sobel 滤波器，σ_{s} 表示空间标准偏差。Y_{S} 和 Y_{T} 分别是空域和时域感知复杂度。视频帧级感知复杂度定义为空域和时域感知复杂度的加权和，具体表示为：

$$Y=\alpha\times Y_{\mathrm{S}}+(1-\alpha)\times Y_{\mathrm{T}} \tag{5}$$

其中，α 是加权参数。为了确定 α 精确取值，分别计算 Y 值与 3 个质量层级 JND 点的相关性。Y 值与 JNQP 的关系如图 4 所示。图 4（a）给出了不同 α 取值下得到的 Y 值与 3 个质量层级之间的皮尔逊线性相关性系数（Pearson linear correlation coefficient，PLCC），计算结果表明 α 取值 0.6 时相关性最高。为了分析 JND 点位置和 Y 值之间的关系，本文还绘制了 Videoset 数据集中 30 个视频 3 个质量层级的 JNQP 以及每个视频的平均感知复杂度，不同视频中感知复杂度与对应的 3 个层级 JND 如图 4（b）所示。此外，本文提供 Videoset 数据集中 3 个典型视频序列以及对应的感知复杂度如图 5 所示。在图 5（a）的序列 001 和序列 089 中，它们的背景纹理是比较复杂且运动变化相对较快，导致了它们有较高的 Y 值（即分别为 26.77 和 33.35）。相反，图 5（a）中序列 037 是纹理平滑且运动缓慢的，它的 Y 值远远小于图 5（a）中的序列 001 和序列 089。

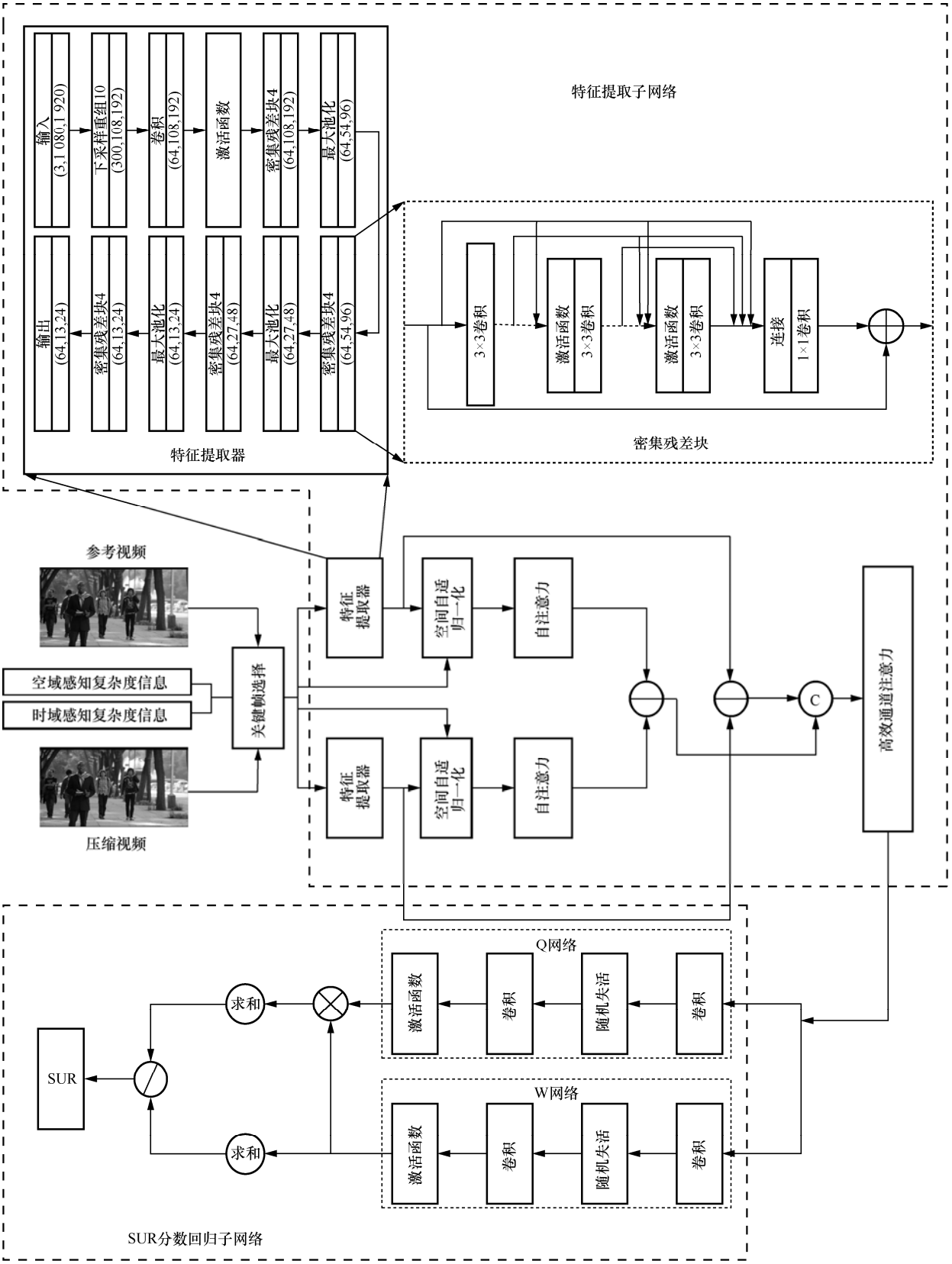
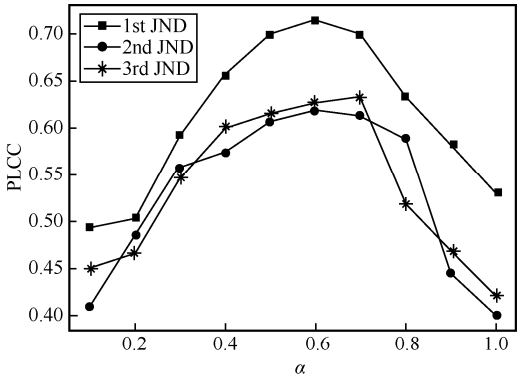
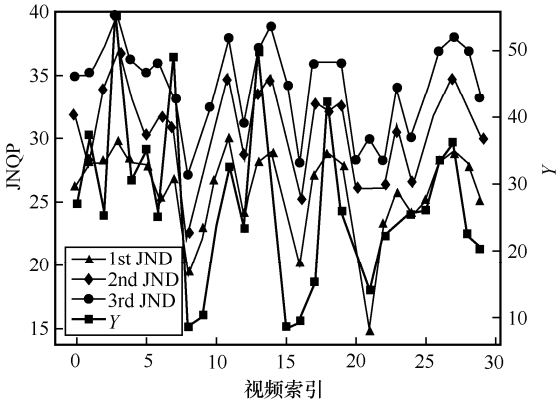


