# 基于迁移学习的无参考视频质量评价

张浩，桑庆兵*

江南大学物联网工程学院，江苏 无锡 214122

**摘要** 视频质量评价主要采用传统的手动提取特征，再利用机器学习预测视频质量分数，导致结果不理想。VGG-16网络在特征提取方面具有非常好的稳健性，借鉴其网络模型，迁移参数构造出用于端到端的视频质量评价网络。LIVE视频数据库的实验结果表明，该方法预测的评价分数与主观评价分数具有较高的一致性。其评价指标斯皮尔曼等级相关系数和皮尔逊线性相关系数分别达到了0.867和0.843，性能优于目前基于手动提取特征进行视频质量评价的大部分算法。

**关键词** 成像系统；视频处理；迁移学习；神经网络；视频质量评价

**中图分类号** TP391.41 **文献标识码** A **doi**：10.3788/LOP55.091101

# No Reference Video Quality Assessment Based on Transfer Learning

Zhang Hao, Sang Qingbing*

*School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu* 214122, *China*

**Abstract** In video quality assessment, most researchers manually extract the features first, and then use machine learning to predict video quality score, which leads to unideal result. Since the VGG-16 net has excellent robustness in feature extraction, we use the network model and migrate parameters to construct the end-to-end video quality assessment network. The experimental results on LIVE video database show that the assessment score of this method is consistent with the subjective assessment score, and its assessment indexes of Spearman rank correlation coefficient and Pearson correlation coefficient reached 0.867 and 0.843, respectively, which indicated that the performance of the proposed method is better than most of the current video quality assessment algorithms based on manual feature extraction.

**Key words** imaging systems; video processing; transfer learning; neural network; video quality assessment

**OCIS codes** 110.4155; 120.1088; 100.4996; 110.3000

## 1 引 言

近年来，随着移动网络的快速发展，用户可以轻松快速地获取视频信息并在社交网站上分享。但视频在采集、编码、传输等过程中通常会产生不同类型的失真[1]，例如采集视频时，对焦问题会引起视频模糊，离散余弦变换(DCT)压缩编码会引起振铃现象，传输过程中会出现模糊噪声等。为了客观地评价失真对视频质量的影响，寻找可以准确评价视频质量的方法依然是一个重要且热门的课题。

根据对未失真原始视频信息的参考度，视频质量评价分为全参考、半参考和无参考方法[2-3]。常见的全参考方法中，Wang等[4]提出了模拟人类视觉系统的结构相似性(SSIM)算法；Vu等[5]提出基于空域和时域切片分析的视频质量评价算法(ViS3)。半参考方法通过对比参考视频的部分信息进行评价，Soundararajan等[6]根据时空熵的差异提出了时空熵差异的半参考评价(RRED)算法。然而，在实际中，获取原参考视频是极其困难的，全参考和半参考视频质量评价的算法应用起来比较困难，因此，应重点研究无参考视频质量评价。

由于无参考方法相较于前两者有一定的难度，所以，当前无参考视频质量评价的方法相对较少。文献[7]利用小波变换进行失真JPEG2000视频的

**收稿日期**：2018-03-30；**修回日期**：2018-04-08；**录用日期**：2018-04-18

**基金项目**：国家自然科学基金(61673194，61672265)、江苏省产学研前瞻性联合研究(BY2016022-17/001)、江苏省自然科学基金(BK20171142)

* **E-mail**：sangqb@163.com

评价；Mittal 等[8]提出了基于时域统计的自然失真视频评价方法；Li 等[9]根据三维(3D)-DCT 频域的统计特征进行视频评价；张淑芳等[10]提出了基于背景消除的小波域视频质量评价模型；Wang 等[11]提出了通过迁移 AlexNet 网络提取视频空域失真特征预测视频评分的方法。

上述方法主要是针对单个失真类型的视频进行质量评价，或者利用迁移学习在失真图像上微调网络后提取视频空域失真特征，融合时域特征进行回归预测，不仅步骤繁琐，而且未实现端到端的预测。本文借鉴 VGG-16 网络[12]结构，迁移部分参数在自己的网络结构中微调，构建只需要输入失真视频即可得到客观视频质量评分的端到端视频质量评价模型。其主要贡献如下：1) 参考 VGG-16 网络卷积层，构建用于视频质量评价的深度卷积神经网络模型；2) 通过迁移 VGG-16 网络卷积层的参数到本文视频质量评价模型中来解决失真视频训练样本不足，且训练时间过长不易收敛的问题；3) 样本处理过程中，通过视频分割切块并进行拼接，解决神经网络直接训练失真视频的难题。本文网络模型不仅能够提取失真视频的空域信息，而且能够提取其时域信息，最终能够像处理图像一样处理视频信息，实现端到端的视频质量评价。实验结果表明，本文方法与主观评分具有很好的一致性。

