A Bioinspired Deep Learning Framework for Saliency-Based Image Quality Assessment

Huasheng Wang, Yueran Ma, Hongchen Tan, Xiaochang Liu, Ying Chen, *Senior Member, IEEE*, and Hantao Liu, *Member, IEEE*

***Abstract*—Advancements in deep learning have led to significant progress in no-reference (NR) image quality assessment (NR-IQA) for evaluating the perceived quality of digital images without relying on a reference. However, existing NR-IQA models remain suboptimal in handling complex and diverse natural images. Visual saliency constitutes a critical element for enhancing the reliability of NR-IQA, but the optimal use of saliency in deep learning-based NR-IQA has not heretofore been significantly explored. In this article, we present a novel method for integrating saliency in NR-IQA, which is motivated by the saliency-based visual search mechanism that different parts of the visual input are visited by the focus of attention (FOA) in the order of decreasing saliency. By dividing saliency into the high and low levels of FOA, we build a bioinspired deep neural network–BioSIQNet–based on a multitask learning (MTL) framework. The network architecture consists of two saliency-specific tasks and one primary image quality assessment (IQA) task. The low and high saliency (HS) are separately encoded and integrated into the early and deeper layers of the IQA network, respectively, analogous to the hierarchical processing in the visual cortex of the brain that allocates low attentional resources to process the simple patterns and high resources to learn intricate representations. We demonstrate that leveraging the synergy between visual attention and image quality perception and joint learning of these interconnected visual tasks can enhance the overall learning capabilities of the primary IQA model. Experiments validate the effectiveness of our proposed BioSIQNet for NR-IQA.**

***Index Terms*—Attention, bioinspired, deep learning, image quality assessment (IQA), saliency.**

I. Introduction

THE widespread use of multimedia technologies encompassing the Internet, social media, and smart devices has transformed our everyday life. These technologies contribute to an unprecedented surge in the creation of digital images. Inevitably, the quality of images is affected by distortions caused by acquisition, compression, transmission, and display, for example. To optimize modern imaging systems toward excellent quality of visual experience, it is critical to have highly reliable image quality assessment (IQA) models in place that can automatically evaluate image quality as perceived by human viewers [1], [2], [3].

IQA models can be classified into full-reference (FR) [4], [5], [6], reduced-reference (RR) [7], and no-reference (NR) types [8], [9], [10], [11]. FR-IQA and RR-IQA require the IQA to use a reference image for comparison. However, in real-world scenarios, access to the original, undistorted reference is often unrealistic. This has led the research attention to shift to the NR-IQA methods, which aim to automatically evaluate the quality of an image without relying on a reference image. The challenge in NR-IQA lies in developing algorithms that can faithfully emulate human perception of image quality. Recently, deep learning using convolutional neural networks (CNNs) has significantly improved the performance of NR-IQA [8], [12], [13], [14], [15]. Many deep learning techniques have been applied to learn complex representations directly from images and distortions, leading to enhanced robustness and generalization capabilities of NR-IQA. The innovative NR-IQA model can be applied in automated systems to enable real-time assessment of images captured by cameras, drones, and other devices, supporting industries like autonomous driving and robotics. Additionally, it can enhance content generation in media and entertainment by evaluating the quality of produced visuals.

Visual saliency that reflects the relative importance of different image regions has been proven to be a crucial element in shaping the perception of image quality [16], [17]. The underlying hypothesis is that certain regions of an image attract more attention from viewers than other regions; hence, these salient regions contribute more toward overall image quality. In deep learning-based NR-IQA, attempts have been made to integrate saliency information into a CNN architecture. For example, the SGDNet model [18] is trained with the addition of saliency information to predict image quality. Unfortunately, existing saliency-based NR-IQA methods exhibit two notable limitations: 1) the saliency information is often generated offline using an off-the-shelf saliency model without supervision in learning the overall model; and 2) the fusion of saliency and IQA is treated in a superficial manner without explicit biologically plausible evidence. The primate visual system employs "*serial processing based on an explicit 2-D map that encodes the saliency objects in the visual environment – competition among neurons in this map gives rise to a single winning location that corresponds*

Received 26 February 2024; revised 27 March 2025 and 8 July 2025; accepted 8 August 2025. *(Corresponding author: Yueran Ma.)*

Huasheng Wang is with the School of Computer Science and Informatics, Cardiff University, CF10 3AT Cardiff, U.K., and also with Alibaba Group, Hangzhou 311121, China.

Yueran Ma and Hantao Liu are with the School of Computer Science and Informatics, Cardiff University, CF10 3AT Cardiff, U.K. (e-mail: may68@cardiff.ac.uk).

Hongchen Tan is with the School of Future Technology, Dalian University of Technology, Dalian 116024, China.

Xiaochang Liu is with the School of Mathematics, Sun Yat-sen University, Guangzhou 510275, China.

Ying Chen is with Alibaba Group, Hangzhou 311121, China.

Digital Object Identifier 10.1109/TNNLS.2025.3598716

to the most salient object, which constitutes the next target [i.e., winner-take-all (WTA) scheme]. Inhibiting this location automatically allows the system to attend to the next most salient location [i.e., inhibition-of-return (IOR) scheme] [19]." This mechanism can be simulated by explicitly decomposing the focus of attention (FOA) into distinct high and low levels of saliency, reflecting the progression of dynamical shifts of FOA in the early visual cortical architectures. This decomposition aligns with the WTA and IOR schemes, which iteratively guide attentional transitions in the primate visual system. By structuring saliency representation in this manner, we can better capture the sequential allocation of attention for improving the performance of IQA models.