图 3　SUR 曲线预测模型的框架



(a) 不同α取值下Y与JNQP的PLCC



(b) Y与3个层级JNQP的位置关系

图 4　Y 值与 JNQP 的关系

可以看出，Y 值和 JNQP 的位置有较好的相关性，视频的 Y 值越低，则其对人眼越敏感，JNQP 点的位置越靠前。为了平衡性能和计算复杂度，每个视频中 Y 值最低的 10 帧被选取为关键帧并输入本文的 SUR 预测网络。

2.2　特征提取和融合模块

如图 3 所示，本文的 SUR 预测模型存在两个输入分支，分别为参考视频 V_{ref} 和压缩视频 V_{dis}。首先经过关键帧选择模块的预处理后，得到一组作为输入的视频帧 $\Gamma_i=(V_{\text{ref}}^i,V_{\text{dis}}^i),i\in[1,10]$。视频帧将通过本文提出的一个密集连接网络提取视频帧内的有效特征。具体而言，每一帧先由下采样重组操作进行快速下采样，原始帧的尺寸为 $c\times h\times w$，其中 c、h 和 w 分别是图像的通道数、高度和宽度。快速下采样后，图像尺寸变为 $(c\times r^2)\times(h/r)\times(w/r)$，$r$ 是高度和宽度的缩小比例。统一缩小图像尺寸并叠加在通道数上的目的是在尽量不丢失图像信息的情况下减小图像尺寸，方便后续的特征提取。为了简化后续的卷积运算，本模型使用一个二维卷积+激活函数 LeakyReLU 的操作，直接压缩特征通道数量至 64，并使用 3 个 RDB 和最大池化来不断降维，提取深度图像特征。RDB 集成了残差块（residual block，RB）和密集块（dense block，DB）的优势[21]，有助于提高模型的表达能力，提取出图像的细致特征。经过三重降维后，本模型额外添加一个 RDB 模块，进一步挖掘图像的语义特征。

实际上，本文认为有监督学习任务的本质是学习两者之间的差异强度，因此所提取的两个分支

(a) 原图

(b) Y值图

图 5　Videoset 数据集中 3 个典型视频序列以及对应的感知复杂度

（即参考视频和压缩视频）特征间的差异强度特征对结果有重要影响。本文所构建模型的两个分支享有共同的权重，特征提取器提取的特征分别为参考帧特征 F_{ref}^{i} 和压缩帧特征 F_{dis}^{i}，如式（6）所示。

$$F_{\text{ref}}^{i},F_{\text{dis}}^{i}=\psi(V_{\text{ref}}^{i},V_{\text{dis}}^{i}),i\in[1,10] \tag{6}$$

