

BPCLIP: A Bottom-up Image Quality Assessment from Distortion to Semantics Based on CLIP

Chenyue Song¹, Chen Hui², Wei Zhang¹, Haiqi Zhu^{1,†}, Shaohui Liu¹, Hong Huang³, Feng Jiang¹

¹Harbin Institute of Technology, China

²Nanjing University of Information Science and Technology, China

³Sichuan University of Science & Engineering, China

{cysong, 19B908079}@stu.hit.edu.cn, chui@nuist.edu.cn,

{haiqizhu, shliu, fjiang}@hit.edu.cn, huanghong@suse.edu.cn

Abstract—Image Quality Assessment (IQA) aims to evaluate the perceptual quality of images based on human subjective perception. Existing methods generally combine multiscale features to achieve high performance, but most rely on straightforward linear fusion of these features, which may not adequately capture the impact of distortions on semantic content. To address this, we propose a bottom-up image quality assessment approach based on the Contrastive Language-Image Pre-training (CLIP, a recently proposed model that aligns images and text in a shared feature space), named BPCLIP, which progressively extracts the impact of low-level distortions on high-level semantics. Specifically, we utilize an encoder to extract multiscale features from the input image and introduce a bottom-up multiscale cross attention module designed to capture the relationships between shallow and deep features. In addition, by incorporating 40 image quality adjectives across six distinct dimensions, we enable the pre-trained CLIP text encoder to generate representations of the intrinsic quality of the image, thereby strengthening the connection between image quality perception and human language. Our method achieves superior results on most public Full-Reference (FR) and No-Reference (NR) IQA benchmarks, while demonstrating greater robustness.

Index Terms—Image Quality Assessment, Bottom-up Approach, CLIP, Multiscale Features, Cross Attention

I. INTRODUCTION

Image Quality Assessment (IQA) aims to estimate perceptual image quality similar to the human visual system (HVS) and is widely applied in various fields such as image processing, compression and storage. Depending on whether an original reference image is required, IQA methods are typically categorized into Full-Reference (FR) IQA and No-Reference (NR) IQA.

Many mathematical models have been developed to assess the appearance or quality of images. Although these tools are effective in quantifying degradations such as noise and blurriness, their integration with human language is still inadequate. Recently, CLIP-based methods have shown promise in IQA tasks by aligning language and vision [1]–[3]. However, CLIP [4] struggles with quality-aware image representations

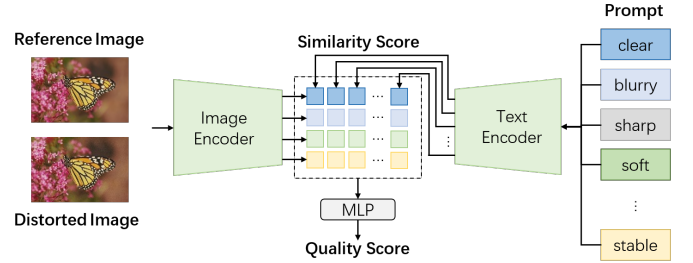


Fig. 1: Overview of the final quality score computation strategy. The quality score is computed by applying the softmax function to the similarity between features of 40 image quality adjectives and image features at different hierarchical levels.

due to its focus on high-level semantics over low-level features [3], [5]–[7]. Building on this, multiscale feature extraction emerges as a crucial strategy for enhancing performance. Traditional methods like MS-SSIM [8], BRISQUE [9], and NIQE [10] downsample images to multiple scales and compute quality scores in parallel, but they often fail to derive meaningful quality representations from low-resolution images. In contrast, top-down methods enhance feature extraction by using high-level features to guide low-level features [11] and building feature pyramids from top to bottom. Despite being more effective for multiscale extraction, top-down methods also have similar limitations: 1) they overlook how low-level distortions can disrupt high-level semantic information; 2) two images with different distortions can exhibit similar high-level semantic features, which complicates the process of using these high-level features to guide the extraction of low-level features.

To address this issue, we propose BPCLIP, a bottom-up quality perception method based on CLIP. A key innovation of our method is the multiscale cross attention (MSCA) mechanism, which effectively propagates image information from low to high levels to Extract the effects of distortion on semantic features. In addition, we developed 40 text descriptors related to image quality factors across six distinct dimensions, enabling the pre-trained CLIP text encoder to generate representations of the image’s intrinsic quality, thereby

[†] Corresponding Author.