## 2 质量评价模型

### 2.1 迁移学习

在计算机视觉领域，迁移学习是有效的，通过学习稀疏标记，能够在适用领域小数据的场景下进行特征学习[13-15]。迁移学习也叫领域适应，通常是将从源领域学习到的东西应用到目标领域。源领域和目标领域之间的数据分布通常有一点差异。简单地讲，迁移学习就是通过对现有表现优异的模型算法进行微调以使其适应于新领域的一种方法。迁移学习方法分为样本迁移、特征迁移、模型迁移和关系迁移。

LIVE 视频数据库中共有 150 个失真视频，属于典型的小样本训练。传统的端到端训练不易拟合，视频质量评价结果与人眼主观性很难实现一致性。本文采用模型迁移的方法进行失真视频训练，不仅适用于小规模的数据集，而且可以减少训练时间，最重要的是效果很好。

### 2.2 网络结构

VGGNet 是牛津大学计算机视觉组和 Google Deep Mind 公司的研究人员共同研发的深度卷积神经网络。VGGNet 研究了卷积神经网络的深度与性能之间的关系，通过使用小型卷积核(3×3)和最大池化层(2×2)不断重复地叠加，成功地构筑了 16 和 19 层深的卷积神经网络。由于 VGGNet 的拓展性很强、结构非常简洁，以及模型参数迁移到其他数据样本上的泛化性非常好，所以，VGGNet 依然经常被用于提取小样本图像数据的特征。本研究借鉴 VGG-16 网络结构，构造实现端到端的视频质量评价网络，并迁移 VGG-16 卷积层参数，对其进行微调。具体网络结构如图 1 所示。

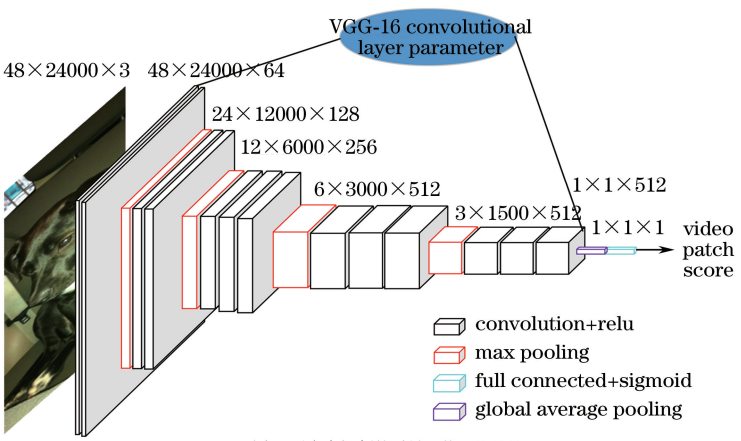


图 1 无参考视频的质量评价网络结构

Fig. 1 Network structure of no reference video quality assessment

### 2.3 实验数据处理

实验采用 LIVE 视频库[16]和 CSIQ 视频库[5]。视频类型为 YUV 格式，这种格式的视频亮度和色度信号是分开的，且网络传输占用极少带宽，即 YUV 格式比 RGB 存储占用空间小。为了提取整个失真视频的特征信息，将 YUV 视频转换成 RGB 图片帧的形式，图 2 为失真视频转换为 RGB 格式图片示例。具体转换公式如下：

$$R=Y+1.403\times(V-128), \tag{1}$$

$$G=Y-0.343\times(U-128)-0.714\times(V-128), \tag{2}$$

$$B=Y+1.770\times(U-128), \tag{3}$$

式中 $Y$ 为亮度信号值，$U$、$V$ 为色度信号值。$R$、$G$、$B$ 和 $Y$、$U$、$V$ 的取值范围均为 0～255。


图 2 失真视频的第一帧 RGB 图片。(a)模糊的天空；(b)步行街；(c)河床；(d)高峰时期

Fig. 2 The first frame RGB image of distortion video. (a) Blur sky; (b) pedestrian street; (c) riverbed; (d) rush hour