其中，$\psi(\cdot)$ 表示特征提取器。

简单地依靠特征提取器提取的最后一层特征可能无法精确地描述视频信息，因此需要对多个不同层级特征进行融合来丰富视频的特征信息。因此，本文首先尝试文献[22]提出的空间自适应归一化（spatially adaptive normalization，SPADE）模块来融合底层图像信息和高层语义特征信息。在模型设计过程中以 SPADE 模块为对比对象设计多尺度密集残差网络（multi-scale dense residual network，MSDRN）进行提取和融合特征。

SPADE 特征融合：本文以原始参考帧作为风格特征，将其与提取的高层语义特征一起输入图 3 中所示的 SPADE 模块中，得到结合底层信息和高级语义特征的双级融合原始特征 F_{ref}^{i}。SPADE 是文献[22]中提出的空间自适应归一化模块，使用该模块的目的是将基础图像信息融合到所提取的高级语义特征中，以确保获得更丰富、更有效的特征表示，有效特征直接影响着最终回归得分。同样地，本文将提取的高层语义特征以压缩帧为风格特征一起输入 SPADE 模块，输出表示压缩帧的双级融合压缩特征 F_{dis}^{i}。

$$F_{\text{ref}}^{i}=\text{SPADE}(F_{\text{ref}}^{i},V_{\text{ref}}^{i}) \tag{7}$$

$$F_{\text{dis}}^{i}=\text{SPADE}(F_{\text{dis}}^{i},V_{\text{dis}}^{i}) \tag{8}$$

MSDRN 特征融合：利用 RDB 堆叠构建的特征提取器来进行多尺度连接融合。现有工作普遍在 RDB 内部直接使用空间金字塔池化来提取多尺度特征并结合传统 RDB 自身产生的多层叠加特征[23]实现多尺度特征融合。从特征提取网络结构来看，该方法并未直接利用 RDB 模块中的多尺度特征。因此，本文尝试设计如图 6 所示的多尺度密集残差网络在不同尺度下使用 RDB 来提取特征，从而将不同尺度下 RDB 提取的特征相加形成多尺度多层级融合特征。

特征提取过程可描述为：

$$F_{\text{ref}}^{i}=\sum_{k=3}^{5}\mathcal{L}(\mathcal{C}(V_{\text{ref}}^{i}))_{k} \tag{9}$$

$$F_{\text{dis}}^{i}=\sum_{k=3}^{5}\mathcal{L}(\mathcal{C}(V_{\text{dis}}^{i}))_{k} \tag{10}$$

其中，$\mathcal{C}(\cdot)$ 表示 1×1 卷积操作，$\mathcal{L}(\cdot)$ 表示上采样操作，k 表示设计网络的层级索引。

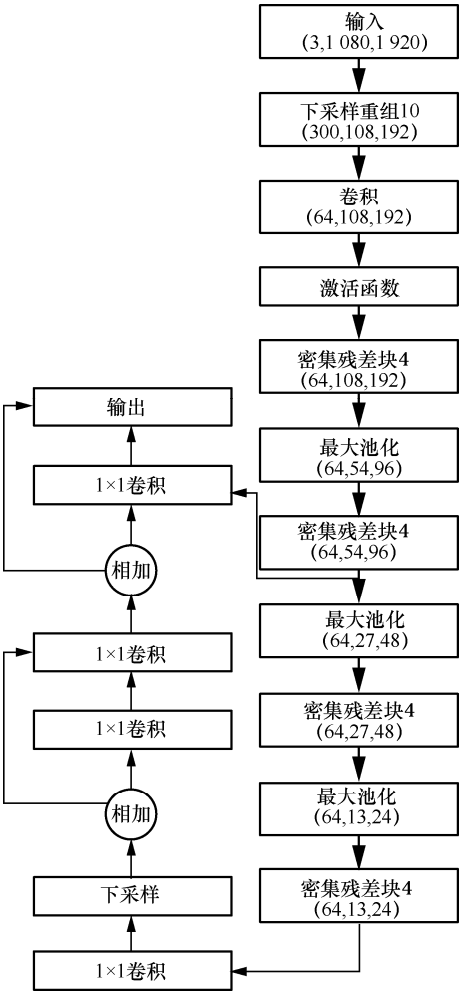


图 6　多尺度密集残差网络

通过特征提取和融合网络获得的两个融合特征分别输入同一个自注意力模块中提取像素间的关联特征。经过自注意力模块后的两个融合特征差称为融合差异特征 F_{md}^{i}，未经过融合的参考帧