This work is supported in part by the Central Guidance for Local Science and Technology Development Fund Projects under grant 2024ZYD0266, in part by the Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology under grant 2025r029.

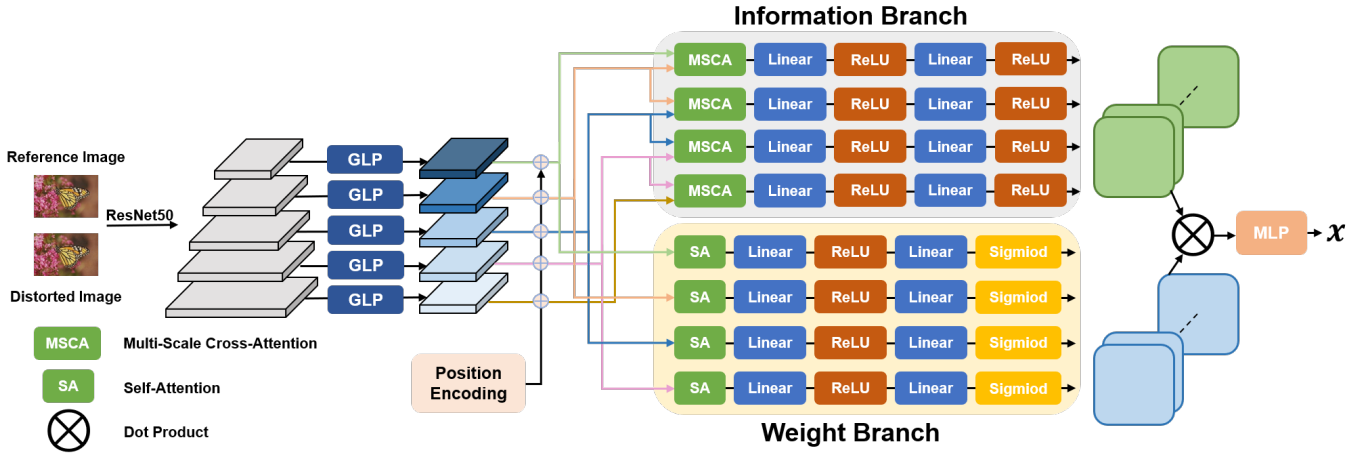


Fig. 2: Architecture overview of the proposed BPCLIP Image Encoder. Five distinct levels of features were extracted from the backbone network and progressively propagated through the information branch in a bottom-up manner.

strengthening the connection between image quality perception and human language. Extensive experiments on FR and NR IQA datasets demonstrate BPCLIP’s superior performance and lower complexity. Our contributions are summarized as follows:

- We introduce a bottom-up method that leverages multiscale features to progressively capture the impact of low-level distortions on high-level semantics. This not only enhances sensitivity to fine-grained details but also strengthens the model’s grasp of high-level semantic structures, leading to more accurate distortion detection and evaluation.
- We integrate a bottom-up image encoder with a CLIP-based text encoder to link the image quality assessment process with human language descriptions. By calculating the similarity of quality-related adjectives from multiple dimensions with image features, we improve the naturalness and explainability of the assessment.
- Our proposed BPCLIP is more efficient than the existing state-of-the-art methods. Utilizing a straightforward ResNet50 [12] backbone, it achieves competitive performance with reduced computational complexity, striking an optimal balance between performance and efficiency.

II. PROPOSED APPROACH

A. Overview

The proposed model is illustrated in Figure 1, which consists of a bottom-up image encoder and a text encoder. This network can be applied to both FR and NR tasks. In this section, we focus on the FR framework, as the NR framework is a simplified version. Using the distorted and reference images as inputs, the image encoder progressively propagates multiscale features in a bottom-up approach and ultimately projects them into the CLIP feature space. The detailed architecture is described in Section II-B. CLIP [4] is a multimodal vision-language model trained with a contrastive

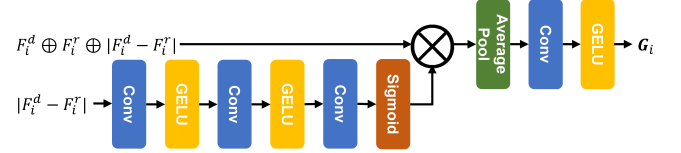


Fig. 3: GLP block comprises a mask branch and a feature branch.

loss function on a large-scale image-text dataset, designed to semantically align images and their corresponding textual descriptions in a shared embedding space. We use the CLIP text encoder to generate feature representations from 40 image quality descriptors across six distinct dimensions and align them with the output of the image encoder by calculating the cosine similarity between these text feature representations and the image features. The final quality score is derived through regression based on these similarities. During training, the pretrained text encoder stays fixed to ensure consistency in text feature representations.

