激光与光电子学进展 先进成像

基于自顶向下视觉机制的立体图像质量评价

李素梅，张慧林*

天津大学电气自动化与信息工程学院，天津 300072

摘要 为了提高立体图像质量客观评价指标与人类主观评价指标的一致性，受人类视觉自顶向下机制的启发，提出一种基于立体注意力的无参考立体图像质量评价方法。在所提立体注意力模块中：首先，通过所提双目融合模块中的能量系数自适应调整双目响应的幅度，并在空间和通道维度对双目特征进行并行处理；其次，所提双目调制模块能够同时实现高层级双目信息对低层级双目信息和低层级单目信息的自顶向下调制。此外，双池化策略分别对双目融合图和双目差异图进行处理，可获取更有利于质量分数回归的关键信息。基于公开的LIVE 3D和WIVC 3D数据库对所提方法进行了性能验证。实验结果表明，所提方法实现了客观评价指标与标签的较高一致性。

关键词 图像处理；立体图像质量评价；人类视觉系统；自顶向下；卷积神经网络

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP241231

Stereo Image Quality Assessment Based on Top-Down Visual Mechanism

Li Sumei, Zhang Huilin*

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract In order to improve the consistency of metrics between the objective assessment and the human subjective evaluation of stereo image quality, inspired by the top-down mechanism of human vision, this paper proposes a stereo attention-based no-reference stereo image quality assessment method. In the proposed stereo attention module. First, the amplitude of binocular response is adaptively adjusted by the energy coefficient in the proposed binocular fusion module, and the binocular features are processed simultaneously in the spatial and channel dimensions. Second, the proposed binocular modulation module realizes the top-down modulation of the high-level binocular information to the low-level bino- and monocular information simultaneously. In addition, the dual-pooling strategy proposed in this paper processes the binocular fusion map and binocular difference map to obtain the critical information that is more conducive to quality score regression. The performance of the proposed method is validated based on the publicly available LIVE 3D and WIVC 3D databases. The experimental results show that the proposed method achieves high consistency between objective assessment indices and labels.

Key words image processing; stereo image quality assessment; human visual system; top-down; convolutional neural network

1 引 言

近年来，立体图像质量评价技术受到越来越多研究者的关注。立体图像质量评价方法主要包括主观评价[1-3]和客观评价[4-6]。主观评价依据受试者的主观感受对立体图像的质量进行评分，虽然可靠性高但耗时耗力。客观评价则利用算法自动、高效地评估立体图像的质量。立体图像质量客观评价方法根据其对参考图像信息的依赖程度不同可以分为全参考、半参考和无参考等3类，本文主要研究无参考立体图像质量评价(NR-SIQA)方法。

依据特征提取方式的不同，NR-SIQA方法可分为基于传统方法的质量评价方法和基于深度学习的质量评价方法。鉴于立体图像与人类主观感知的密切相关性[7-9]，在基于传统方法的质量评价中，较多研究人员基于人类视觉系统(HVS)特性对质量特征进行手工提取[6,10-11]，但该类方法所提取特征不够准确。

随着深度学习技术的飞速发展，卷积神经网络

收稿日期：2024-05-07；**修回日期**：2024-05-13；**录用日期**：2024-05-16；**网络首发日期**：2024-05-17

基金项目：国家自然科学基金面上项目(61971306)

通信作者：*hl_zhang@tju.edu.cn

(CNN)成为质量评价的主流模型[12-18],一些研究人员开始基于CNN和部分视觉特性搭建立体图像质量评价模型。例如:Zhou等[15]和Yan等[16]分别提出了双流网络和多级特征融合网络来模拟HVS的多级交互现象,二者均在网络的不同阶段通过融合单目特征得到双目特征;Shen等[17]通过交叉堆叠操作模拟视觉皮层V1区的左右单目信号的首次交互融合;Si等[18]设计了双目交互模块和双目融合模块,分别模拟视觉感知过程中持续存在的双目交互与融合现象。上述方法通过对部分视觉机制的模拟取得了较好的质量评价性能。

然而事实上,人类大脑中同时存在自底向上[19]和自顶向下[20]这两种推理机制。上述方法只对自底向上机制进行了模拟,未考虑自顶向下的反馈调制机制。随着研究的不断深入,研究人员开始尝试在CNN模型中模拟自顶向下特性。例如:Chang等[21]从由粗到细的角度模拟了双目信息对单目信息的自顶向下反馈调制;Chang等[22]则模拟了双目信息之间的自顶向下调制。实际上,单、双目信息在HVS中会被同时处理[23-24],而Chang的方法未同时模拟高级双目信息对低级双目信息、高级双目信息对低级单目信息的自顶向下机制。

为了更好地模拟该特性,本文提出一种基于立体注意力的NR-SIQA方法。在立体注意力模块中设计的双目调制模块可同时实现高层级双目信息对低层级双目信息和低层级单目信息的自顶向下调制。所提双目融合模块中的能量系数可自适应调整双目信息的幅度,以模拟双目响应幅度小于两个单目响应幅度之和的实际现象[23-24]。此外,双池化策略可筛选出关键信息用于质量回归。实验结果表明,所提方法与主观评价具有高度一致性。