特征 F_{ref}^{i} 和压缩帧特征 F_{dis}^{i} 之间的差被称作本质差异特征 F_{ed}^{i}，如式（11）和式（12）所示。最后，将计算获得的两个差异特征拼接作为SUR分数回归网络的输入，得到的输出 F_{final}^{i} 是最终的融合差异特征，如式（13）所示。

$$\overline{F_{\text{ref}}^{i}}, \overline{F_{\text{dis}}^{i}} = \text{SA}\left(F_{\text{ref}}^{i}, F_{\text{dis}}^{i}\right) \tag{11}$$

$$F_{\text{md}}^{i} = \overline{F_{\text{ref}}^{i}} - \overline{F_{\text{dis}}^{i}}, F_{\text{ed}}^{i} = F_{\text{ref}}^{i} - F_{\text{dis}}^{i} \tag{12}$$

$$F_{\text{final}}^{i} = \text{ECA}(\text{Concat}(F_{\text{md}}^{i}, F_{\text{ed}}^{i})) \tag{13}$$

其中，SA(·)表示经过自注意力模块操作，ECA(·)表示经过通道注意力（efficiency channel attention，ECA）模块操作，Concat(·)表示通道维度上的连接操作。

2.3 SUR分数回归模块

差异特征拼接导致特征通道的数量成倍增加，不可避免引入特征冗余。为了消除特征冗余对回归结果的影响，本文使用式（13）中的ECA模块来增强有效特征的激活，ECA通过一维卷积生成通道注意力，是个超轻量级的即插即用模块[24]，可以达到降低冗余特征影响的效果。本文设计的回归网络与其他网络的不同之处在于它并不使用全连接层来确定最终的分数，而是使用两个双层卷积网络作为Q网络和W网络。为了确保最终得分在0和1之间，Q网络的最后一层需要使用Sigmoid函数激活，W网络的最后一层使用ReLU函数激活。如式（14）和式（15）所示，Q_{final} 是最终的帧级SUR分数，而整个视频的SUR分数是对所有的帧级分数平均得来。

$$\begin{aligned} Q &= \text{Sigmoid}(\text{Conv}(F_{\text{final}}^{i})), \\ W &= \text{ReLU}(\text{Conv}(F_{\text{final}}^{i})) \end{aligned} \tag{14}$$

$$Q_{\text{final}} = \frac{\sum(Q \otimes W)}{\sum W} \tag{15}$$

其中，Q 和 W 分别是Q网络和W网络得到的分数和权重，⊗表示点乘操作，∑(·)表示求和操作。

在所有的训练和测试过程中，本文分别使用L2损失函数和L1损失函数。L2损失是均方误差（mean square error，MSE）损失函数，可以使训练损失快速下降，让网络快速拟合。而L1损失为平均绝对值误差（mean absolute error，MAE）损失函数，能直接反映出预测结果和真实标签（ground truth，GT）之间的误差。

3 实验结果与分析

3.1 数据集介绍

本文使用的数据集是VideoSet[9]，包括4种分辨率（即 1 920×1 080、1 280×720、960×540 和640×360，单位为像素）的220个源视频，共880个视频。每个源视频都使用 x.264/AVC 和 51 个QP（1～51）压缩成对应的51个失真视频，并且通过主观测试获得了每个视频的3个JND点。本文将220个源视频按照6:2:2的比例随机分成训练集、验证集和测试集，而它们的压缩视频也被同样划分。在测试过程中，我们把从 5 个不同训练批次中获得的最佳模型结果平均值作为最终预测结果，每个训练批次共100个epoch。本文所有的实验均在NVIDIA GTX1080Ti上进行测试。

3.2 消融实验

本文设计如下消融实验来验证本文模型各模块的有效性。关键帧选择模块：根据视觉感知机制，本文采用空时域感知复杂度代替PSNR指标选择视频帧中质量最差的 10 帧为关键帧。不同模块在VideoSet上的性能比较见表2，其中WQFS表示文献[12]采用基于PSNR指标的关键帧选择方式，KFS表示本文关键帧选择方式。特征提取模块：本文实验发现通过RDB模块堆叠有助于提高模型精度，且当RDB模块超过4个时出现过拟合现象。因此，消融实验中本文提供了 KFS+ add_RDB 测试，其中add_RDB表示RDB模块堆叠操作，RDB数量为4。特征融合模块：在此模块，本文提出两套方案进行特征融合，其一为RDB模块堆叠和SPADE的组合，另外一种是本文设计的多尺度密集残差网络。因此，消融实验中本文测试了 KFS+add_RDB+SPADE 和KFS+MSDRN两种配置的预测精度。

表 2　不同模块在 VideoSet 上的性能比较

方法	SUR		QP	
	平均值	方差	平均值	方差
WQFS [12]	0.076	0.009 6	3.01	3.45
KFS	0.072	0.009 2	2.67	2.89
KFS+add_RDB	0.069	0.008 2	2.55	2.25
KFS+add_RDB+SPADE	0.052	0.006 6	2.41	1.93
KFS+MSDRN	0.051	0.006 2	2.22	1.83