B. Image Encoder: Bottom-up Network

We develop the BPCLIP image encoder network using a bottom-up approach. As shown in Figure 2, the image encoder takes the distorted and reference images as inputs, using ResNet [12] as the backbone network to extract multiscale features from both the distorted and reference images. We then apply gated local pooling (GLP) [11] to normalize the multiscale features to a consistent spatial size. The information branch uses multiscale cross attention (MSCA) blocks to process features progressively from low-level to high-level, capturing the effects of low-level distortions on high-level semantics. The weight branch incorporates self-attention (SA) blocks to direct the network’s focus towards regions with more significant semantic content. Finally, MLP maps the aggregated features to the CLIP feature space.

TABLE I: FR and NR IQA Datasets used for training and evaluation. We resize the SPAQ images so that the shorter side is 448.

Type	Dataset	Ref	Dist	Dist Type	Rating	Split	Original Size WH	Resize (shorter size)	Train Size (cropped patch)
FR	LIVE	29	779	Synthetic	25k	6:2:2	768×512(typical)	-	384×384
	CSIQ	30	866	Synthetic	5k	6:2:2	512×512	-	384×384
	TID2013	25	3k	Synthetic	5k	6:2:2	512×384	-	384×384
	KADID-10k	81	10.1k	Synthetic	30.4k	6:2:2	512×384	-	384×384
NR	CLIVE	-	1.2k	Authetic	350k	8:2	500×500	-	384×384
	KonIQ-10k	-	10k	Authetic	1.2M	8:2	512×384	-	384×384
	SPAQ	-	11k	Authetic	-	8:2	4k(typical)	448	384×384

1) *Gated Local Pooling*: The input image pair is $(I^d, I^r) \in \mathbb{R}^{3 \times H \times W}$, and the backbone features for block i are $(F_i^d, F_i^r) \in \mathbb{R}^{C_i \times H_i \times W_i}$, where H_i and W_i represent height and width, respectively, and C_i denotes the channel dimension. For ResNet50, $i \in \{1, 2, \dots, n\}$ with $n = 5$. Generally, low-level features are roughly twice the size of adjacent high-level features, so we have $H_i = H/2^i$. To enhance simplicity and efficiency, we resize F_i to match the shape of the highest-level feature F_n . For the FR task, we define gated convolution as:

$$F_i^{mask} = \sigma(\phi_i(|F_i^d - F_i^r|)) \cdot (F_i^d \oplus F_i^r \oplus |F_i^d - F_i^r|) \quad (1)$$

where σ is the sigmoid activation function that maps mask values to the range $[0, 1]$, ϕ_i denotes the bottleneck convolution block, and \oplus indicates the concatenation operation. For a detailed illustration, refer to Figure 3. For NR tasks, we apply the same gated convolution formula as described below:

$$F_i^{mask} = \sigma(\phi_i(F_i)) \cdot (W_f F_i) \quad (2)$$

the masked features F_i^{mask} undergo window average pooling and linear dimensionality reduction to produce the features $G_i \in \mathbb{R}^{D \times H_n \times W_n}$ for the subsequent block, where D represents the dimensionality of the features.

2) *Attention Modules*: We employ scaled dot-product attention [13] as the core component of our attention module. Given three feature vector sets (query, key, value), the attention mechanism first computes the similarity between the query (Q) and key (K) vectors, and then produces the weighted sum of the values (V), as shown in the following formula:

$$Attn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

we apply (3) in different ways to support the IQA task. The attention modules are divided into two branches: the information branch and the weight branch. The outputs of these branches are multiplied to produce the weighted features. In addition, in Attention Modules, positional information is crucial as an additional factor for querying features across different scales. Therefore, after GLP, we apply the same learnable positional encoding to all feature maps G_i , as illustrated in Figure 2.