2　立体图像质量评价网络

所提网络的总体结构如图1所示,主要分为4个部分:初级特征提取、单目分支、立体注意力块(SAT)、双池化及质量回归。

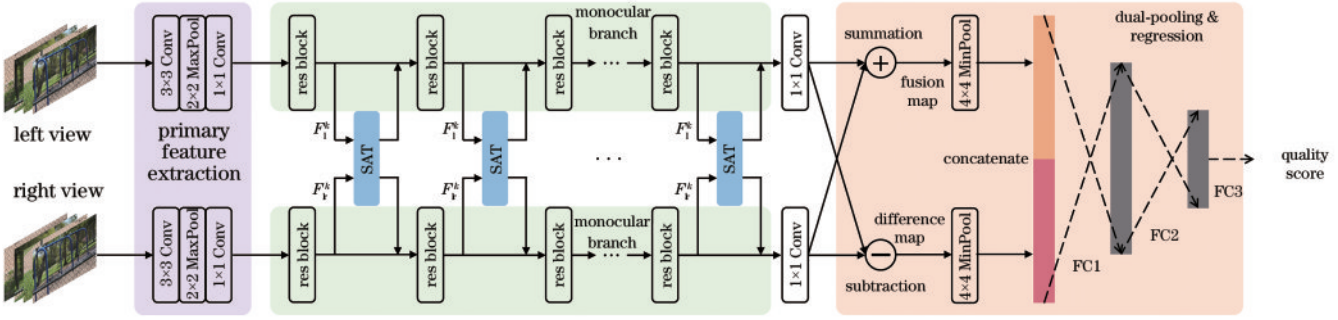

图1　所提方法结构图

Fig. 1　The architecture of the proposed method

对立体图像的处理流程如下:首先,通过初级特征提取部分分别提取左右视点的初级特征,并对特征图进行一次下采样。接着,采用ResNet[25]作为两个单目分支的骨架网络,将每一对残差块及其跟随的立体注意力块链式堆叠,完成单目信息的提取和双目信息的聚合,以及较高层级双目信息对较低层级双目信息、较高层级双目信息对较低层级单目信息的自顶向下调制。随后,分别使用两个卷积层对左右两个单目分支的输出解码,将左右两路特征图分别进行加、减操作,得到双目融合图和双目差异图。最后,使用双池化策略对双目特征图进行下采样,通过全连接层实现拼接特征到质量分数的映射。

2.1　初级特征提取

初级特征提取部分在左右支路上均由两个卷积层和一个最大池化层组成,且两支路不共享参数。两层卷积层的卷积核尺寸分别为3×3和1×1,可提取到左右两支路的初级特征。利用两个卷积层中间的一个2×2最大池化层对特征图进行下采样,以提高后续处理效率。

2.2　单目分支

为了提高特征的表征与学习能力,避免梯度消失问题,采用7个残差块链式堆叠而成的ResNet作为左右单目分支的骨架网络。而且,与最初的残差块[25]不同,骨架网络采用带有预激活的残差块,即在卷积操作之前进行非线性激活,可以增强恒等映射[26],进而提高网络性能。所提方法在同一层级左右单目支路的残差块后插入了一个立体注意力模块,用来将单目特征聚合为双目注意力图,并完成两种自顶向下的调制,故网络中立体注意力模块的数量也为7。

2.3　立体注意力

人类大脑在进行信息处理时同时存在自底向上和自顶向下两种推理机制。自底向上是指来自外界刺激的视觉信号由眼睛经视神经传输到初级视觉皮层(V1)进而到达高级视觉皮层(V2、V3、V4、MT),形成初步的视觉感知[19];自顶向下则是指在高级视觉皮层中得到的先验信息反馈至较低的视觉皮层,进一步对视觉信息进行筛选和处理,以便更好地完成视觉感知任务[20]。大脑在两种视觉机制的共同作用下实现对信

息的准确感知。然而，除去文献[21-22]外，在NR-SIQA领域尚未有对两种视觉机制同时模拟的方法。其中，文献[21]模拟了高层级双目信息对低层级单目信息的反馈指导，文献[22]模拟了高层级双目信息对低层级双目信息的反馈指导。事实上，在各级视觉皮层中既有单目细胞又有双目细胞，即自顶向下的反馈指导可能存在多种形式，不确定是高层级双目信息对低层级单目信息或高层级双目信息对低层级双目信息的反馈指导。基于此，所提方法提出了一种全新的立体注意力，它能够同时实现高层级双目信息对低层级双目信息、低层级单目信息的自顶向下调制。如图2所示，SAT包含两个部分：双目融合模块(BFM)和双目调制模块(BMM)。