本文测试了分辨率为 540p 数据集的第一个 JND 点预测结果，指标分别为 SUR 和 QP 的平均误差和平均方差误差。实验结果表明，采用空时域感知复杂度作为关键帧选择指标有效降低了 SUR 和 QP 的预测误差和方差，从而证明了本文关键帧选择模块的有效性。在关键帧选择基础上，通过 RDB 模块堆叠操作进一步提高了模型的预测精度。通过 SPADE 模块的特征融合，SUR 和 QP 的预测误差大幅下降，从而证明 SPADE 特征融合模块在本任务中的有效性。此外，采用 MSDRN 模块替换 add_RDB 和 SPADE 模块，预测精度最高，充分说明了本文模型各模块的有效性。

3.3　性能评估

为了综合评估本文提出的两个预测模型，根据特征提取和融合方式的不同，分别命名为 SUR-SPADE 和 SUR-MSDRN，并与文献[12]提出的 3 个模型进行性能比较。文献[12]提出的 3 个模型中，VW-SSUR 考虑到图像空间纹理在视频质量评价（video quality assessment，VQA）中起关键作用，通过深度神经网络提取空间纹理信息回归预测 SUR 分数。VW-STSUR-QF 和 VW-STSUR-FF 则是在考虑空间纹理信息的基础上，引入光流作为时间运动参考信息。不同的是 VW-STSUR-QF 是在空间信息和时间运动信息两个分支分别预测各自的分数后加权融合得到最终的分数，而 VW-STSUR-FF 是直接将空间特征和时间特征串联起来进行预测。预测值和真实标签之间的绝对误差被当作性能评估指标（即|ΔSUR|、|ΔQP|、|ΔPSNR|、|ΔSSIM|），QP 表示视频编码中的量化参数，峰值信噪比（peak signal-to-noise ratio，PSNR）和结构相似度测量（structure similarity index measure，SSIM）常用于压缩视频的质量评价指标。指标|ΔPSNR|和|ΔSSIM|可以用式（16）和式（17）计算得出：

$$\left|\Delta\mathrm{PSNR}\right| = \left|\mathrm{PSNR_O} - \mathrm{PSNR_P}\right| \tag{16}$$

$$\left|\Delta\mathrm{SSIM}\right| = \left|\mathrm{SSIM_O} - \mathrm{SSIM_P}\right| \tag{17}$$

其中，$\mathrm{PSNR_O}$、$\mathrm{SSIM_O}$ 分别表示原始 JNQP（经过高斯分布拟合后的 SUR 曲线，当 SUR=0.75 时取得的 QP）对应的编码视频与源视频之间计算的 PSNR、SSIM，$\mathrm{PSNR_P}$、$\mathrm{SSIM_P}$ 分别表示本文方法预测得到 JNQP 对应的编码视频与源视频之间计算的 PSNR、SSIM。本文模型与其他模型在 SUR 和 JND 预测误差比较见表 3。

均值误差越小则表明该模型的预测精度越高，较小的方差误差表示模型稳定性更强。如表 3 中所示，不同预测模型在不同指标上略有差异。整体上，本文提出的 SUR-SPADE 和 SUR-MSDRN 在误差均值方面小于 VW-SSUR，但略高于 VW-STSUR- QF 和 VW-STSUR-FF；在误差方差上与最好的模型基本一致。文献[12]中性能更好的模型 VW-SSUR 和 VW-SSUR-FF 考虑了光流作为时空信息特征，模型预测时间相对更长。相比较而言，本文提出的方法相比减少约 8.1%的时间消耗。

此外，本文在 VideoSet[9]数据集上测试了 1 080p、720p、540p 以及 360p 分辨率的视频，多分辨率下参考视频的 SUR 和 QP 预测误差比较见表 4。测试结果与 VW-SSUR 进行了对比，其中性能指标选择|ΔSUR|和|ΔQP|的均值误差。结果表明本文提出模型明显优于 VW-SSUR，且 SUR-MSDRN 优于 SUR-SPADE。VideoSet 中每个 1 080p 测试视频的 1st JND 点 SUR 预测误差分布如图 7 所示，可以观察到 SUR-SPADE 的大部分误差都集中在 0.04 左右，SUR-MSDRN 的误差基本都在 0.04 以

表 3　本文模型与其他模型在 SUR 和 JND 预测误差比较

指标	方法	|ΔSUR|	|ΔQP|	|ΔPSNR|	|ΔSSIM|	时间/s
平均值	VW-SSUR[12]	0.066	1.90	0.91	1.63×10^{-3}	15.51
	VW-STSUR-QF[12]	0.056	1.69	0.84	1.56×10^{-3}	19.28
	VW-STSUR-FF[12]	0.049	1.86	0.90	1.71×10^{-3}	20.33
	SUR-SPADE	0.055	1.89	0.90	1.60×10^{-3}	14.26
	SUR-MSDRN	0.053	1.86	0.89	1.61×10^{-3}	14.04
方差	VW-SSUR[12]	0.006 4	1.94	0.46	1.90×10^{-6}	—
	VW-STSUR-QF[12]	0.004 8	1.69	0.51	2.17×10^{-6}	—
	VW-STSUR-FF[12]	0.005 7	2.07	0.60	3.61×10^{-6}	—
	SUR-SPADE	0.007 2	2.01	0.53	2.55×10^{-6}	—
	SUR-MSDRN	0.005 6	1.66	0.51	2.11×10^{-6}	—