a) *Information Branch(MSCA)*: After applying GLP, we obtain a set of features at different scales, denoted as $G_1, \dots, G_n \in \mathbb{R}^{(H_n \times W_n) \times D}$. In this context, the query feature Q in formula (3) inherently guides the output computation. Consequently, our multiscale cross attention mechanism is designed by generating Q , K , and V from features at various scales, as follows:

$$\begin{aligned} G'_i &= MSCA(G_i, G_{i+1}) \\ &= Attn(W_q G_i, W_k G_{i+1}, W_v G_{i+1}) + G_i \end{aligned} \quad (4)$$

where $i \in \{1, \dots, n-1\}$. The MSCA block intuitively identifies the impact of lower-level distortions on the semantic integrity of the image, by detecting the correlation between shallow-level distortions in G_i and deep-level semantics in G_{i+1} .

b) *Weight Branch(SA)*: We employ self-attention blocks to enhance G_i , and then multiply the output of the information branch processed by the MLP with the weight branch to derive the weighted features:

$$G''_i = SA(G_i) = Attn(W_q G_i, W_k G_i, W_v G_i) + G_i \quad (5)$$

$$G'''_i = MLP(G'_i) \cdot MLP(G''_i) \quad (6)$$

G'''_i suppresses irrelevant information while emphasizing critical regions that more significantly influence image quality from a human perspective. It is important to note that the last layer of the weight branch uses a Sigmoid activation function to constrain the weight values to $(0, 1)$.

3) *CLIP Feature Space Mapping*: The feature mapping formula is as follows:

$$x_i = MLP(G'''_i) \quad (7)$$

C. Score Regression

The final score is computed from the image encoder feature x and text encoder feature y :

$$s_i = \frac{x_i \odot y_i}{\|x_i\| \cdot \|y_i\|} \quad (8)$$

$$\hat{p} = MLP(s_1 \oplus s_2 \oplus \dots \oplus s_{n-1}) \quad (9)$$

where \odot denotes the dot product, $\|\cdot\|$ represents the 2-norm. The cosine similarity s_i between the i -th layer feature and the corresponding text prompt is then calculated, and \hat{p} denotes the final score.

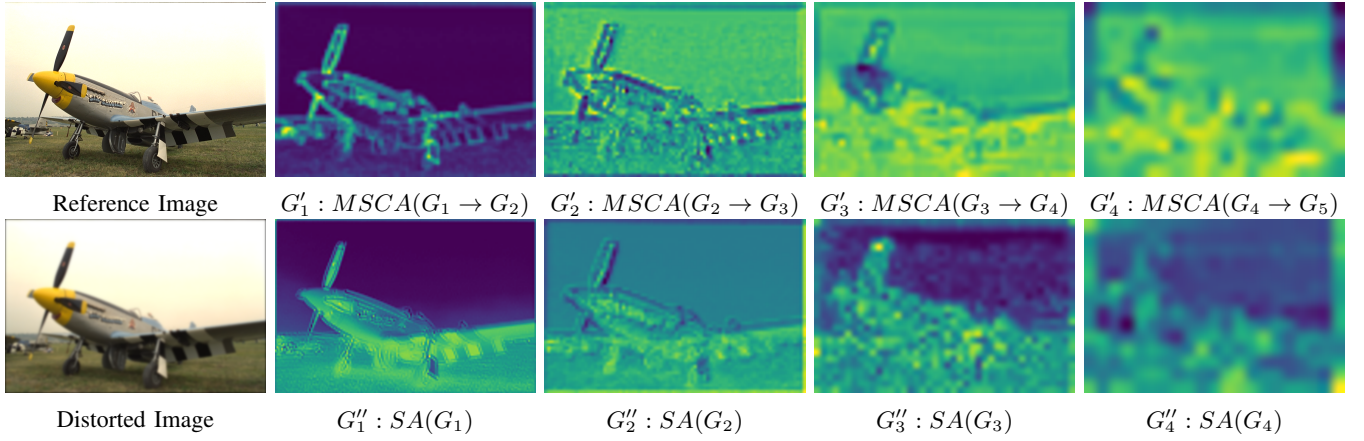


Fig. 4: Attention-based visualization of Gaussian blur distortion from the LIVE dataset. First row: the output G'_i from the information branch in Eq. 4; Second row: the output G''_i from the weight branch in Eq. 5.

TABLE II: Quantitative comparison with related works on public FR benchmarks, including the traditional LIVE, CSIQ, TID2013 and KADID-10k with MOS labels. The best and second results are colored in red and blue, while “-” indicates a non-applicable score.