本实验有两大难点：1) 样本数据少且复杂；2) 如何将样本传送给卷积神经网络进行端到端的训练。具体解决方法如下。

1) 视频分块。LIVE 视频库中视频分辨率大小为 768 pixel×432 pixel。对每一帧图片进行不重叠的分割切块，视频块的大小取 48 pixel×48 pixel。视频库中有 3 种视频帧数，即 217，250，500 frame。为了方便之后的拼接、不丢失原始失真信息，统一按照 500 frame 等长的方式进行切割，不足的帧数用 0 填充。视频帧长和宽采样点的个数分别为 16 和 9，最终采样生成 144 个小的视频块。

2) 视频块拼接。视频分块后的视频块大小为 48 pixel×48 pixel×1500 pixel。为了适应网络层 3 通道的输入要求，对视频块进行通道分离，建立一个 48 pixel×48 pixel×3 pixel×500 pixel 的 4 维矩阵进行存放视频块拼接。分离后，将每一帧图像块依次保存到矩阵中，用于本次网络的端到端的训练。

图 3 所示为视频数据处理流程。

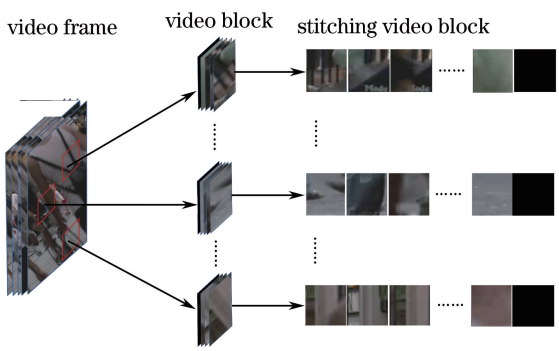


图 3 视频数据处理流程

Fig. 3 Flow chart of video data processing

### 2.4 网络微调和参数设置

参考 VGG-16 网络进行微调，以适应端到端的视频质量评价。对于卷积神经网络[17]，低层的特征信息往往具有通用性，而高层的信息则具有抽象性。根据不同的任务，如分割、识别、定位等，微调不同卷积层来提取有效的抽象信息。

VGG-16 网络一共有 16 层，而本实验用到的只有全卷积部分的参数，所以需要删除后面的 3 层全连接层，剩下 13 层卷积，利用高层的视频帧的失真特征来进行视频质量分数的拟合。在本文网络中，对前 10 层卷积网络不进行训练，只微调后面 3 层卷积。这样针对端对端的训练不仅可以提取到有效失真特征，而且因为训练参数减少而缩短训练时间，达到快速收敛的目的。本实验参考 VGG-16 网络参数设置如表 1 所示。

表 1 卷积层参数

Table 1 Convolution layer parameters

| Index | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Kernel size | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Feature map | 64 | 64 | 128 | 128 | 256 | 256 | 256 | 512 | 512 | 512 | 512 | 512 | 512 |
| Training | F | F | F | F | F | F | F | F | F | F | T | T | T |

表 1 中 C 代表卷积层，卷积层之间都采用填充为 2 的措施，每一次卷积滑动窗口的步长为 1。C2C3、C4C5、C7C8、C10C11 卷积层中间的池化操作全部采用“max pooling”，激活函数全部采用修正线性单元(Relu)，

$$o_k=\max(x_{kij}),i,j\leqslant r, \tag{4}$$

式中 $k$ 代表第 $k$ 个通道，$x_{kij}$ 代表池化区域$(i,j)$的值，$r$ 代表池化大小。

$$o_k=\begin{cases}x_k, & \text{if } (x_k)>0\\ 0, & \text{if } (x_k)\leqslant 0\end{cases}, \tag{5}$$