表 4　多分辨率下参考视频的 SUR 和 QP 预测误差比较

方法	分辨率			
	1 080p	720p	540p	360p
VW-SSUR[12]	0.066/1.90	—	0.056/2.27	—
SUR-SPADE	0.055/1.89	0.058/2.21	0.052/2.41	0.061/2.30
SUR-MSDRN	0.053/1.86	0.057/2.32	0.051/2.22	0.059/2.48

下，这得益于本文设计的关键帧选择方法以及融合差异特征，提取的特征能够准确地表征压缩视频与源视频之间的差异强度。VideoSet 中每个 1 080p 测试视频的 1st JND 点 QP 预测误差分布如图 8 所示，QP 误差越接近 0 表示预测得越准确，能够发现本文所构建的两个模型在 QP 预测上表现良好，大部分均在 0 附近。

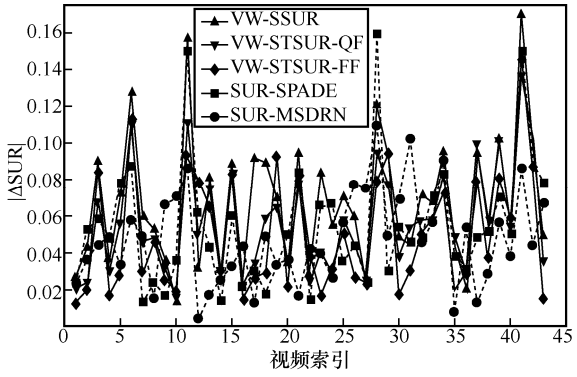


图 7　VideoSet 中每个 1 080p 测试视频的 1st JND 点 SUR 预测误差分布

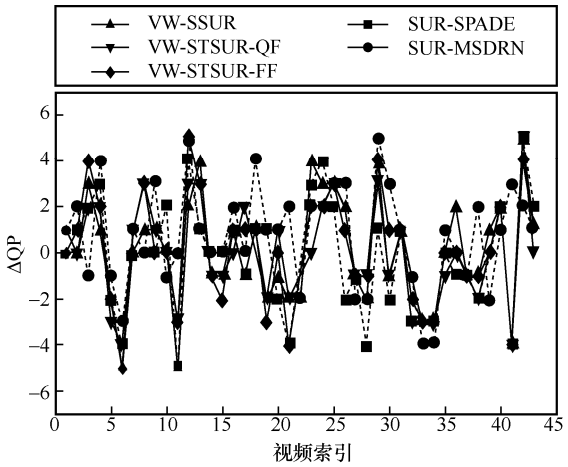


图 8　VideoSet 中每个 1 080p 测试视频的 1st JND 点 QP 预测误差分布

基于本文提出的预测模型，本文测试了第二层级（2nd）以及第三层级（3rd）JND 预测。本文模型在 VideoSet 中分辨率为 1 080p 视频上的第二和第三层级的 JND 预测结果见表 5。相较于第一层级 JND 预测，SUR 的预测误差相差不大，而

表 5　本文模型在 VideoSet 中分辨率为 1080p 视频上的第二和第三层级的 JND 预测结果

方法	指标	层级	|ΔSUR|	|ΔQP|	|ΔPSNR|	|ΔSSIM|
SUR-SPADE	平均值	2nd JND	0.049	1.65	0.71	2.01×10^{-3}
		3rd JND	0.052	1.67	0.75	2.25×10^{-3}
SUR-MSDRN		2nd JND	0.051	1.40	0.69	1.97×10^{-3}
		3rd JND	0.050	1.53	0.72	2.27×10^{-3}
SUR-SPADE	方差	2nd JND	0.004 1	1.32	0.56	6.36×10^{-6}
		3rd JND	0.005 6	1.96	0.47	5.23×10^{-6}
SUR-MSDRN		2nd JND	0.003 9	1.54	0.49	4.61×10^{-6}
		3rd JND	0.004 6	1.80	0.51	6.78×10^{-6}

QP 和 PSNR 的预测误差降低了不少。图 9～图 12 所示分别为 SUR-SPADE 和 SUR-MSDRN 3 个层级的 SUR 和 QP 预测误差分布，可以看出，随着层级的增加，QP 预测得越准确，这可能是因为第二层级和第三层级的 JND 已经处于较低的质量范围，其失真对于 HVS 更加明显，因此模型能够正确地区分视频质量的差异，从而使得预测结果更加准确。