We hereby propose a computer implementation of the key organizational principles (i.e., WTA and IOR) of the attentional selection scheme based on the cortical visual hierarchy. The aim is to integrate the hierarchical saliency representation into a deep learning-based computational architecture for IQA. We address the problem of how the biologically plausible attentional selection mechanism–the FOA selects attended image locations in order of decreasing saliency–can be incorporated in IQA algorithms to enhance their performance. To this end, we build a bioinspired deep neural network–BioSIQNet–based on a multitask learning (MTL) framework, enabling the integration of hierarchical saliency information in a serial fashion to IQA. We implement an intuitively simple saliency decomposition method in which a threshold is applied to generate two new saliency maps with one representing strong intensity responses and one representing weakly activated locations, as shown in Fig. 1. The BioSIQNet learns to predict the hierarchical saliency representation (i.e., high and low levels of FOA), using discriminative saliency feature expression to facilitate the primary IQA task. The contributions of the work are as follows.

1) We propose a first-of-its-kind IQA method that emulates and incorporates the hierarchical saliency-based attentional selection mechanism. An end-to-end saliency-based NR-IQA model is built, where three tasks, including two auxiliary saliency prediction tasks and one primary IQA task, are simultaneously trained with their respective ground-truth labels. The learned visual representations across tasks are jointly optimized to improve the overall performance of the model.
2) A novel bioinspired fusion scheme is proposed to integrate the low and high saliency (HS) to the early and deeper layers of the IQA network, respectively, simulating the cortical visual hierarchy that allocates low processing resources to deal with simple patterns and high processing resources to handle abstract and complex information.
3) Extensive experiments are conducted to demonstrate the superior performance of the proposed bioinspired saliency-based IQA architecture based on deep learning. This provides insights into computational modeling of early vision mechanisms in visual tasks.

II. Related Work

A. NR Image Quality Assessment

Traditional NR-IQA methods are based on calculating handcrafted features including natural scene statistics (NSS) [20], [21], pixel-based features [22], and artifact-specific features [23], [24]. NSS-based approaches assume that different distortion types in natural images have inherent statistical properties; for example, these statistical models [24], [25] were used to extract quality-related features using locally normalized coefficients. Some NR-IQA methods extract pixel-based features directly from the pixel intensities of an image. For instance, the codebook approaches [26], [27] were utilized to derive features from local image patches. The work in [28] combined the features of the semantic obviousness of an image and its local characteristic features to boost the performance of the NR-IQA. The performance of these methods, however, varies depending on the characteristics of images, specific distortion types present in the images, and application contexts. In recent years, deep learning-based methods using CNN have been proven powerful in automatically learning relevant features for IQA. Particularly, the CNN architecture using an MTL framework can learn multiple and diverse quality-related features, leading to robust performance for NR-IQA [18], [29], [30]. The principle of MTL is that multiple related tasks are simultaneously learned to leverage shared features and representations amongst tasks to improve the overall performance. In implementing the MTL framework, IQA is regarded as the primary task, and other related tasks are considered auxiliary tasks. For example, in [31], NR-IQA is divided into two closely related subtasks with one task to classify the type of distortions and one task to predict the image quality score, and two subtasks being jointly learned. To compress the model parameters, this NR-IQA does not allow interactions between the two subtasks. In [32], an MTL-based model is proposed to increase the connection between image quality estimation and distortion identification. The IQA model in [11] adopts semantic information as the auxiliary task to enhance the primary image quality prediction task. The challenges for these methods lie in identifying perceptually relevant auxiliary tasks for IQA and obtaining reliable ground-truth labels for both the auxiliary and IQA tasks. Recently, an NR-IQA framework [11] that leverages vision-language correspondence within a multitask learning paradigm was proposed, and the method in [33] trains multimodal models to align with text-defined quality levels. However, both approaches continue to face challenges in reliably capturing the subtle, continuous spectrum of human perceptual judgments.

B. Saliency Prediction

Various computational models have been produced for saliency prediction. Earlier research focused on using visual features such as color, intensity, and orientation [34] or some heuristic saliency priors [35] to predict a saliency map. Due to the lack of higher level semantics of salient objects, these methods are rather limited in handling complex natural scenes. Recently, deep learning techniques have been applied in saliency prediction, leading to significant improvements in model performance. The ensembles of Deep Networks (DNs) [36] used shallow CNNs to detect visual saliency. After that, many deep learning-based saliency prediction models have emerged and achieved remarkable success. The model in [37] applied AlexNet [38] and VGGNet [39] on pre-trained networks to extract relevant features for saliency prediction. In [40], GoogleNet was applied for saliency feature extraction.

Fig. 1. Illustration of saliency decomposition method–a threshold is applied to a saliency map to generate two new saliency maps with one representing strong intensity responses and one representing weakly activated locations. The first column shows an image, the second column illustrates its HSM, and the third column represents the LS map.

It was found in [41] that VGGNet is more effective than AlexNet and GoogleNet for saliency prediction. Deep Visual Attention (DVA) [42] used three VGGNet-based decoders to generate multiscale feature representations for saliency detection. Besides, MSI-Net [43] employed VGGNet as the backbone in conjunction with a skip architecture to extract multiscale features, which are combined by Atrous spatial pyramid pooling. To simulate explicit properties of the human attention mechanism, a Long Short-Term Memory (LSTM) module was integrated into a saliency prediction model [44].

C. Datasets of Ground Truth: CUID-CUDAS

To achieve a highly reliable saliency-based NR-IQA, apart from the above-mentioned challenges for the design of a deep learning-based architecture inspired by the biological vision system, one of the bottlenecks is obtaining datasets of reliable ground-truth labels. The IQA literature lacks holistic datasets that represent a fully controlled psychophysical study that derives both eye-tracking data and IQA ratings for the same set of visual stimuli. A recent contribution [45], [46] created a first-of-its-kind dataset–CUID-CUDAS–that contains both image quality ratings and saliency data for a set of 600 images of varying degrees of perceived quality. Rigorous psychophysical experimentation was conducted to collect reliable human behavioral responses with eliminated subject biases. In the CUID-CUDAS dataset, each image is associated with an IQA label/score representing the image quality as perceived by an average human and a saliency label/map representing the ground-truth stimulus-driven visual attention. In this article, the CUID-CUDAS dataset is used to train the proposed NR-IQA model.