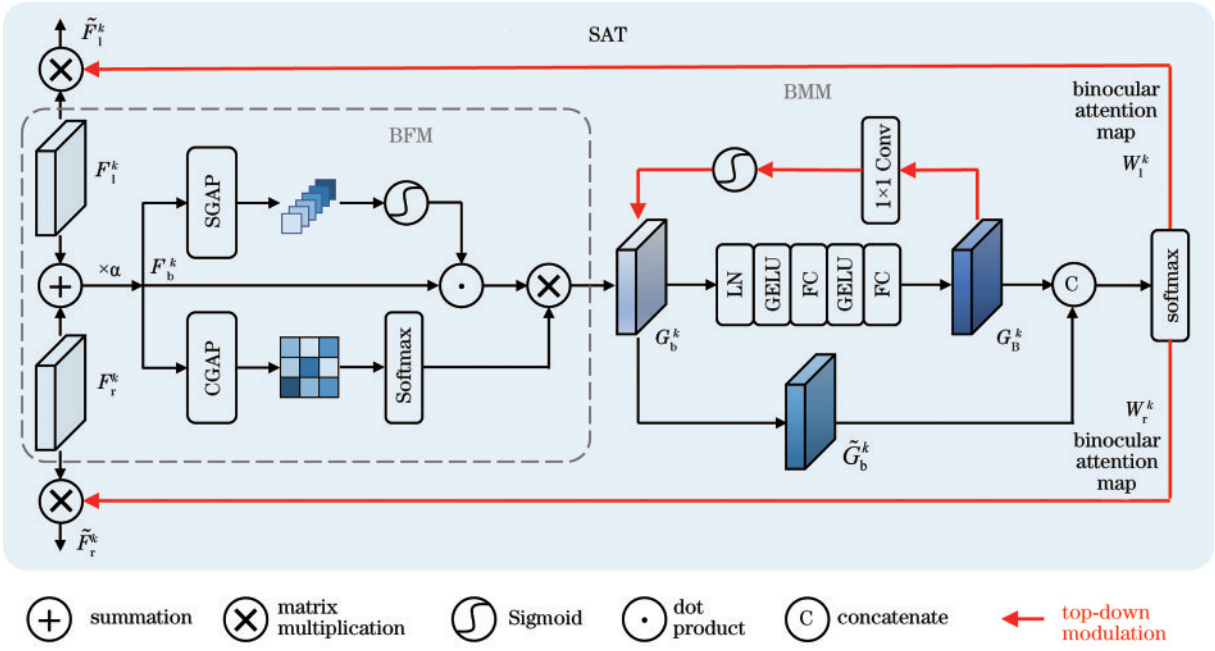

图2　立体注意力结构图

Fig. 2　The architecture of the SAT

双目融合模块可实现两个功能：将左右单目分支输出的两个单目特征动态融合为双目特征，并进一步提取关键信息。生物学的最新研究指出，双目响应的幅度应小于两个单目响应的幅度之和[23-24]，因此，为了对此现象进行模拟，在双目融合过程中引入了能量系数来动态控制双目特征的幅度。具体地，对于第 k ($k=1,2,\cdots,7$)个SAT的左右单目特征 F_{l}^{k} 和 F_{r}^{k}，将其求和后与能量系数 α 相乘得到第 k 个SAT的双目特征 F_{b}^{k}：

$$F_{\mathrm{b}}^{k}=\alpha\times\left(F_{\mathrm{l}}^{k}+F_{\mathrm{r}}^{k}\right) \tag{1}$$

式中：α 是一个值被Sigmoid函数限制在0～1之间的可学习的参数；×表示像素级乘法；+表示像素级加法。α 可动态变化，例如SAT在网络中处于不同的位置(k)，或输入特征 F_{l}^{k} 和 F_{r}^{k} 的不同，都会使 α 收敛到不同的值。因此在模型推理时，α 可以自适应地控制双目响应的幅度，使其更符合人类的视觉感知过程，提高模型性能。

在动态融合之后，需要进一步对得到的双目特征 F_{b}^{k} 进行特征提取。将特征提取后的双目特征记为 G_{b}^{k}，提取过程如下：

$$G_{\mathrm{b}}^{k}=\mathrm{Softmax}\left[\mathrm{CGAP}\left(F_{\mathrm{b}}^{k}\right)\right]\otimes\left\{F_{\mathrm{b}}^{k}\odot\sigma\left[\mathrm{SGAP}\left(F_{\mathrm{b}}^{k}\right)\right]\right\} \tag{2}$$

式中：CGAP(·)指通道全局平均池化；SGAP(·)指空间全局平均池化；σ 表示Sigmoid函数；⊗与⊙分别表示矩阵乘法与点乘操作。此过程实现了对双目特征在通道和空间维度的并行提取，对双目信息进行了聚合与增强。

双目调制模块通过反馈支路实现较高层级双目信息对较低层级双目信息与较低层级单目信息的自顶向下调制。具体地，将双目特征图 G_{b}^{k} 通过前馈支路的非线性变换得到更高层级的双目特征 G_{B}^{k}，再将 G_{B}^{k} 进行一次非线性拟合并通过反馈支路作用于较低层级双目特征 G_{b}^{k}，得到矫正后的双目特征图 $\tilde{G}_{\mathrm{b}}^{k}$：

$$G_{\mathrm{B}}^{k}=g\left(G_{\mathrm{b}}^{k}\right) \tag{3}$$

$$\tilde{G}_{\mathrm{b}}^{k}=G_{\mathrm{b}}^{k}\odot\sigma\left[\mathrm{Conv}\left(G_{\mathrm{B}}^{k}\right)\right] \tag{4}$$

式中：$g(\cdot)$ 表示前馈支路的一系列非线性拟合操作，包括层归一化(LN)、激活函数GeLU、FC、GeLU、FC；σ 表示Sigmoid函数；Conv(·)表示1×1卷积；⊙表示点乘操作。$g(\cdot)$ 采用预归一化与预激活的方式更好地对双目特征进行非线性变换，增强特征的表达能力。

如图2所示，将 $\tilde{G}_{\mathrm{b}}^{k}$ 与 G_{B}^{k} 拼接后送入SAT的输出函数中，生成高级的双目注意力图。与大多数注意力模块(例如SE[27]和CBAM[28])中采用Sigmoid激活函数不同，所提方法采用Softmax函数。通过指定

Softmax 函数的输出个数为 2，可生成两个和为 1 的权重 W_{l}^{k} 和 W_{r}^{k}，分别对应于左右两个单目支路的较高层级双目注意力图。上述推导过程可表示为

$$\left(W_{\mathrm{l}}^{k}, W_{\mathrm{r}}^{k}\right)=\mathrm{Softmax}\left(\left[\!\left[G_{\mathrm{B}}^{k}, \tilde{G}_{\mathrm{b}}^{k}\right]\!\right]\right),\ W_{\mathrm{l}}^{k}+W_{\mathrm{r}}^{k}=1 \tag{5}$$