式中 $x_k$ 代表第 $k$ 个神经元的输入。

在最后一层卷积后，未使用 VGG-16 网络原始的“max pooling”池化方式，而是采用全局平均值池化[18]，其主要思想是对每一个特征图进行一次特征映射。在最顶层卷积后的特征图上，取其平均值得到结果向量，然后直接送入全连接层进行回归预测。全局平均值池化的好处在于：一是通过在特征映射和训练回归之间建立对应关系，更适合于提取特征的卷积结构；二是在全局平均值池化过程中不需要对参数进行训练学习，可以减少过拟合的风险；三是全局平均值池化可以计算出空间信息，对翻译输入的空间更加稳健；四是可以降低最后一层的纬度空间，减少训练参数，加快收敛效果。全局平均值池化计算公式为

$$o_k=\text{Average}\left(\sum x_{kij}\right),i,j\leqslant w,h, \tag{6}$$

式中 $k$ 代表第 $k$ 个通道，$x_{kij}$ 代表特征图区域$(i,j)$的值，$w$ 代表特征图的宽，$h$ 代表特征图的高。

在全局平均池化层后接入一个神经元用于预测视频质量的评分。原始给定的差异平均意见分(DMOS)取值范围为 0～10，对其除以 10 进行归一化处理。激活函数采用 sigmoid 函数预测视频质量评分：

$$o_k=\frac{1}{1+\exp(-x_k)}, \tag{7}$$

式中 $x_k$ 代表第 $k$ 个神经元的输入。本次视频质量评价网络的参数约有 150 万，因为只需要微调后面 3 层卷积层和 1 层全连接层，所以网络需要训练的参数约为 47 万，缩短了训练时间。

过拟合是很多机器学习的通病，在训练集上过拟合得到的模型在测试集上的表现效果很差。机器学习中通常采用集成的方法解决过拟合问题，但是训练多个模型进行集成组合，将浪费大量的时间和空间。在深度学习中，“dropout”可以很好地解决这些问题，每一次“dropout”相当于从原始大的网络中以一定概率随机挑选一个小的网络进行参数更新。因此，对于一个有 $N$ 个节点的网络层来说，使用“dropout”技术后，可以看作 $2N$ 个小网络的集合。类似于集成，但此时要训练更新的参数数目是不变的，训练时间和参数占用的空间并没有增加。本实验采用该技术来减少过度拟合的风险。实验中“dropout”取值为 0.5，即每一个神经元参数将会有 0.5 的概率被随机采样参与网络训练更新 。

本次实验的优化器采用自适应时刻估计(Adam)优化算法[19]。在深度学习实际应用中，Adam 方法表现效果良好。与其他自适应学习率算法，如 Adagrad、AdaDelta 相比，其收敛速度更快，学习效果更为有效，而且可以纠正其他优化技术中存在的问题，如学习率消失、收敛过慢，以及高方差的参数更新导致损失函数波动较大等问题。由于本研究采用 500 frame 等长的视频数据作为输入，输入数据集比较稀疏，常见的随机梯度下降(SGD)、Nesteroy 梯度加速(NAG)和动量项等方法效果可能不佳，所以选择 Adam 优化算法，该方法可以进行自适应调整学习率。由于是微调网络，所以本研究未采用默认初始学习率，而是将其设置为 0.0001，批量(batch_size)设置为 10。

$$\hat{m}_t=\frac{m_t}{1-\beta_1^t}, \tag{8}$$

$$\hat{v}_t=\frac{v_t}{1-\beta_2^t}, \tag{9}$$

$$\theta_{t+1}=\theta_t-\frac{\eta}{\sqrt{\hat{v}_t}+\varepsilon}\hat{m}_t, \tag{10}$$

式中$\hat{m}_t$ 为梯度的第一时刻平均值，$\hat{v}_t$ 为梯度的第二时刻非中心方差值，$\theta$ 为梯度更新值，$\beta_1=0.9$，$\beta_2=0.9999$，$\varepsilon=10^{-8}$。

损失函数使用均方误差进行计算：

$$E_{\text{loss}}=\frac{1}{2N}\sum_{n=1}^{N}\left\|Y_n-\hat{y}_n\right\|_2^2, \tag{11}$$

式中 $E_{\text{loss}}$ 代表损失值，$Y_n$ 代表视频真实的主观得分，$\hat{y}_n$ 代表网络的预测值，$N$ 表示样本个数。

## 3 实验结果与分析

### 3.1 视频数据库

在德克萨斯州奥斯汀分校图像和视频工程实验室公布的 LIVE 视频数据库上进行实验。视频库中包含 160 个视频，其中 10 个为原始视频，每一个原始视频有 15 个失真视频。LIVE 视频库包含无线网络传输失真、互联网协议(IP)传输失真、H.264 压缩失真、MPEG-2 压缩失真等 4 种失真类型。所有的视频文件都以 YUV 4:2:0 格式存储，分辨率为 768 pixel×432 pixel，每个视频都提供 DMOS 和其标准偏差值。