III. Methodology

A. BioSIQNet: Overall Architecture

Our goal is to predict the quality score of an input image, incorporating hierarchical saliency information to emulate key organizational principles of the attentional selection mechanism. The schematic overview of the proposed architecture of BioSIQNet is shown in Fig. 2, in which an input image is denoted as $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$ with H and W being the height and width. It should be noted that our model is designed as a *conceptual framework* to demonstrate the biologically plausible mechanism of hierarchical saliency integration; we deliberately keep the choice of encoder backbone flexible depending on specific applications. For example, popular options include VGGNet, ResNet, and Vision Transformer (ViT). In the baseline framework, we use VGGNet as the encoder backbone for extracting features from the input image (note the impact of different encoder backbones will be discussed in detail in Section IV-D). Generally, an encoder backbone network contains a series of convolutional layers, with early layers for capturing simple patterns, and deeper layers for capturing more complex representations. Let TB_{ls}, TB_{iq}, and TB_{hs} represent the module for low saliency (LS) prediction, image quality prediction, and HS prediction, respectively. For TB_{ls} and TB_{hs}, the last three fully connected layers of the VGGNet were replaced with an independent decoder, i.e., D_{ls} for TB_{ls} and D_{hs} for TB_{hs} to achieve the intended task. The biologically plausible fusion scheme is implemented by merging the feature representations of TB_{ls} and TB_{hs} into TB_{iq} to generate saliency-enhanced feature representations for the IQA task. Let F_{ls} and F_{iq_e} denote the feature maps generated by the third pooling layer in TB_{ls} and TB_{iq}, respectively. We set two parameters α and β that can be derived to assist in obtaining the fused feature map $\overline{F}$, combining F_{ls} and F_{iq_e}

$$\overline{F} = \alpha F_{ls} + \beta F_{iq\_e}. \tag{1}$$

Similarly, we fuse the deeper feature maps F_{hs} and F_{iq_d} of TB_{hs} and TB_{iq} with derivable γ and σ to generate $\hat{F}$

$$\hat{F} = \gamma F_{hs} + \sigma F_{iq\_d}. \tag{2}$$

The parameters α, β, γ, and σ are trainable scalars that dynamically control the fusion of feature maps in different stages of hierarchical saliency integration. These parameters are automatically updated via backpropagation during training to minimize the overall loss function, ensuring that the model learns an optimal combination of saliency-enhanced feature representations. The adaptive nature of these parameters enables the IQA branch to selectively absorb low-level F_{ls} and high-level F_{hs} saliency information to complement its feature representations F_{iq_e} and F_{iq_d}.

Unlike existing saliency-based IQA methods such as SGDNet [18] that generate an off-the-shelf saliency map offline by a saliency model and use it as the spatial attention mask for weighting IQA features, the proposed architecture aims to learn saliency and IQA simultaneously and directly fuse learned representations of both tasks. In our feature fusion strategy, the low saliency representation is merged into the early feature map of the IQA network, facilitating learning the basic patterns related to IQA; and the HS representation is combined with the deeper feature map of the IQA network, supporting learning complex semantics for the IQA task. By doing so, hierarchical levels of saliency information are systematically integrated into the IQA network to generate a saliency-based image quality score.

In order to adapt the pre-trained network to the saliency prediction task, similar to previous approaches [41], we remove all the fully connected layers in TB_{ls} and TB_{hs} for the feature extraction phase. All encoders of TB_{ls}, TB_{hs}, and TB_{iq} are each pre-trained on ImageNet, gaining general representations