Method	LIVE		CSIQ		TID2013		KADID-10k	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
PSNR	0.865	0.873	0.819	0.810	0.677	0.687	0.675	0.676
SSIM [14]	0.937	0.948	0.852	0.865	0.777	0.727	0.717	0.724
MS-SSIM [8]	0.940	0.951	0.889	0.906	0.830	0.786	0.820	0.826
VIF [15]	0.960	0.964	0.913	0.911	0.771	0.677	0.687	0.679
FSIMc [16]	0.961	0.965	0.919	0.931	0.877	0.851	0.850	0.854
MAD [17]	0.968	0.967	0.950	0.947	0.827	0.781	0.799	0.799
VSI [18]	0.948	0.952	0.928	0.942	0.900	0.897	0.687	0.679
DeepQA [19]	0.982	0.981	0.965	0.961	0.947	0.939	0.891	0.897
PieAPP [20]	0.979	0.977	0.975	0.973	0.946	0.945	0.836	0.836
LPIPS [21]	0.934	0.932	0.927	0.903	0.749	0.670	0.839	0.843
DISTS [22]	0.980	0.975	0.973	0.965	0.947	0.943	0.886	0.887
CVRKD [23]	0.965	0.960	0.965	0.958	0.935	0.928	0.959	0.957
MANIQA [24]	0.983	0.982	0.968	0.961	0.943	0.937	0.946	0.944
JSLP [25]	0.983	0.980	0.970	0.977	0.949	0.940	0.962	0.960
AHIQ [26]	0.989	0.981	0.976	0.973	0.965	0.961	-	-
TOPIQ [11]	0.982	0.981	0.979	0.976	0.956	0.953	0.960	0.956
BPCLIP	0.985	0.985	0.982	0.982	0.969	0.967	0.964	0.963
std	±0.003	±0.004	±0.003	±0.002	±0.013	±0.011	0.014	0.013

D. Loss Function

We use datasets labeled with Mean Opinion Score (MOS), and we first normalize the MOS to the range $[0, 1]$, and then apply the Mean Squared Error (MSE) loss function:

$$L(\hat{p}, p) = \frac{1}{N} \sum_{i=1}^N \|\hat{p} - p\|^2 \quad (10)$$

III. EXPERIMENTS

A. Implementation Details

1) *Datasets and Metrics*: We conduct experiments on several public benchmarks, detailed in Table I. For FR datasets, we use LIVE [27], CSIQ [17], TID2013 [28], and KADID-10k [29]. For NR datasets, we use CLIVE [30], KonIQ-10k [31], and SPAQ [32]. We resize the SPAQ images so that the shorter side is 448 as shown in Table I. We perform random splits 10 times and report the mean and standard deviation.

TABLE III: Comparison of cross-dataset performance on public benchmarks.

Train dataset	KADID-10k					
	LIVE		CSIQ		TID2013	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
PieAPP [20]	0.908	0.919	0.877	0.892	0.859	0.876
LPIPS-VGG [21]	0.934	0.932	0.896	0.876	0.749	0.670
DISTS [22]	0.954	0.954	0.928	0.929	0.855	0.830
AHIQ [26]	0.952	0.970	0.955	0.951	0.899	0.901
TOPIQ [11]	0.955	0.966	0.962	0.966	0.916	0.915
BPCLIP	0.961	0.975	0.965	0.970	0.922	0.918

For FR datasets, image segmentation is performed based on the reference image to prevent overlapping content. We evaluate performance using Spearman’s rank-order correlation coefficient (SRCC) and Pearson’s linear correlation coefficient (PLCC), which assess the monotonicity and accuracy of predictions, respectively. Higher values for SRCC and PLCC indicate better performance.

2) *Training Details*: We utilized ResNet50 pretrained on ImageNet as the backbone, freezing the batch normalization layers while fine-tuning the other parameters. Additionally, the pretrained CLIP text encoder was frozen to maintain consistency in text feature representation. Data augmentation techniques, including random cropping and horizontal/vertical flipping, were applied to preserve the intrinsic quality of the images and enhance the model’s generalization ability. The AdamW optimizer was used with a weight decay of 10^{-5} across all experiments. The initial learning rate was set to 10^{-4} for FR datasets and 3×10^{-5} for NR datasets. A cosine annealing scheduler was configured with $T_{max} = 50$, $\eta_{min} = 0$, and $\eta_{max} = lr$, where T_{max} represents the maximum number of iterations, and η_{min} and η_{max} define the minimum and maximum learning rates, respectively. Training is performed for 200 epochs, with the model implemented in PyTorch and trained on a NVIDIA GeForce RTX 3060 GPU.