式中：$[\![\cdot]\!]$ 表示拼接操作。生成的较高层级双目注意力图 W_{l}^{k} 和 W_{r}^{k} 作为双目自顶向下调制信号，分别作用于两个低级单目特征 F_{l}^{k} 和 F_{r}^{k} 上，并指导其特征矫正。该调制过程可表示为

$$\tilde{F}_{m}^{k}=W_{m}^{k}\otimes F_{m}^{k},\ m\in\{\mathrm{l},\mathrm{r}\} \tag{6}$$

式中：$\tilde{F}_{m}^{k}$ 是矫正后的单目特征，作为立体注意力模块的输出；$\otimes$ 表示矩阵乘法操作。在推理阶段，高级双目注意力图与 α 均会根据输入的不同而自适应改变，还会根据该 SAT 在网络中所处的位置（k）不同而动态调整，这种设计的内在合理性尤其有助于网络处理非对称失真的立体图像。

2.4 双池化及质量回归

在网络中最后一个立体注意力模块完成对单目分支输出特征的自顶向下调制后，采用两个卷积核大小为 1×1 的卷积层分别对左右两个输出特征进行解码，以便进一步对其进行处理与回归。

类似于大多数 NR-SIQA 方法，所提方法对左右解码特征进行求和与相减操作，分别得到融合图与差异图。融合图包含左右视点的失真信息，差异图包含大量的视差信息，二者都含有失真图像的语义信息。因此，采用双池化策略对融合图与差异图进行处理。具体地，对融合图进行 4×4 最小池化，对差异图进行 4×4 最大池化。双池化策略可以在减少参数的同时筛选出最关键的质量信息用于质量分数回归。

最后，将池化后的两个质量特征拉平成一维向量并拼接在一起，使用三层全连接层将向量回归为该立体图像的预测质量得分。前两层全连接层后都有 ReLU 激活函数对特征进行非线性变换，同时采用 dropout（dropout 率为 0.5）来避免网络过拟合。

3 实验设计与结果分析

3.1 数据库及评价指标介绍

本实验选取 LIVE 3D 与 WIVC 3D 数据库。LIVE 3D 数据库包含 LIVE 3D Phase I[29] 与 LIVE 3D Phase II[30]，二者均包含 5 种失真类型：高斯模糊（BLUR）、快衰落（FF）、JPEG 2000 压缩（JP2K）、JPEG 压缩（JPEG），以及加性高斯白噪声（WN）。立体图像分辨率均为 640×360，均以差异平均意见分数（DMOS）作为主观质量评价结果，DMOS 值越小，图像质量越好。LIVE 3D Phase I 数据库包含 20 个参考立体图像及其深度图和视差图，还包含 365 个对称失真的立体图像。LIVE 3D Phase II 数据库包含 8 个参考立体图像及其深度图和视差图，对于每种失真类型，均包含 3 个对称失真和 6 个非对称失真的立体图像，因此共有 120 个对称失真和 240 个非对称失真的立体图像。

WIVC 3D 数据库包含 WIVC 3D Phase I[31] 与 WIVC 3D Phase II[32]，二者均包含 3 种失真类型：WN、BLUR 和 JPEG，每个参考图像的每种失真类型均包含 4 种不同失真程度的失真图像，均包含对称失真与非对称失真，且均提供平均意见分数（MOS）为主观质量评价结果，MOS 值越大，图像质量越好。WIVC 3D Phase I 包含 6 个参考立体图像和 324 个失真立体图像，图像分辨率为 1390×1080 和 1342×1080。WIVC 3D Phase II 包括 10 个参考立体图像和 450 个失真立体图像，其图像分辨率为 1920×1080。

采用的 3 个评价指标分别为皮尔逊线性相关系数（PLCC）、斯皮尔曼秩序相关系数（SROCC）与均方根误差（RMSE）。PLCC 用于衡量预测结果与标签之间的线性相关性，SROCC 用于衡量预测结果与标签之间的单调一致程度，RMSE 用于衡量预测结果与标签之间的数值偏差程度。PLCC 与 SROCC 的绝对值越大，RMSE 越趋近于 0，表明算法的预测结果与主观评分之间的一致性越好。

3.2 网络训练

所提方法基于 PyTorch 框架搭建模型，并在 NVIDIA GeForce RTX 4090 GPU（24 GB 内存）上训练 100 个迭代周期。初始学习率设置为 0.0001，并采用带预热的余弦退火衰减策略更新学习率。批量大小设置为 64，使用 Adam 优化器更新网络参数，动量值采用默认值，$\beta_1=0.9$、$\beta_2=0.999$，权重衰减设置为 0.0004。所提方法中在训练过程中采取 k 折交叉验证，取 $k=10$ 并采用其均值。损失函数为 L2 损失。此外，在计算评价指标时采用了五参数的逻辑函数[33]进行非线性映射。

3.3 性能比较与分析

为了验证所提方法的性能，进行了多组实验并与现有方法进行性能比较。鉴于多数方法均未开源，且不同方法的实验条件与数据划分各不相同，所以在各类实验结果中，对比方法的指标均来自相应论文，其中，性能最优的指标用加粗字体表示。首先，对所提方法与现有方法在 LIVE 3D 和 WIVC 3D 数据库上进行了整体性能的比较，实验结果如表 1 和表 2 所示。

从表 1 和表 2 可以看出，所提方法在 LIVE 3D 和 WIVC 3D 数据库上均取得了较优的性能，与最新的方法相比，各项指标均具有很大的竞争力。值得注意的是，文献[18]方法也具有优良的性能，这归功于其复杂的双目交互模块设计与较大的计算开销，文献[21]、[22]方法也拥有具有竞争力的性能指标，这得益于二者对自顶向下推理机制的模拟、较复杂的模块设计与高昂的计算成本，上述方法的复杂度比较见第 3.6 节。