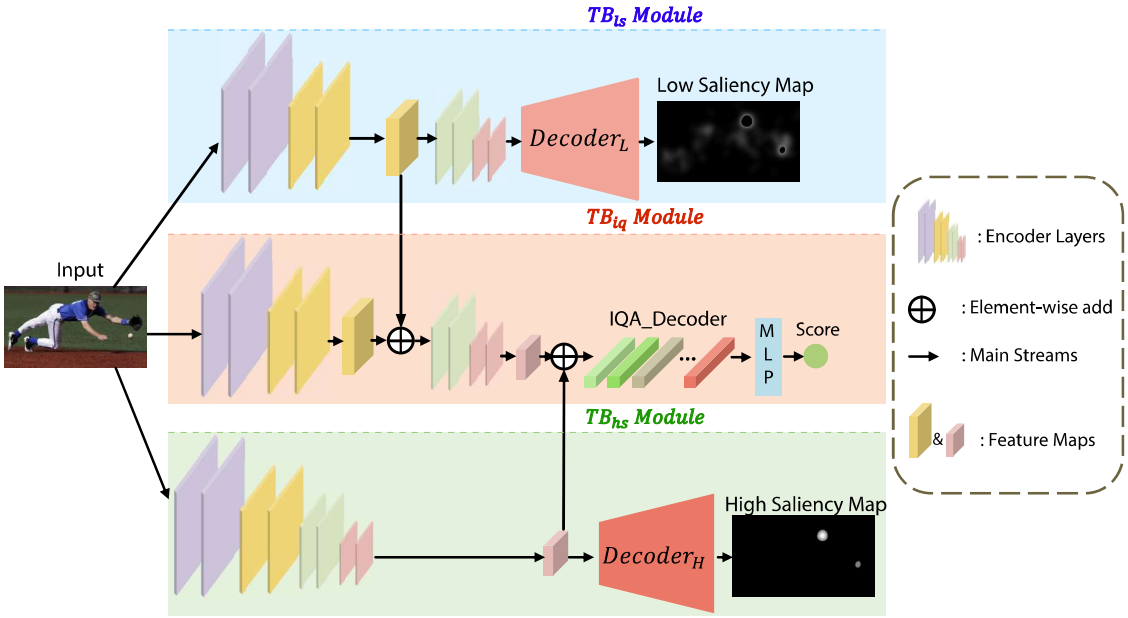

Fig. 2. Schematic overview of the proposed architecture of BioSIQNet, incorporating hierarchical saliency information to emulate key organizational principles of the attentional selection mechanism. TB_{ls}, TB_{iq}, and TB_{hs} represent the module for low saliency prediction, image quality prediction, and HS prediction, respectively. TB_{ls} and TB_{hs} focus on extracting low and HS-related features. TB_{iq} aims to generate saliency-enhanced image quality predictions by integrating features from both the low- and high-saliency pathways. BioSIQNet is trained end-to-end, where the hierarchical saliency fusion mechanism is embedded within the architecture.

of images which are transferred to learn a target task. We employ two decoders to respectively predict low and HS maps (HSMs) as the outputs of TB_{ls} and TB_{hs} networks, being $\mathcal{S}_l$ and $\mathcal{S}_h$. The detailed information on the specific decoders is described in Section III-B. In addition, these three specific tasks, being low saliency prediction, HS prediction and IQA are constrained by loss functions, which will be described in detail in Section III-C.

B. Decoding Mechanisms

To enhance the adaptability of the IQA decoder to various backbone encoder selections, we devise two distinct decoding options, i.e., transformer-tailored and CNN-tailored, as illustrated in Fig. 3. It should be noted that each pixel in the deep feature map is derived from various patches of the input image, and each patch uniquely influences the perception of overall image quality. To effectively leverage these deep feature maps, we propose an adaptive multilayer perceptron (MLP) regression module for generating IQA scores, ensuring that the extracted features from different encoder architectures (CNN-based or transformer-based) are optimally mapped to the quality prediction space, i.e., the CNN-tailored decoder leverages spatial feature hierarchies, while the transformer-tailored decoder maintains compatibility with tokenized feature representations. Given an input feature map to the MLP module $\hat{F}_{iq}$ with C channels, for the transformer-tailored decoding option, it passes through two linear projection branches. One branch computes the probability for each pixel in the feature map, while the other branch calculates the corresponding attention map. The quality score S is obtained through a weighted summation

$$S = \frac{s \odot w}{\sum w} \tag{3}$$

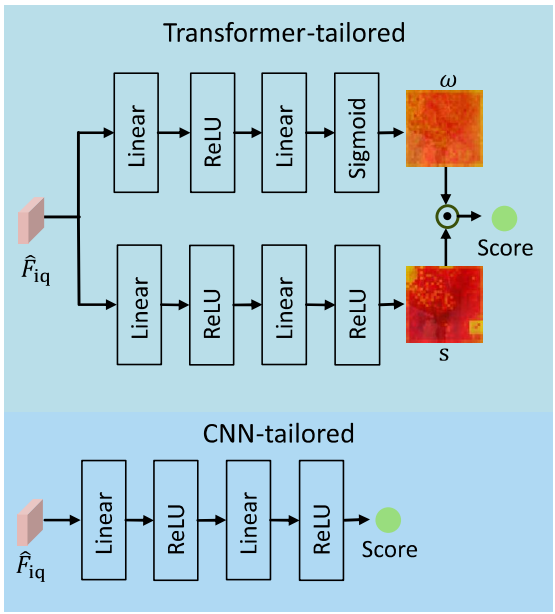

Fig. 3. Illustration of the adaptive MLP regression module for generating IQA scores.

where s of dimensions $28 \times 28 \times C$ represents the probability map, w of dimensions $28 \times 28 \times C$ represents the corresponding attention map, and $\odot$ denotes the element-wise Hadamard product that preserves the dimensionality of the channels. For the CNN-tailored decoding option, we only employ ReLU and fully connected layers, and the output parameters of the last fully connected layer are set to 1 for predicting the quality score S.

For TB_{ls} and TB_{hs}, the input image is initially downscaled by the encoder network by a factor of 32. To obtain a saliency map of the same size as the input, a 2-scale upsampling is performed five times on the feature maps in the decoder, employing nearest-neighbor interpolation. Before each upsampling layer, a convolutional layer and a ReLU activation layer are applied to adjust the channel dimensions of the feature maps.

C. Loss Functions

Due to the advances in saliency prediction methods using deep learning [46], [47], many loss functions have been investigated and proven effective. In this article, we employ a linear combination of three saliency evaluation metrics as a loss function to predict the saliency map, including Kullback–Leibler Divergence (KLD), Linear Pearson's Correlation Coefficient (CC), and Similarity (SIM). The loss function leverages their complementary properties in evaluating saliency maps. This approach aligns with prior work [46], [47] in saliency prediction, where these metrics are widely used for both evaluation and optimization. KLD measures the divergence between predicted and ground-truth saliency distributions, ensuring probabilistic alignment. CC captures linear correlation, emphasizing structural consistency. SIM quantifies overall SIM, reinforcing global alignment. By combining these metrics, the loss function provides a comprehensive supervision signal, balancing distributional accuracy, structural correlation, and perceptual SIM. Now, we first describe the loss function for the prediction of HS. We denote y_{hs} and $\hat{y_{hs}}$ as the predicted HS map and the ground truth, and i indicates the ith pixel of y_{hs} and $\hat{y_{hs}}$. The loss function is defined as