为验证本文算法的通用性，同时采用俄克拉荷马州立大学视觉实验室提供的主观视频质量数据库 CSIQ[2] 进行算法验证。CSIQ 视频数据库共包含 228 个视频，其中 12 个为原始视频，216 个为失真视频。该数据库包含 6 种类型的失真：H.264 压缩、HEVC 压缩、动态 JPEG 压缩(MJPEG)、基于小波的 SNOW 压缩、无线传输、白噪声。所有的视频文件都以 YUV 4:2:0 格式存储，分辨率为 832 pixel×480 pixel，每个视频都提供 DMOS 和其标准偏差值。

### 3.2 评价方法指标

为了衡量本文算法在客观视频质量评价上的性能，将预测结果与主观视频评价分数进行比较，并引入常用的线性皮尔逊相关系数(LPCC)和斯皮尔曼等级相关系数(SROCC)来检验视频的客观预测值与主观评价值之间的相关性。计算公式为

$$D_{\text{LPCC}}=\frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}}, \tag{12}$$

式中 $X_i$ 和 $Y_i$ 分别表示第 $i$ 个视频的主观评价值和客观预测值，$\bar{X}$ 和 $\bar{Y}$ 分别表示主观评价值和客观预测值的均值，$N$ 表示失真视频的数量。LPCC 评价的是预测的精度，即准确度，取值范围为[0,1]，越接近 1，表明预测的准确度越高，反之，表明预测的准确度越低。

$$D_{\text{SROCC}}=1-\frac{6\sum_{i=1}^{N}d_i^2}{N(N^2-1)}, \tag{13}$$

式中 $N$ 表示失真视频的数量，$d_i$ 表示第 $i$ 个失真视频的主观评价值与客观预测值之间的差异。SROCC 评价预测的单调性，取值范围为[−1,1]，绝对值越接近 1，表明单调性越好。

### 3.3 实验结果与分析

在视频数据库中随机选取 80%的视频作为训练集，10%作为验证集，10%作为测试集，测试集和训练集之间无交叉。实验中为了达到扩充数据集的目的，切割每个视频块大小为 48 pixel×48 pixel，因而每幅视频就可以固定点不重叠采样 144 块，并将这些视频块拼接起来。视频质量评价需要考虑视频帧与帧之间的时序关系，拼接之后，用 VGG-16 进行滑动卷积，以解决视频帧的时序问题。因此，一共得到 21600 个小视频拼接块，可以直接送入网络进行端到端的训练。为了加快样本收敛，对视频拼接块进行“min-max”归一化：

$$\hat{x}=\frac{x-x_{\min}}{x_{\max}-x_{\min}}, \tag{14}$$

式中 $\hat{x}$ 为归一化后输入数据，$x$ 为原始视频数据，$x_{\min}$ 和 $x_{\max}$ 分别为原始视频数据中的最小值和最大值。

把训练好的模型在测试集上进行预测。首先，预测每个失真视频块的质量得分，然后，通过求取其所有子块的平均分得到整个视频质量的预测得分。为了验证算法的稳健性，采用不同的随机数种子来打乱数据，进行重新训练，然后取平均值作为本次实验的最终结果。最后，分别测试了 LIVE 视频数据库的不同失真类型，以及整体混合失真视频的客观得分。

将本文方法与主流的全参考和无参考视频质量评价算法进行比较，其中，全参考算法包括峰值信噪比(PSNR)、SSIM、视觉信息保真度(VIF)[20]、时空最明显失真模型(STMAD)[21] 和 ViS3，无参考算法包括无参考视频质量评价方法(V-Blinds)[22]、卷积神经网路＋多元回归(CNN＋MR)[10]、基于背景减法的小波域(BSWQ)[11] 等。从表 2、3 评价指标数据中可以看出，本文算法在 H.264 和 MPEG-2 压缩失真上与目前最好的指标相差无几，在 Wireless 和 IP 失真类型上优于其他算法。整体上，无论是拟合精度还是单调性都优于基于传统方法的全参考视频质量评价算法，以及无参考算法 V-Blinds、BSWQ。与最新利用卷积神经网络提取特征的 CNN-MR 算法相比，预测准确性相当，但拟合单调性有较大进步；同时，本文算法采用端到端的训练预测，优于 CNN-MR 算法先提取特征再进行预测的分段算法，预测效率也因此得到提升。