本文还在 LIVE 3D 数据库上进行了单失真类型

表 1 不同方法在 LIVE 3D 数据库上的性能比较

Table 1 Performance comparison on LIVE 3D database

Method	LIVE 3D Phase I			LIVE 3D Phase II		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Ref. [30]	0.8950	0.8910	7.2470	0.8800	0.8800	5.1020
Ref. [12]	0.9470	0.9430	5.3360			
Ref. [15]	0.9730	0.9650	3.6820	0.9570	0.9470	3.2700
Ref. [16]	0.9620	0.9500		0.9460	0.9380	
Ref. [17]	0.9720	0.9620		0.9530	0.9510	
Ref. [18]	**0.9779**	0.9656	**2.6077**	**0.9717**	0.9529	**2.2771**
Ref. [21]	0.9705	0.9725	4.0457	0.9624	0.9557	3.4598
Ref. [22]	0.9746	0.9716	3.5526	0.9671	**0.9627**	2.9853
Proposed method	0.9773	**0.9735**	3.9677	0.9674	**0.9627**	3.0062

表 2 不同方法在 WIVC 3D 数据库上的性能比较

Table 2 Performance comparison on WIVC 3D database

Method	WIVC 3D Phase I			WIVC 3D Phase II		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Ref. [18]	0.9690	0.9599	**2.9508**	0.9580	0.9501	3.0506
Ref. [21]	0.9525	0.9502	3.3431	0.9665	0.9524	4.2889
Ref. [22]	0.9739	0.9608	3.4537	**0.9768**	**0.9765**	**2.8933**
Proposed method	**0.9742**	**0.9725**	3.3066	0.9733	0.9740	3.5029

的实验。表 3 和表 4 给出了所提方法与其他方法对单失真类型的性能比较结果。所提方法在预测特定失真类型的图像质量时表现出色，在大部分单失真类型中均取得了较优的性能，尤其是失真类型为 BLUR 和 WN 时，明显优于其他方法。

此外，所提方法评价不准确的样本如图 3 所示，左侧是左视图，右侧是右视图。该样本来自 LIVE 3D 数据库，图 3 中的标注为立体图像名称/失真类型/标签

表 3 LIVE 3D 数据库上单失真类型的 PLCC 指标对比

Table 3 PLCC comparison of individual distortion types on two LIVE 3D databases

Method	LIVE 3D Phase I					LIVE 3D Phase II				
	BLUR	FF	JP2K	JPEG	WN	BLUR	FF	JP2K	JPEG	WN
Ref. [30]	0.9170	0.7350	0.9070	0.6950	0.9170	0.9410	0.9320	0.8990	0.9010	0.9470
Ref. [12]	0.9300	0.8830	0.9260	0.7400	0.9440					
Ref. [15]	0.9740	**0.9650**	0.9880	0.9160	0.9880	0.9550	**0.9940**	0.9050	**0.9330**	0.9720
Ref. [17]	0.9880	0.9390	0.9840	0.9060	0.9470	0.9880	0.9640	0.9560	0.8250	0.9540
Ref. [21]	0.9715	0.9221	**0.9916**	**0.9406**	0.9601	0.9987	0.9609	0.9712	0.8754	0.9637
Ref. [22]	0.9928	0.9425	0.9876	0.9173	0.9724	0.9969	0.9686	**0.9810**	0.9046	0.9650
Proposed method	**0.9958**	0.9590	0.9638	0.9366	**0.9959**	**0.9990**	0.9651	0.9324	0.9053	**0.9960**

表 4 LIVE 3D 数据库上单失真类型的 SROCC 指标对比

Table 4 SROCC comparison of individual distortion types on two LIVE 3D databases

Method	LIVE 3D Phase I					LIVE 3D Phase II				
	BLUR	FF	JP2K	JPEG	WN	BLUR	FF	JP2K	JPEG	WN
Ref. [30]	0.8780	0.6520	0.8630	0.6170	0.9190	0.9000	0.9330	0.8670	0.8670	0.9500
Ref. [12]	0.9090	0.8340	0.9310	0.6930	0.9460					
Ref. [15]	0.8550	0.9170	0.9610	0.9120	0.9650	0.6000	0.9510	0.8740	0.7470	0.9420
Ref. [17]	0.9450	0.9000	0.9650	0.8790	0.9210	0.9510	**0.9690**	0.9540	0.8160	0.9230
Ref. [21]	0.9667	0.9473	**0.9780**	**0.9393**	0.9455	0.9231	0.9636	0.9525	0.8555	0.9341
Ref. [22]	0.9833	**0.9604**	0.9604	0.9036	0.9701	0.8811	0.9455	**0.9608**	0.8658	0.9341
Proposed method	**0.9903**	0.9176	0.9547	**0.9393**	**0.9966**	**0.9771**	0.9588	0.9439	**0.8674**	**0.9762**

图 3　LIVE 3D数据库中的困难样本

Fig. 3　The hard sample in LIVE 3D database

值/预测值。该数据集中较好质量和中等质量的样本较多，而失真严重样本的数量较少。由图3可知，该样本失真严重，图像质量较差，因此在测试集数据中属于离群点，网络对离群点的拟合不够准确，但所提方法对其质量的排序预测正确，具有较高的SROCC值。