TABLE IV: Comparison of NR benchmarks:CLIVE, KonIQ-10k and SPAQ.

Method	CLIVE		KonIQ-10k		SPAQ	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
DIIVINE [33]	0.591	0.588	0.558	0.546	0.600	0.599
BRISQUE [10]	0.629	0.629	0.685	0.681	0.817	0.809
NIQE [10]	0.493	0.451	0.534	0.526	0.712	0.713
ILNIQE [34]	0.508	0.508	0.537	0.523	0.721	0.713
CONTRIQUE [35]	0.422	0.394	0.630	0.637	0.680	0.676
Re-IQA [36]	0.444	0.418	0.550	0.558	0.618	0.616
ARNIQA [37]	0.558	0.484	0.760	0.741	0.797	0.789
CL-ML [38]	0.525	0.507	0.645	0.645	0.702	0.701
MUSIQ [39]	-	-	0.928	0.916	0.918	0.915
TOPIQ [11]	0.880	0.842	0.935	0.922	0.917	0.914
GRePQ [1]	0.769	0.740	0.788	0.768	0.809	0.805
CLIP-IQA [2]	0.593	0.611	0.733	0.699	0.728	0.733
LIQE [3]	0.749	0.770	0.806	0.807	0.813	0.816
BPCLIP	0.883	0.847	0.939	0.928	0.923	0.920
std	± 0.014	± 0.013	± 0.002	± 0.004	± 0.003	± 0.003

TABLE V: Results of KonIQ-10k using official split.

Method	DIIVINE [33]	BRISQUE [9]	KonCepT512 [31]	MUSIQ [39]	TOPIQ [11]	BPCLIP
PLCC	0.612	0.707	0.937	0.937	0.939	0.945
SRCC	0.589	0.705	0.912	0.924	0.926	0.932

B. Visualization of Multiscale Feature Maps

In this section, we visualize multiscale attention feature maps to demonstrate how BPCLIP operates in a bottom-up fashion. The BPCLIP image encoder consists of two branches: i) the information branch, which processes features progressively from low-level to high-level, capturing the impact of low-level distortions on higher-level semantics, and ii) the weight branch, which directs the network’s attention to regions containing more semantically significant content.

Figure 4 visualizes the multiscale attention features from $G'1$ to $G'4$ in the information branch and the region importance in multiscale features $G''1, \dots, G''4$ in the weight branch.

We observe that the MSCA block effectively captures the contours of the airplane at lower levels, while at higher levels, it extracts more abstract features. Additionally, the SA block directs the model’s attention to regions with more semantically significant content. This observation indicates that the model progressively extracts distortions in semantic regions in a coarse-to-fine manner, which aligns with human visual perception, and demonstrates that BPCLIP effectively extracts distortion features from critical semantic regions.

C. Comparison with FR Methods

To validate the superiority of the bottom-up approach, we compare the proposed BPCLIP with various traditional and deep learning methods on the FR benchmark (see Table I). Our evaluation covers within-dataset experiments, cross-dataset experiments, and computational complexity comparisons.

1) *Intra-dataset Results of Public Benchmarks:* We performed within-dataset experiments on four benchmarks: LIVE, CSIQ, TID2013 and KADID-10k. Table II presents the results showing that BPCLIP achieves significant performance on the FR IQA task, demonstrating the effectiveness of the bottom-up feature fusion approach.

TABLE VI: Ablation study for different components in BPCLIP. Experiments are done for both FR and NR datasets. IB is the Information Branch and WB is the Weight Branch.

Information Branch	Weight Branch	Multiscale Cross Attention	CLIP Text Encoder	KADID-10k(FR)		KonIQ-10k(NR)	
				PLCC	SRCC	PLCC	SRCC
IB		top-down	×	0.953	0.950	0.921	0.913
IB&WB		top-down	×	0.955	0.952	0.924	0.914
IB&WB		bottom-up	×	0.960	0.958	0.931	0.919
IB&WB		bottom-up	✓	0.964	0.963	0.939	0.928

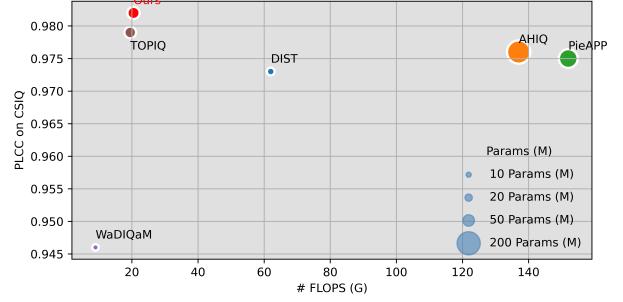


Fig. 5: Computational cost (FLOPS) vs. Performance (PLCC) on CSIQ [17]. The number of parameters is indicated by the circle radius.