$$L_{hs}\left(y_{hs}, \hat{y_{hs}}\right) = \lambda_1 L_{\mathrm{KLD}}\left(y_{hs}, \hat{y_{hs}}\right) + \lambda_2 L_{\mathrm{CC}}\left(y_{hs}, \hat{y_{hs}}\right) + \lambda_3 L_{\mathrm{SIM}}\left(y_{hs}, \hat{y_{hs}}\right) \tag{4}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weights of individual loss functions, and

$$L_{\mathrm{KLD}}\left(y_{hs}, \hat{y_{hs}}\right) = \sum_i \hat{y_{hs_i}} \log\left(\frac{\hat{y_{hs_i}}}{\epsilon + y_{hs_i}} + \epsilon\right) \tag{5}$$

where ϵ is a regularization constant and set to 1×10^{-8}

$$L_{\mathrm{CC}}\left(y_{hs}, \hat{y_{hs}}\right) = \frac{\mathrm{cov}\left(y_{hs}, \hat{y_{hs}}\right)}{\sigma\left(y_{hs}\right)\sigma\left(\hat{y_{hs}}\right)} \tag{6}$$

where $\mathrm{cov}(\cdot)$ represents covariance and $\sigma(\cdot)$ represents standard deviation;

$$L_{\mathrm{SIM}}\left(y_{hs}, \hat{y_{hs}}\right) = \sum_i \mathrm{Min}\left(y_{hs_i}, \hat{y_{hs_i}}\right). \tag{7}$$

In implementation, y_{hs} and $\hat{y_{hs}}$ are normalized so that $\sum_i y_{hs_i} = \sum_i \hat{y_{hs_i}} = 1$. As the method detailed in [46], for the visual saliency loss, $\lambda_1, \lambda_2, \lambda_3$ are initially set to balance the sublosses as they operate on different scales. Since lower KLD (or higher CC/SIM) indicates better saliency prediction, λ_1 is set positive, while λ_2, λ_3 are negative, following prior studies [8], [46], [47]. We then refine these weights using a grid search approach, adjusting one weight while keeping the others fixed to optimize validation performance [48]. After this process, the final values are set to $\lambda_1 = 5$, $\lambda_2 = -1$, $\lambda_3 = -1$.

Similarly, we denote y_{ls} and $\hat{y_{ls}}$ as the predicted low saliency map (LSM) and the ground truth. The loss function for predicting LSM is defined as

$$L_{ls}\left(y_{ls}, \hat{y_{ls}}\right) = \lambda_1 L_{\mathrm{KLD}}\left(y_{ls}, \hat{y_{ls}}\right) + \lambda_2 L_{\mathrm{CC}}\left(y_{ls}, \hat{y_{ls}}\right) + \lambda_3 L_{\mathrm{SIM}}\left(y_{ls}, \hat{y_{ls}}\right) \tag{8}$$

where $\lambda_1, \lambda_2, \lambda_3$ are set the same values as (4).

Finally, we denote y_{iq} and $\hat{y_{iq}}$ as the predicted image quality score and the ground truth. We utilize the L_2 loss function to constrain the prediction of image quality. Consequently, the overall loss function for the model is defined as

$$L_{\mathrm{total}} = L_{hs}\left(y_{hs}, \hat{y_{hs}}\right) + L_{ls}\left(y_{ls}, \hat{y_{ls}}\right) + L_2\left(y_{iq}, \hat{y_{iq}}\right). \tag{9}$$

IV. Experiment

A. Datasets and Experimental Protocols

As already mentioned in Section II-C, the CUID-CUDAS dataset is the only holistic dataset available in the literature that provides reliable ground-truth labels of both saliency and IQA for the same set of stimuli. Therefore, we use the CUID-CUDAS dataset to derive the proposed BioSIQNet model and validate its effectiveness. First, we conduct an ablation study (*EXP1*) to verify the significance of the proposed saliency integration scheme. Second, we perform a comparative study (*EXP2*) to analyse the performance of our proposed BioSIQNet model in comparison with the state-of-the-art (SOTA) NR-IQA models. In addition, to demonstrate the generalization capability of our BioSIQNet model, we perform a series of experiments (*EXP3*) on widely recognized IQA datasets, including LIVE [49], CSIQ [50], TID2013 [51], and KADID-10K [52]. Note, these datasets only contain ground-truth IQA labels without saliency maps. To implement BioSIQNet, we use a SOTA saliency model [47] to generate saliency maps, serving as proxies for ground truth to supervise the training process.

These experiments represent two distinct validation scenarios for BioSIQNet. *EXP1* and *EXP2* use the full CUID-CUDAS dataset (IQA + eye-tracking) to assess the impact of the proposed saliency integration strategy. A CNN-based feature extractor is chosen as an interpretable "baseline" to isolate the true contribution of the proposed framework. This scenario aims to demonstrate the gain that comes from the saliency integration rather than the backbone. *EXP3* uses existing datasets (IQA only) to evaluate the adaptability and superiority of the BioSIQNet framework. Transformer-based backbone is used with the attempt to combine a powerful backbone with our saliency integration method to achieve SOTA performance on public IQA datasets. This scenario aims to demonstrate the combined gain that comes from both the saliency integration and a powerful backbone.

For the CUID-CUDAS dataset, we adopt a 9:1 train-test split, resulting in 540 images for training and 60 images for testing. This choice ensures a sufficiently large training set to facilitate robust model learning while preserving a representative test set for evaluation on small-scale IQA datasets [53], [54]. For the other four IQA datasets, we follow the widely used 8:2 train-test split, aligning with established practices in IQA research [9], [10]. This ratio balances training efficacy

with evaluation reliability, as demonstrated in benchmark IQA studies. For the CNN-tailored encoding and decoding option for BioSIQNet, the input image is randomly cropped into dimensions of 288 $\times$ 384 pixels. For the transformer-tailored encoding and decoding option for BioSIQNet, the input image is randomly cropped into dimensions of 224 $\times$ 224 pixels. The ViT-B/8 [55] serves as the pre-trained model for feature extraction, acting as the encoder. We set the patch size to 8 and the embedding dimension to 384, resulting in the channels of $\hat{F}_{iq}$ being equal to 384.