表 2 LIVE 数据库实验的 SROCC 相关性

Table 2 SROCC correlations of experiments on LIVE database

| Algorithm | SROCC | | | | |
|---|---|---|---|---|---|
| | Wireless | IP | H.264 | MPEG-2 | Mix |
| PSNR | 0.667 | 0.485 | 0.500 | 0.393 | 0.532 |
| SSIM | 0.738 | 0.600 | 0.631 | 0.512 | 0.664 |
| VIF | 0.571 | 0.486 | 0.762 | 0.619 | 0.573 |
| STMAD | 0.806 | 0.768 | **0.904** | 0.842 | 0.829 |
| ViS3 | 0.845 | 0.788 | 0.757 | 0.730 | 0.816 |
| V-Blinds | 0.691 | 0.771 | 0.738 | 0.881 | 0.798 |
| BSWQ | 0.848 | 0.802 | 0.756 | 0.731 | 0.826 |
| CNN-MR | 0.833 | 0.771 | 0.857 | 0.881 | 0.821 |
| Proposed | **0.857** | **0.845** | 0.868 | **0.893** | 0.867 |

表 3 LIVE 数据库实验的 LCC 相关性

Table 3 LCC correlations of experiments on LIVE database

| Algorithm | LCC | | | | |
|---|---|---|---|---|---|
| | Wireless | IP | H.264 | MPEG-2 | Mix |
| PSNR | 0.831 | 0.853 | 0.741 | 0.668 | 0.594 |
| SSIM | 0.822 | 0.769 | 0.705 | 0.534 | 0.700 |
| VIF | 0.789 | 0.884 | 0.858 | 0.756 | 0.637 |
| STMAD | 0.806 | 0.769 | 0.904 | 0.847 | 0.829 |
| ViS3 | 0.846 | 0.816 | 0.773 | 0.746 | 0.829 |
| V-Blinds | 0.806 | 0.816 | 0.892 | 0.904 | 0.831 |
| BSWQ | 0.871 | 0.835 | 0.807 | 0.753 | 0.843 |
| CNN-MR | 0.917 | 0.913 | **0.929** | 0.933 | 0.852 |
| Proposed | **0.912** | **0.920** | 0.895 | 0.915 | 0.843 |

### 3.4 算法通用性

为了验证本文算法的可靠性，在 CSIQ 视频数据库中进行验证。具体训练方法和预测采用与 LIVE 视频库上相同的方法。根据表 4 评价指标数据，本文算法模型在其他视频库上仍然可以很好地预测失真视频的客观分数，优于目前大部分视频质量评价算法。这进一步说明了本文算法具有很好的通用性，预测视频客观得分能够和人眼主观评价结果达成很好的一致性，并且实现了端到端的训练和预测，省去了分阶段特征提取。本文端到端网络结构简洁明了、容易实现，便于以后改进网络结构，以更好地适用于视频质量评价。

表 4 CSIQ 数据库实验的 SRCC 和 LCC 相关性

Table 4 SRCC and LCC correlations of experiments on CSIQ database

| Algorithm | SRCC | LCC |
|---|---|---|
| PSNR | 0.578 | 0.583 |
| SSIM | 0.701 | 0.718 |
| VIF | 0.653 | 0.642 |
| STMAD | 0.817 | 0.826 |
| ViS3 | 0.841 | 0.830 |
| V-Bliinds | 0.836 | 0.832 |
| BSWQ | 0.841 | 0.826 |
| Proposed | **0.859** | **0.845** |

## 4 结 论

提出一种基于参考 VGG-16 网络，迁移参数进行端到端学习的视频质量评价模型。由于 VGG-16 卷积层具有出色的特征提取能力，仅通过对其参数进行微调，即可很好地解决视频训练样本过少的问题，同时加快了训练速度，提高了算法的性能和稳定性，为后续模型的改进和特征提取打下了基础。实验结果表明，本文方法优于现有的一些视频质量评价模型，结构简单，且容易训练。但是，卷积操作并不能充分反映帧与帧的时序关系，而循环神经网络能够很好地体现时序问题，所以，今后将采用长短期记忆网络提取视频帧之间的特征关系，并开展相关实验验证。

## 参 考 文 献

[1] Hou C P, Ma T T, Yue G H, *et al*. Multiply-distorted image quality assessment based on high-order phase congruency[J]. Laser & Optoelectronics Progress, 2017, 54(7): 071001.
侯春萍, 马彤彤, 岳广辉, 等. 基于高阶相位一致性的混合失真图像质量评价[J]. 激光与光电子学进展, 2017, 54(7): 071001.