2) *Cross Dataset Experiments:* Furthermore, BPCLIP demonstrates significantly improved generalization with fewer parameters, as shown by the cross dataset experiments presented in Table III. Comparing the results in Table II and Table III, we observe that BPCLIP exhibits greater robustness across various datasets.

3) *Comparison of computational complexity:* Figure 5 shows a comparison of computational complexity for recent deep learning-based FR methods. It is evident that BPCLIP achieves the best performance. Moreover, BPCLIP uses ResNet50 as the backbone, and compared to transformer-based methods like AHQ [26], its inference time is almost half. Overall, BPCLIP strikes the optimal balance between performance and computational complexity.

D. Comparison with NR Methods

Due to the absence of reference images, NR-IQA presents greater challenges than FR-IQA.

1) *Results on NR Benchmarks:* The primary NR datasets for real distortions are CLIVE, KonIQ-10k, and SPAQ. The three categories of methods in Table IV are traditional methods, learning-based methods, and CLIP-based methods, with BPCLIP outperforming the other methods, demonstrating that the proposed BPCLIP can effectively handle real distortions even without reference images. In addition, other CLIP-based methods focus solely on optimizing training strategies or loss functions, further confirming the effectiveness of bottom-up feature propagation.

2) *Results on the KonIQ-10k Official Split:* We present the results from 10 random splits of KonIQ10k in Table IV. However, [31] provides a fixed split in their official code and reports the results. We also present the results obtained using the same

settings in Table V. It is evident that BPCLIP achieves the best performance, further demonstrating the effectiveness and robustness of the proposed BPCLIP.

E. Ablation Study

Table VI shows that each key component of Bottom: 1) Information Weight Double Branch, 2) Multiscale Cross Attention, and 3) CLIP Text Encoder. We observe that all three components positively impact the results. The MSCA block, in particular, shows the most significant improvement, highlighting the effectiveness of bottom-up feature propagation.

IV. CONCLUSION

In this work, we introduce BPCLIP, a bottom-up image quality assessment method based on CLIP. We develop a bottom-up network that progressively captures the influence of distortions on semantic information. The core component of BPCLIP is MSCA, which effectively transfers low-level distortion features to higher semantic layers. In addition, by integrating 40 image quality adjectives across six dimensions, we demonstrate the feasibility of using a pre-trained CLIP text encoder to bridge human language with image quality assessment and improving the model's interpretability. Finally, we conduct extensive experiments on various public benchmarks for both full-reference (FR) and no-reference (NR) scenarios, where BPCLIP consistently achieves state-of-the-art or competitive performance.