In this article, whether employing a transformer-tailored or CNN-tailored network architecture, a fixed learning rate of 5×10^{-5} is applied, and it is multiplied by 0.1 for every 10 epochs. Models undergo training with a batch size of 4 for 50 epochs, using an early stopping patience of 5 epochs. At each stop, the model that performs best is saved and used for testing. Hyperparameters $\lambda_1, \lambda_2, \lambda_3$ are set to 5, -2, -1, respectively. In addition, $\alpha, \beta, \gamma, \sigma$ represent learnable hyperparameters, initially set to a default value of 1. The network is implemented using the PyTorch framework, and training is conducted on a single RTX 3060 GPU. We mitigate overfitting through dropout (introducing randomness to prevent co-adaptation of neurons), data augmentation (expanding the dataset with augmentation techniques, such as rotations and translations), and transfer learning with pretrained models (fine-tuning a backbone model pretrained on a larger dataset), ensuring robustness without an explicit regularization term in the loss function.

For IQA model performance evaluation, we employ three widely used metrics, including Spearman's rank-order CC (SROCC), Pearson's linear CC (PLCC), and root mean squared error (RMSE). Both SROCC and PLCC range from 0 to 1, where a higher value indicates better performance. The RMSE metric ranges from 0 to positive infinity. A value of 0 indicates perfect alignment between the predicted results and the ground truth. In general, a lower RMSE value signifies better predictive performance of the model. In our study, each experiment was repeated 10 times with different random seeds to account for variability due to stochastic optimization. For each performance metric, we report the average across all runs, ensuring the results are robust and not overly influenced by any particular random initialization.

In addition, we conduct hypothesis testing to verify the statistical significance of performance differences between model variants and across different IQA models, using the statistical methodology described in [56]. The significance testing is performed on the test set (i.e., comprising 20% of the entire dataset) of each IQA dataset under study. For example, on the test set of the TID2013 dataset, each model produces 600 data points per run, representing the residuals between the ground-truth and predicted image quality scores. The comparison between two models is then based on their respective sets of residuals aggregated over all runs. When both residual samples satisfy normality assumptions, we apply either a paired t-test (for comparisons of model variants) or an independent samples t-test (for comparisons of different IQA models). In the case where normality is not satisfied, we instead employ the nonparametric Wilcoxon signed-rank test (for model variants) or the Mann–Whitney U test (for different IQA models).

TABLE I

ABLATION STUDY TO VERIFY THE IMPACT OF: 1) LS INTEGRATION; 2) HS INTEGRATION; 3) ES INTEGRATION; AND 4) HRS INTEGRATION ON THE DEEP LEARNING-BASED NR-IQA USING THE CUID-CUDAS DATASET. BASEIQNET REPRESENTS A BASELINE MODEL THAT ADOPTS THE VGGNET ONLY FOR THE IQA PREDICTION TASK. STATISTICAL SIGNIFICANCE (SIG): "*" MEANS THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT ($P < 0.05$ AT THE 95% CONFIDENCE LEVEL). "-" MEANS THE DIFFERENCE IS NOT SIGNIFICANT

Method Variant	CUID-CUDAS			Sig
	PLCC	SROCC	RMSE	
v1: BaseIQNet (baseline)	0.861	0.845	0.188	
v2: BaseIQNet + LS (low)	0.883	0.875	0.158	v2 vs v1: *
v3: BaseIQNet + HS (high)	0.892	0.886	0.154	v3 vs v2: *
v4: BaseIQNet + ES (entire)	0.912	0.898	0.169	v4 vs v3: *
v5: BioSIQNet (HRS)	**0.926**	**0.920**	**0.138**	**v5 vs v4: ***

B. EXP1: Ablation Study

To verify the effectiveness of the proposed bioinspired saliency integration scheme, we conduct an ablation study to systematically quantify the contribution of key components of our BioSIQNet model. More specifically, we investigate the impact of: 1) low saliency integration; 2) HS integration; 3) entire saliency (ES) integration; and 4) hierarchical saliency (HRS) integration on the deep learning-based NR-IQA. To this end, five IQA model variants are constructed to demonstrate the relative added value of these key components. BaseIQNet represents a baseline model that adopts the VGGNet only for the IQA prediction task (i.e., the middle primary network of Fig. 2). BaseIQNet + LS represents a model combining the VGGNet and the low saliency module (i.e., the middle primary network combined with the top auxiliary network of Fig. 2). BaseIQNet + HS represents a model combining the VGGNet and the HS module (i.e., the middle primary network combined with the bottom auxiliary network of Fig. 2). BaseIQNet + ES represents a model combining the VGGNet and a saliency integration module for predicting an entire saliency map. To implement this, we use the middle primary network combined with the bottom auxiliary network of Fig. 2, but let the auxiliary network be supervised to learn the entire saliency map (instead of learning the HS map). The results are shown in Table I, demonstrating the superiority of utilizing the bioinspired hierarchical saliency integration for NR-IQA.

To validate the proposed saliency feature fusion strategy, we conduct an ablation study, testing different placements of LSM and HSM across different VGG19 Blocks (from shallower to deeper layers). The best-performing configuration places LSM in Block 2 and HSM in Block 4, achieving the highest PLCC (0.926) and SROCC (0.920), as detailed in Table II. The results substantiate that our hierarchical saliency fusion not only aligns with human perceptual mechanisms but also empirically enhances feature representation, leading to improved correlation with subjective IQA judgments.