[2] Zhang Y, Jin W Q. Assessment method of fusion image quality in wavelet domain structural similarity [J]. Chinese Journal of Lasers, 2012, 39(s1): s109007.
张勇, 金伟其. 小波域结构相似度融合图像质量评价方法[J]. 中国激光, 2012, 39(s1): s109007.

[3] Xue X B, Yu M, He M L. Stereoscopic image-quality-assessment method based on visual cell model [J]. Laser & Optoelectronics Progress, 2016, 53(4): 041004.
薛小波, 郁梅, 何美伶. 基于仿视觉细胞模型的立体图像质量评价方法[J]. 激光与光电子学进展, 2016, 53(4): 041004.

[4] Wang Z, Bovik A C, Sheikh H R, *et al*. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[5] Vu P V, Chandler D M. ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices [J]. Journal of Electronic Imaging, 2014, 23(1): 013016.

[6] Soundararajan R, Bovik A C. Video quality assessment by reduced reference spatio-temporal entropic differencing [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 23(4): 684-694.

[7] Vu P V, Chandler D M. A no-reference quality assessment algorithm for JPEG2000-compressed images based on local sharpness[J]. Proceedings of SPIE, 2013, 8653: 865302.

[8] Mittal A, Saad M, Bovik A C. Assessment of video naturalness using time-frequency statistics[C]. IEEE International Conference on Image Processing, 2014: 571-574.

[9] Li X, Guo Q, Lu X. Spatiotemporal statistics for video quality assessment[J]. IEEE Transactions on Image Processing, 2016, 25(7): 3329-3342.

[10] Zhang S F, Huang X Q. Video quality assessment model in wavelet domain based on background subtraction [J]. Journal of Tianjin University (Science and Technology), 2017, 50(12): 1255-1261.
张淑芳, 黄小琴. 基于背景消除的小波域视频质量评价模型[J]. 天津大学学报(自然科学与工程技术版), 2017, 50(12): 1255-1261.

[11] Wang C, Su L, Huang Q. CNN-MR for no reference video quality assessment[C]. 2017 4th International Conference on Information Science and Control Engineering (ICISCE), 2017: 224-228.

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

[13] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[14] Patricia N, Caputo B. Learning to learn, from transfer learning to domain adaptation: A unifying perspective[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1442-1449.

[15] Courty N, Flamary R, Tuia D, *et al*. Optimal transport for domain adaptation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(9): 1853-1865.

[16] Seshadrinathan K, Soundararajan R, Bovik A C, *et al*. Study of subjective and objective quality assessment of video[J]. IEEE Transactions on Image Processing, 2010, 19(6):1427-1441.

[17] Li S M, Lei G Q, Fan R. Depth maps super-resolution reconstruction based on convolutional neural networks[J]. Acta Optica Sinica, 2017, 37(12): 1210002.
李素梅, 雷国庆, 范如. 基于卷积神经网络的深度图超分辨率重建[J]. 光学学报, 2017, 37(12): 1210002.

[18] Lin M, Chen Q, Yan S. Network in network[J]. arXiv, 2013: 1312.4400.

[19] Kingma D P, Ba J. Adam: a method for stochastic optimization[J]. arXiv, 2014: 1412.6980.

[20] Sheikh H R, Bovik A C. A visual information fidelity approach to video quality assessment[C]// The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2005: 23-25.

[21] Vu P V, Vu C T, Chandler D M. A spatiotemporal most-apparent-distortion model for video quality assessment [C] // 2011 18th IEEE International Conference on Image Processing (ICIP), 2011: 2505-2508.

[22] Saad M A, Bovik A C, Charrier C. Blind prediction of natural video quality[J]. IEEE Transactions on Image Processing, 2014, 23(3):1352-1365.