3.4　跨数据库实验与分析

为了验证所提方法的泛化性能，表5和表6分别给出了其在LIVE 3D和WIVC 3D数据库上的跨数据库结果，表中用“训练集/测试集”的格式表示在本实验中使用的训练集和测试集，例如“LIVE 3D Phase I/II”表示使用LIVE 3D Phase I进行训练，使用LIVE 3D Phase II进行测试，其他同理。

表5　不同方法在LIVE 3D数据库上的跨数据库实验结果对比

Table 5　Performance comparison of cross-dataset validation on two LIVE 3D databases

Method	LIVE 3D Phase I/II		LIVE 3D Phase II/I	
	PLCC	SROCC	PLCC	SROCC
Ref. [15]	0.7100		**0.9320**	
Ref. [17]	0.8480	0.8330	0.9150	**0.9210**
Ref. [18]	**0.8595**	0.7972	0.9184	0.9149
Ref. [21]	0.8471	0.8074	0.9101	0.9061
Ref. [22]	0.8212	0.7625	0.9052	0.9133
Proposed method	0.8573	**0.8480**	0.9226	0.9209

表6　不同方法在WIVC 3D数据库上的跨数据库实验结果对比

Table 6　Performance comparison of cross-dataset validation on two WIVC 3D databases

Method	WIVC 3D Phase I / II		WIVC 3D Phase II / I	
	PLCC	SROCC	PLCC	SROCC
Ref. [18]	0.7816	0.7502	0.8981	0.8712
Ref. [21]	**0.9223**	**0.9117**	**0.9364**	**0.9490**
Ref. [22]	0.9171	0.9023	0.9309	0.9274
Proposed method	0.9150	0.9088	0.9323	0.9289

可以发现，相较于训练集和测试集来自同一数据库的情况，所有方法跨数据库的实验性能均出现了退化，这是因为不同数据库的数据分布存在较大差别。所提方法在“LIVE 3D Phase I/II”上取得了最高的SROCC。而文献[21]方法因更多的参数量和较高的复杂度而在WIVC 3D的跨库实验中取得了较所提方法更好的性能，复杂度情况见第3.6节。因此，综合性能和复杂度，所提方法具有一定的竞争力。

3.5　消融实验与分析

为了验证所提方法各部分的有效性，在LIVE 3D数据库上进行了一系列消融实验，实验结果见表7～9，其中，每个消融实验的最优指标均用加粗表示，Param. 表示参数量。

3.5.1　立体注意力模块的分析。

为验证所提SAT模块的有效性，设置了3类对照组，实验结果见表7。3类对照实验的说明如下：1）w/o SAT表示在网络中移除所有SAT块，保留其余组件，以此来验证SAT模块的有效性，此时网络中没有任何注意力机制，即No attention。2）选用普通注意力模块SE[27]、CBAM[28]、GC[34]作为对比，以验证SAT的有效性。普通注意力模块均为单输入单输出结构，因此只能对单目特征或双目特征进行校正。Attention in monocular指在w/o SAT的基础上，在左右单目特征提取支路插入普通注意力模块。例如，使用SE的实验用M-SE表示，其他注意力模块同理。Attention in binocular指在w/o SAT的基础上，使用普通注意力模块对单目特征图拼接成的双目特征图进行处理。例如，使用SE的实验用B-SE表示，其他注意力模块同理。3）所提SAT既可以实现高级双目信息对低级双目信息（B2B）的自顶向下调制，也可以实现高级双目信息对低级单目信息（B2M）的自顶向下调制。为了验证这两种自顶向下调制的有效性，设置本对照组。表7中的B2B表示在SAT中保留B2B，移除B2M；B2M表示在SAT中保留B2M，移除B2B；B2B＋B2M表示所提方法。

从表7可知，移除SAT后网络性能出现了大幅度退化，验证了SAT模块的有效性。相较于移除SAT，在单目分支中插入普通注意力模块，或使用普通注意力模块处理双目特征，均能提高性能指标，且这两种方法（使用相同普通注意力模块时）带来的性能增益相近。相较于移除SAT、使用普通注意力模块处理单目

表 7 立体注意力模块消融实验结果

Table 7 Performance comparison of SAT ablation study on two LIVE 3D databases

Scheme		Param. /10^6	LIVE 3D Phase I		LIVE 3D Phase II	
			PLCC	SROCC	PLCC	SROCC
No attention	w/o SAT	7.07	0.9415	0.9404	0.9345	0.9304
Attention in monocular	M-SE	7.88	0.9534	0.9530	0.9515	0.9510
	M-CBAM	7.88	0.9579	0.9538	0.9566	0.9565
	M-GC	7.88	0.9554	0.9540	0.9522	0.9522
Attention in binocular	B-SE	7.47	0.9536	0.9520	0.9515	0.9511
	B-CBAM	7.47	0.9588	0.9540	0.9567	0.9563
	B-GC	7.47	0.9552	0.9544	0.9525	0.9511
SAT	B2B	7.47	0.9654	0.9653	0.9639	0.9600
	B2M	7.45	0.9671	0.9660	0.9658	0.9611
	B2B+B2M	7.47	**0.9773**	**0.9735**	**0.9674**	**0.9627**

特征、使用普通注意力模块处理双目特征，仅采用B2B或仅采用B2M均使性能得到了较大提升。所提方法同时使用B2B和B2M，网络的性能指标再次提升。上述实验均说明所提SAT中两种自顶向下调制的有效性，且其复杂度并未增加。

3.5.2 能量系数的分析

为了验证能量系数自适应策略的有效性，分别将α设置为固定值0.1、0.5、0.9和1，实验结果如表8所示。表中，w/o α表示在SAT的双目融合模块中不引入能量系数α（即将左右输入相加后直接进行下一步操作），等价于$\alpha=1$。由表8可知，将α设置为固定值时，模型性能会有一定程度的下降，PLCC和SROCC值均低于0.9600，而所提方法的PLCC和SROCC值均大于0.9600。因此，将α设置为可学习的参数对双目特征的幅度进行自适应调整更符合人类视觉感知机制。