C. EXP2: Comparative Study on CUID-CUDAS Dataset

We implement SOTA deep learning-based NR-IQA models and evaluate their performance on the CUID-CUDAS dataset under the same experimental conditions as described above in Section IV-A. Table III shows the performance of the

TABLE II

ABLATION STUDY TO TEST DIFFERENT PLACEMENTS OF LSM AND HSM ACROSS DIFFERENT VGG19 BLOCKS (FROM EARLY TO DEEPER LAYERS)

LSM placement	HSM placement	PLCC	SROCC
Block 1	Block 1	0.880	0.872
Block 1	Block 2	0.895	0.887
Block 1	Block 3	0.892	0.885
Block 1	Block 4	0.910	0.902
Block 2	Block 1	0.902	0.895
Block 2	Block 2	0.915	0.908
Block 2	Block 3	0.918	0.911
Block 3	Block 1	0.899	0.892
Block 3	Block 2	0.910	0.903
Block 3	Block 3	0.902	0.896
Block 3	Block 4	0.918	0.911
Block 4	Block 1	0.893	0.886
Block 4	Block 2	0.904	0.897
Block 4	Block 3	0.912	0.905
Block 4	Block 4	0.919	0.913
Block 2	**Block 4**	**0.926**	**0.920**

TABLE III

PERFORMANCE COMPARISON OF OUR BIOSIQNET MODEL AND THE SOTA NR-IQA MODELS ON THE CUID-CUDAS DATASET, USING A 9:1 TRAIN-TEST SPLIT. STATISTICAL SIGNIFICANCE (SIG): "*" MEANS THE DIFFERENCE IN PERFORMANCE BETWEEN THE CURRENT MODEL AND BIOSIQNET IS STATISTICALLY SIGNIFICANT ($P < 0.05$ AT THE 95% CONFIDENCE LEVEL). "-" MEANS THE DIFFERENCE IS NOT SIGNIFICANT

Method	SROCC	PLCC	RMSE	Sig
DIVINE [24]	0.757	0.776	0.245	*
BRISQUE [25]	0.772	0.782	0.234	*
WaDIQaM [57]	0.846	0.859	0.201	*
TIQA [58]	0.867	0.875	0.193	*
MetaIQA [14]	0.881	0.889	0.175	*
SGDNet [18]	0.901	0.905	0.172	*
TReS [59]	0.904	0.908	0.171	*
LIQE [11]	0.905	0.910	0.171	*
Q-align [33]	0.905	0.911	0.171	*
DOR-IQA [8]	0.905	0.911	0.171	*
MANIQA [9]	0.909	0.914	0.168	*
BioSIQNet (Ours)	**0.920**	**0.926**	**0.141**	

SOTA NR-IQA models (including the saliency-based model SGDNet) and our proposed BioSIQNet. It tends to suggest the importance of saliency integration and, more critically, the necessity of adeptly designing an integration scheme tailored for the deep learning architecture. For example, the performance of the saliency-based model SGDNet is comparable to that of MANIQA without saliency information. This could be attributed to the suboptimal utilization of saliency information in SGDNet, as it is solely generated offline without any supervision integrated throughout the entire network. Our BioSIQNet outperforms both MANIQA and SGDNet, implying that superior integration of saliency within a deep learning architecture yields notable improvements for the IQA prediction task.

In our proposed BioSIQNet model, the two auxiliary saliency networks TB_{ls} and TB_{hs} generate respective outputs. While these outputs are low and HSMs are not directly used, they provide insights into how networks are learning to predict the hierarchical saliency. To visualize the effectiveness of these

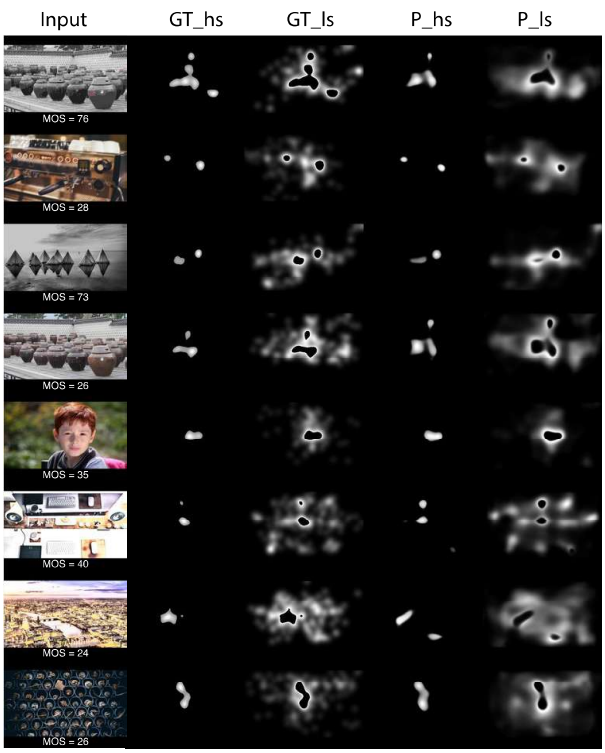

Fig. 4. Illustration of visual comparison between the ground truth of low and HSMs and the corresponding maps generated by TB_{ls} and TB_{hs} of the BioSIQNet model. The first column displays the input image, the second and third columns show the ground truth of high and LSMs (GT_hs and GT_ls), the fourth and fifth columns represent the predicted high and LSMs (P_hs and P_ls).