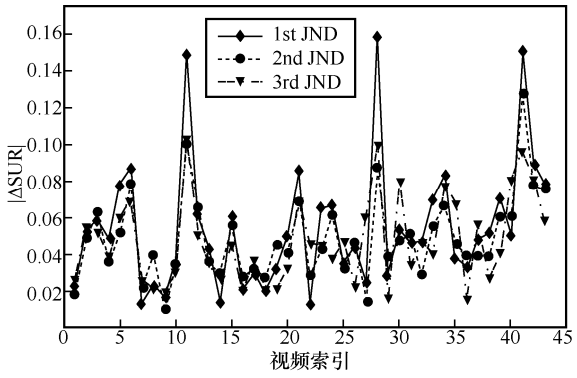


图 9　SUR-SPADE 预测的 3 个层级 SUR 绝对误差分布

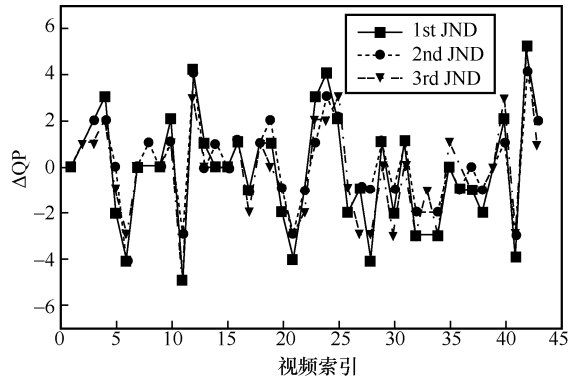


图 10　SUR-SPADE 预测的 3 个层级 QP 误差分布

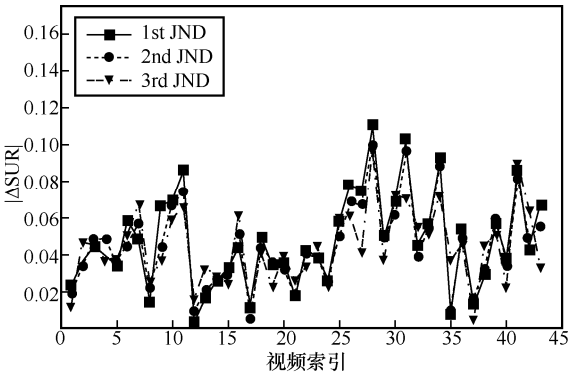


图 11　SUR-MSDRN 预测的 3 个层级 SUR 绝对误差分布

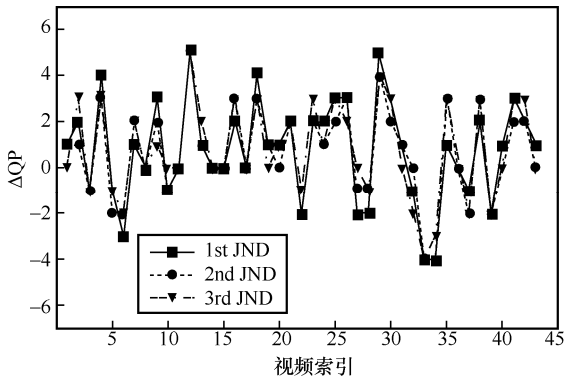


图 12　SUR-MSDRN 预测的 3 个层级 QP 误差分布

4　结束语

本文针对视频 JND 阈值预测中多层级预测问题，提出了一种基于特征融合的 SUR 曲线预测模型，实现视频级的多层级 JND 阈值预测。在关键帧选择过程中，本文创新性地提出空时域感知复杂度，并以此作为关键帧判决指标。在特征提取和融合过程中，本文探索了 RDB 模

块迭代对预测精度的影响、SPADE 模块对特征融合的作用，在此基础上提出了多尺度密集残差网络实现特征的提取和多尺度融合。实验结果表明，本文提出的多层级 JND 预测模型整体上优于现有单层级 JND 预测模型，在运行效率上平均降低 8.1%的时间成本。同时，测试结果也表明了本模型在第二、第三层级 JND 预测中良好的预测效果。未来将在本文工作基础上继续探索帧级以及块级的多层级 JND 阈值预测，并将其应用于视频编码过程，实现视频的多层级感知编码压缩。

参考文献：

[1] LIN W S, GHINEA G. Progress and opportunities in modelling just-noticeable difference (JND) for multimedia[J]. IEEE Transactions on Multimedia, 2021(24): 3706-3721.

[2] WU J, LI L, DONG W, et al. Enhanced just noticeable difference model for images with pattern complexity[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2682-2693.