表 8 能量系数消融实验结果

Table 8 Performance comparison of α ablation study on two LIVE 3D databases

Method	LIVE 3D Phase I		LIVE 3D Phase II	
	PLCC	SROCC	PLCC	SROCC
$\alpha=0.1$	0.9454	0.9475	0.9405	0.9407
$\alpha=0.5$	0.9576	0.9572	0.9546	0.9534
$\alpha=0.9$	0.9579	0.9572	0.9529	0.9513
w/o α($\alpha=1$)	0.9590	0.9573	0.9523	0.9515
Proposed method	**0.9773**	**0.9735**	**0.9674**	**0.9627**

3.5.3 双池化策略的分析

如表9所示，采用平均池化、最大池化和最小池化等3种常见的池化方式分别对融合图和差异图进行处理，且表中+和−分别表示融合图和差异图。例如，表9第一行中的+和−符号都出现在Avg一列中，表示融合特征和差异特征在质量回归前都经过了平均池化处理，以此类推。

由实验结果可知，仅使用平均池化操作会导致模型性能下降。虽然平均池化在整合上下文信息方面表现出色，但它会平滑学习到的质量特征（多为像素值过大或过小的失真特征），造成模型性能退化。大多数研究一般采用对融合图和差异图都进行最大池化处理的策略（表9第二行），其PLCC和SROCC两项指标均低于所提方法提出的双池化策略（表9最后一行）。这说明所提方法可以筛选出关键的失真信息进行质量回归，因此取得了最佳性能。

3.6 复杂度分析

表10给出了在LIVE 3D Phase I数据库上所提方法及文献[18]、[21]、[22]方法在同一环境下的模型参数量和处理一幅立体图像的平均推理时间。

从表10可知，所提方法的参数量为7.47×10^6，在LIVE 3D Phase I数据库上测试一对立体图像只需要0.0075 s，即可以在1 s内评估133对立体图像，在消费级显卡上可以满足高帧率3D电影实时处理的需求。相比之下，文献[18]方法的网络参数量为3.83×10^9，远超所提方法。由于GPU内存容量（24 GB）有限，无法训练和测试这个庞大的模型，故无法提供其推理时间。文献[21]和文献[22]方法的参数量与推理时间均比所提方法大一个数量级。综合来看，所提方法在模型性能与计算开销上均取得了较好的平衡，较上述3种方法更加高效。

表 9 不同池化策略的消融实验结果

Table 9 Performance comparison of different pooling strategies on two LIVE 3D databases

Pooling strategy			LIVE 3D Phase I		LIVE 3D Phase II	
Avg	Max	Min	PLCC	SROCC	PLCC	SROCC
＋－			0.9527	0.9503	0.9468	0.9430
	＋－		0.9676	0.9653	0.9589	0.9584
		＋－	0.9599	0.9582	0.9497	0.9486
	＋	－	0.9642	0.9614	0.9587	0.9569
	－	＋	**0.9773**	**0.9735**	**0.9674**	**0.9627**

表 10 复杂度比较结果

Table 10 Complexity comparison on the LIVE 3D Phase I database

Method	Param. /10^6	Inference time /s
Ref. [18]	3830.00	
Ref. [21]	28.57	0.0963
Ref. [22]	35.18	0.1712
Proposed method	**7.47**	**0.0075**

4 结 论

根据人类视觉自顶向下机制，提出一种基于立体注意力的无参考立体图像质量评价方法。立体注意力模块可以实现高级双目信息对低级双目信息、高级双目信息对低级单目信息的自顶向下调制，使所提方法更符合人类质量感知过程。实验结果证明，所提方法算法相比目前主流方法，具有与人类主观质量评价结果更高的一致性。

参 考 文 献

[1] Union I T. Methodology for the subjective assessment of the quality of television pictures: ITU-R BT. 500-11[S]. Geneva: ITU-R, 2002.

[2] Union I T. Subjective assessment of stereoscopic television pictures: ITU-R BT. 1438[S]. Geneva: ITU-R, 2000.

[3] Mohona S S, Au D, Kio O G, et al. Subjective assessment of stereoscopic image quality: the impact of visually lossless compression[C]//2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), May 26-28, 2020, Athlone, Ireland. New York: IEEE Press, 2020.

[4] 侯春萍, 林洪湖. 基于小波变换与结构特征的立体图像质量评价[J]. 激光与光电子学进展, 2018, 55(6): 061005.
Hou C P, Lin H H. Stereoscopic image quality assessment based on wavelet transform and structure characteristics[J]. Laser & Optoelectronics Progress, 2018, 55(6): 061005.

[5] 黄姝钰, 桑庆兵. 基于图像融合的无参考立体图像质量评价[J]. 激光与光电子学进展, 2019, 56(7): 071004.
Huang S Y, Sang Q B. No-reference stereo image quality assessment based on image fusion[J]. Laser & Optoelectronics Progress, 2019, 56(7): 071004.

[6] 叶蒙梦, 胡晋滨, 王雪津, 等. 基于双目神经元响应的无参考立体图像质量评价[J]. 激光与光电子学进展, 2021, 58(24): 2410007.
Ye M M, Hu J B, Wang X J, et al. No-reference stereoscopic image quality assessment based on binocular neuron response[J]. Laser & Optoelectronics Progress, 2021, 58(24): 2410007.