REFERENCES

- [1] Suhas Srinath, Shankhanil Mitra, et al., "Learning generalizable perceptual representations for data-efficient no-reference image quality assessment," in *WACV*, 2024, pp. 22–31.
- [2] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy, "Exploring clip for assessing the look and feel of images," in *AAAI*, 2023, vol. 37, pp. 2555–2563.
- [3] Weixia Zhang, Guangtao Zhai, et al., "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *CVPR*, 2023, pp. 14071–14081.
- [4] Alec Radford, Jong Wook Kim, et al., "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.
- [5] Zhexin Liang, Chongyi Li, et al., "Iterative prompt learning for unsupervised backlit image enhancement," in *ICCV*, 2023, pp. 8094–8103.
- [6] Yuqi Li, Yao Lu, et al., "Sgip: A similarity guided fast layer partition pruning for compressing large deep models," *arXiv preprint arXiv:2410.14720*, 2024.
- [7] Yuqi Li, Xingyou Lin, et al., "Fedkd-hybrid: Federated hybrid knowledge distillation for lithography hotspot detection," *arXiv preprint arXiv:2501.04066*, 2025.
- [8] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *ACSSC*. Ieee, 2003, vol. 2, pp. 1398–1402.
- [9] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE TIP*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [10] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a 'completely blind' image quality analyzer," *IEEE SPL*, vol. 20, no. 3, pp. 209–212, 2012.
- [11] Chaofeng Chen, Jiadi Mo, et al., "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE TIP*, 2024.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
- [13] Cem Subakan, Mirco Ravanelli, et al., "Attention is all you need in speech separation," in *ICASSP*. IEEE, 2021, pp. 21–25.
- [14] Zhou Wang, Alan C Bovik, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.
- [15] Hamid R Sheikh and Alan C Bovik, "Image information and visual quality," *IEEE TIP*, vol. 15, no. 2, pp. 430–444, 2006.
- [16] Lin Zhang, Lei Zhang, et al., "Fsim: A feature similarity index for image quality assessment," *IEEE TIP*, vol. 20, no. 8, pp. 2378–2386, 2011.
- [17] Eric C Larson and Damon M Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *JEI*, vol. 19, no. 1, pp. 011006–011006, 2010.
- [18] Lin Zhang, Ying Shen, and Hongyu Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE TIP*, vol. 23, no. 10, pp. 4270–4281, 2014.
- [19] Jongyoo Kim and Sanghoon Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *CVPR*, 2017, pp. 1676–1684.
- [20] Ekta Prashnani, Hong Cai, et al., "Pieapp: Perceptual image-error assessment through pairwise preference," in *CVPR*, 2018, pp. 1808–1817.
- [21] Richard Zhang, Phillip Isola, et al., "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.
- [22] Keyan Ding, Kede Ma, et al., "Image quality assessment: Unifying structure and texture similarity," *IEEE T-PAMI*, vol. 44, no. 5, pp. 2567–2581, 2020.
- [23] Guanghao Yin, Wei Wang, et al., "Content-variant reference image quality assessment via knowledge distillation," in *AAAI*, 2022, vol. 36, pp. 3134–3142.
- [24] Sidi Yang, Tianhe Wu, et al., "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *CVPR*, 2022, pp. 1191–1200.
- [25] Yue Cao, Zhaolin Wan, et al., "Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment," in *CVPR*, 2022, pp. 5851–5861.
- [26] Shanshan Lao, Yuan Gong, et al., "Attentions help cnns see better: Attention-based hybrid image quality assessment network," in *CVPR*, 2022, pp. 1140–1149.
- [27] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE TIP*, vol. 15, no. 11, pp. 3440–3451, 2006.
- [28] Nikolay Ponomarenko, Oleg Ieremeiev, et al., "Color image database tid2013: Peculiarities and preliminary results," in *EUVIP*. IEEE, 2013, pp. 106–111.
- [29] Hanhe Lin, Vlad Hosu, and Dietmar Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *QoMEX*. IEEE, 2019, pp. 1–3.
- [30] Deepti Ghadiyaram and Alan C Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE TIP*, vol. 25, no. 1, pp. 372–387, 2015.
- [31] Vlad Hosu, Hanhe Lin, et al., "Konqi-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE TIP*, vol. 29, pp. 4041–4056, 2020.
- [32] Yuming Fang, Hanwei Zhu, et al., "Perceptual quality assessment of smartphone photography," in *CVPR*, 2020, pp. 3677–3686.
- [33] Anush Krishna Moorthy and Alan Conrad Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE TIP*, vol. 20, no. 12, pp. 3350–3364, 2011.
- [34] Lin Zhang, Lei Zhang, and Alan C Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE TIP*, vol. 24, no. 8, pp. 2579–2591, 2015.
- [35] Pavan C Madhusudana, Neil Birkbeck, et al., "Image quality assessment using contrastive learning," *IEEE TIP*, vol. 31, pp. 4149–4161, 2022.
- [36] Avinab Saha, Sandeep Mishra, and Alan C Bovik, "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *CVPR*, 2023, pp. 5846–5855.
- [37] Lorenzo Agnolucci, Leonardo Galteri, et al., "Arniqa: Learning distortion manifold for image quality assessment," in *WACV*, 2024, pp. 189–198.
- [38] Nithin C Babu, Vignesh Kannan, and Rajiv Soundararajan, "No reference opinion unaware quality assessment of authentically distorted images," in *WACV*, 2023, pp. 2459–2468.
- [39] Junjie Ke, Qifei Wang, et al., "Musiq: Multi-scale image quality transformer," in *ICCV*, 2021, pp. 5148–5157.