[3] WANG H, YU L, LIANG J, et al. Hierarchical predictive coding-based JND estimation for image compression[J]. IEEE Transactions on Image Processing, 2020(30): 487-500.

[4] BAE S H, KIM M. A novel generalized DCT-based JND profile based on an elaborate CM-JND model for variable block-sized transforms in monochrome images[J]. IEEE Transactions on Image Processing, 2014, 23(8): 3227-3240.

[5] 骆琼华, 王鸿奎, 殷海兵, 等. 基于熵掩蔽的 DCT 域恰可察觉失真模型[J]. 电信科学, 2023, 39(2): 59-70.
LUO Q H, WANG H K, YIN H B, et al. Just noticeable distortion model based on entropy masking in DCT domain[J]. Telecommunications Science, 2023, 39(2): 59-70.

[6] 邢亚芬, 殷海兵, 王鸿奎, 等. 基于视频时域感知特性的恰可察觉失真模型[J]. 电信科学, 2022, 38(2): 92-102.
XING Y F, YIN H B, WANG H K, et al. Video temporal perception characteristics based just noticeable difference model[J]. Telecommunications Science, 2022, 38(2): 92-102.

[7] JIN L N, LIN J Y, HU S D, et al. Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis[J]. Electronic Imaging, 2016, 28(13): 1-9.

[8] WANG H Q, GAN W H, HU S D, et al. MCL-JCV: a JND-based H.264/AVC video quality assessment dataset[C]//Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2016: 1509-1513.

[9] WANG H Q, KATSAVOUNIDIS I, ZHOU J T, et al. VideoSet: a large-scale compressed video quality dataset based on JND measurement[J]. Journal of Visual Communication and Image Representation, 2017(46): 292-302.

[10] HUANG Q, WANG H Q, LIM S C, et al. Measure and prediction of HEVC perceptually lossy/lossless boundary QP values[C]//Proceedings of the 2017 Data Compression Conference (DCC). Piscataway: IEEE Press, 2017: 42-51.

[11] ZHANG X, YANG C, WANG H, et al. Satisfied-user-ratio modeling for compressed video[J]. IEEE Transactions on Image Processing, 2020(29): 3777-3789.

[12] ZHANG Y, LIU H H, YANG Y, et al. Deep learning based just noticeable difference and perceptual quality prediction models for compressed video[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1197-1212.

[13] LIN J Y, JIN L N, HU S D, et al. Experimental design and analysis of JND test on coded image/video[C]//SPIE Optical Engineering + Applications. Applications of Digital Image Processing XXXVIII. [S.l.: s.n.], 2015: 324-334.

[14] FAN C L, ZHANG Y, HAMZAOUI R, et al. Learning-based satisfied user ratio prediction for symmetrically and asymmetrically compressed stereoscopic images[J]. IEEE MultiMedia, 2021, 28(3): 8-20.

[15] LIU X H, CHEN Z H, WANG X, et al. JND-pano: database for just noticeable difference of JPEG compressed panoramic images[C]//Pacific Rim Conference on Multimedia. Berlin: Springer, 2018: 458-468.

[16] SHEN X L, NI Z K, YANG W H, et al. A JND dataset based on VVC compressed images[C]//Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Piscataway: IEEE Press, 2020: 1-6.

[17] LIN H H, CHEN G G, JENADELEH M, et al. Large-scale crowdsourced subjective assessment of picturewise just noticeable difference[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 5859-5873.

[18] TIAN T, WANG H L, ZUO L X, et al. Just noticeable difference level prediction for perceptual image compression[J]. IEEE Transactions on Broadcasting, 2020, 66(3): 690-700.

[19] LIU H H, ZHANG Y, ZHANG H, et al. Deep learning based picture-wise just noticeable distortion prediction model for image compression[J]. IEEE Transactions on Image Processing: a

Publication of the IEEE Signal Processing Society, 2019(29): 641-656.

[20] ITU-T P.910. Subjective video quality assessment methods for multimedia applications[S]. Geneva: ITU-T Publications, 2022.

[21] ZHANG Y, TIAN Y, KONG Y, et al. Residual dense network for image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2472-2481.

[22] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2337-2346.

[23] BAO L, YANG Z, WANG S, et al. Real image denoising based on multi-scale residual dense block and cascaded U-Net with block-connection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2020: 448-449.

[24] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway: IEEE Press, 2020: 11534-11542.

[作者简介]

徐海峰（1999− ），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为感知视频编码。

王鸿奎（1990− ），男，博士，杭州电子科技大学通信工程学院讲师，主要研究方向为感知视频编码。

殷海兵（1974− ），男，博士，杭州电子科技大学通信工程学院教授，主要研究方向为数字视频编解码。

陈楚翘（1993− ），女，博士，杭州电子科技大学网安学院讲师，主要研究方向为智能信息处理。