[7] 李素梅, 常永莉, 段志成. 基于卷积神经网络的立体图像舒适度客观评价[J]. 光学学报, 2018, 38(6): 0610003.
Li S M, Chang Y L, Duan Z C. Objective assessment of stereoscopic image comfort based on convolutional neural network[J]. Acta Optica Sinica, 2018, 38(6): 0610003.

[8] 常永莉, 李素梅, 胡佳洁, 等. 基于显著区域的立体图像饱和度舒适范围测定[J]. 光学学报, 2018, 38(7): 0710003.
Chang Y L, Li S M, Hu J J, et al. Measurement of comfortable range of stereo image saturation based on salient region[J]. Acta Optica Sinica, 2018, 38(7): 0710003.

[9] 胡佳洁, 李素梅, 常永莉, 等. 基于显著区域的立体图像舒适视差范围研究[J]. 光学学报, 2018, 38(8): 0811001.
Hu J J, Li S M, Chang Y L, et al. Comfortable disparity range of stereo image based on salient region[J]. Acta Optica Sinica, 2018, 38(8): 0811001.

[10] Fang Y M, Yan J B, Wang J H, et al. Learning a no-reference quality predictor of stereoscopic images by visual binocular properties[J]. IEEE Access, 2019, 7: 132649-132661.

[11] Liu Y, Yan W Q, Zheng Z, et al. Blind stereoscopic image quality assessment accounting for human monocular visual properties and binocular interactions[J]. IEEE Access, 2020, 8: 33666-33678.

[12] Zhang W, Qu C F, Ma L, et al. Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network[J]. Pattern Recognition, 2016, 59: 176-187.

[13] Yue G H, Cheng D, Li L D, et al. Semi-supervised authentically distorted image quality assessment with consistency-preserving dual-branch convolutional neural network[J]. IEEE Transactions on Multimedia, 2022, 25: 6499-6511.

[14] Shi J S, Gao P, Qin J. Transformer-based No-reference image quality assessment via supervised contrastive learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(5): 4829-4837.

[15] Zhou W, Chen Z B, Li W P. Dual-stream interactive

networks for no-reference stereoscopic image quality assessment[J]. IEEE Transactions on Image Processing, 2019, 28(8): 3946-3958.

[16] Yan J B, Fang Y M, Huang L P, et al. Blind stereoscopic image quality assessment by deep neural network of multi-level feature fusion[C]//2020 IEEE International Conference on Multimedia and Expo (ICME), July 6-10, 2020, London, UK. New York: IEEE Press, 2020.

[17] Shen L L, Chen X F, Pan Z Q, et al. No-reference stereoscopic image quality assessment based on global and local content characteristics[J]. Neurocomputing, 2021, 424: 132-142.

[18] Si J W, Huang B X, Yang H, et al. A no-reference stereoscopic image quality assessment network based on binocular interaction and fusion mechanisms[J]. IEEE Transactions on Image Processing, 2022, 31: 3066-3080.

[19] Kosslyn S M, Alpert N M, Thompson W L, et al. Visual mental imagery activates topographically organized visual cortex: PET investigations[J]. Journal of Cognitive Neuroscience, 1993, 5(3): 263-287.

[20] Bar M. A cortical mechanism for triggering top-down facilitation in visual object recognition[J]. Journal of Cognitive Neuroscience, 2003, 15(4): 600-609.

[21] Chang Y L, Li S M, Liu A Q, et al. Coarse-to-fine feedback guidance based stereo image quality assessment considering dominant eye fusion[J]. IEEE Transactions on Multimedia, 2023, 25: 8855-8867.

[22] Chang Y L, Li S M, Liu A Q, et al. Bidirectional feature aggregation network for stereo image quality assessment considering parallax attention-based binocular fusion[J]. IEEE Transactions on Broadcasting, 2024, 70(1): 278-289.

[23] Mitchell B A, Dougherty K, Westerberg J A, et al. Stimulating both eyes with matching stimuli enhances V1 responses[J]. iScience, 2022, 25(5): 104182.

[24] Zhang S H, Zhao X N, Tang S M, et al. Ocular dominance-dependent binocular combination of monocular neuronal responses in macaque V1[EB/OL]. (2023-10-27) [2024-02-04]. https://www. biorxiv. org/content/10.1101/2023.10.27.564359v1.

[25] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[26] He K M, Zhang X Y, Ren S Q, et al. Identity mappings in deep residual networks[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 630-645.

[27] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[28] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[29] Chen M J, Su C C, Kwon D K, et al. Full-reference quality assessment of stereopairs accounting for rivalry[J]. Signal Processing: Image Communication, 2013, 28(9): 1143-1155.

[30] Chen M J, Cormack L K, Bovik A C. No-reference quality assessment of natural stereopairs[J]. IEEE Transactions on Image Processing, 2013, 22(9): 3379-3391.

[31] Wang J H, Wang Z. Perceptual quality of asymmetrically distorted stereoscopic images: the role of image distortion types[EB/OL]. [2024-02-05]. https://ece. uwaterloo. ca/~z70wang/publications/vpqm14.pdf.

[32] Wang J H, Rehman A, Zeng K, et al. Quality prediction of asymmetrically distorted stereoscopic 3D images[J]. IEEE Transactions on Image Processing, 2015, 24(11): 3400-3414.

[33] Sheikh H R, Sabir M F, Bovik A C. A statistical evaluation of recent full reference image quality assessment algorithms[J]. IEEE Transactions on Image Processing, 2006, 15(11): 3440-3451.

[34] Cao Y, Xu J R, Lin S, et al. GCNet: non-local networks meet squeeze-excitation networks and beyond[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), October 27-28, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 1971-1980.