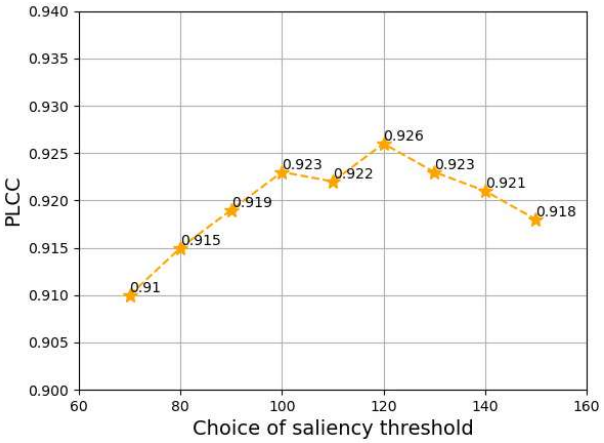

Fig. 5. Illustration of the impact of the choice of saliency threshold (separating the low and HS) on the performance (i.e., measured by PLCC) of the BioSIQNet.

saliency networks, we show some representative outputs in Fig. 4. The figure illustrates the visual comparison between the ground truth of low and HSMs and the corresponding maps generated by TB_{ls} and TB_{hs} of the BioSIQNet model. It can be seen that our model can effectively learn the hierarchical saliency information.

We also investigate the impact of the choice of saliency threshold (separating the low and HS) on the performance

TABLE IV

PERFORMANCE COMPARISON OF OUR BIOSIQNET MODEL AND THE SOTA NR-IQA MODELS ON PUBLIC IQA DATASETS INCLUDING LIVE, CSIQ, TID2013, AND KADID-10K, USING A 8:2 TRAIN-TEST SPLIT

Method	LIVE			CSIQ			TID2013			KADID-10k		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
DIVINE [24]	0.908	0.892	0.168	0.776	0.804	0.234	0.567	0.643	0.279	0.435	0.413	0.284
BRISQUE [25]	0.944	0.929	0.143	0.748	0.812	0.235	0.571	0.626	0.278	0.567	0.528	0.279
ILNIQE [60]	0.906	0.902	0.169	0.865	0.822	0.191	0.648	0.521	0.262	0.558	0.528	0.280
BIECON [61]	0.961	0.958	0.129	0.823	0.815	0.209	0.762	0.717	0.243	0.648	0.623	0.262
MEON [32]	0.955	0.951	0.134	0.864	0.852	0.191	0.824	0.808	0.215	0.691	0.604	0.254
WaDIQaM [57]	0.955	0.960	0.134	0.844	0.852	0.200	0.855	0.835	0.189	0.752	0.739	0.246
DBCNN [62]	0.971	0.968	0.120	0.959	0.946	0.131	0.865	0.816	0.181	0.856	0.851	0.188
TIQA [58]	0.965	0.949	0.125	0.838	0.825	0.203	0.858	0.846	0.185	0.855	0.850	0.189
MetalQA [14]	0.959	0.960	0.130	0.908	0.899	0.168	0.868	0.856	0.172	0.775	0.762	0.236
P2P-BM [63]	0.958	0.959	0.132	0.902	0.899	0.171	0.856	0.862	0.186	0.849	0.840	0.208
SGDNet [18]	0.965	0.969	0.125	0.903	0.883	0.171	0.861	0.843	0.184	-	-	-
HyperIQA [64]	0.966	0.962	0.124	0.942	0.923	0.145	0.858	0.840	0.188	0.845	0.852	0.219
TReS [59]	0.968	0.969	0.121	0.942	0.922	0.145	0.883	0.863	0.167	0.858	0.915	0.185
LIQE [11]	0.972	0.970	0.118	0.936	0.938	0.140	0.883	0.863	0.167	0.863	0.860	0.181
Q-align [33]	0.975	0.977	0.110	0.961	0.944	0.129	0.893	0.891	0.158	0.876	0.874	0.177
DOR-IQA [8]	0.978	0.977	0.110	0.961	0.945	0.129	0.901	0.887	0.154	0.885	0.883	0.166
MANIQA [9]	0.983	0.982	0.103	0.968	0.961	0.110	0.943	0.937	0.125	0.943	0.937	0.125
BioSIQNet (Ours)	**0.985**	**0.983**	**0.101**	**0.974**	**0.969**	**0.103**	**0.956**	**0.949**	**0.112**	**0.948**	**0.943**	**0.118**

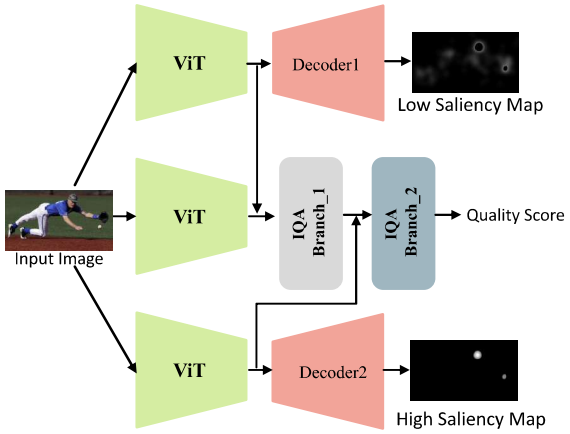

Fig. 6. Illustration of ViT-based network architecture of BioSIQNet. The encoder-decoder mechanism as adopted in MANIQA [9] is employed to predict low and HS as well as IQA.

of the BioSIQNet. To this end, we calculate the model's performance (i.e., measured by PLCC) on varying threshold selections. More specifically, we choose saliency intensity values within the range of 70 to 150 with an increment of 10. This range of intensity values is determined through empirical experimentation to achieve a balance between low and HS levels. As shown in Fig. 5, the performance of the BioSIQNet exhibits variation with the different selections of saliency threshold, reaching its peak at the saliency intensity of 120. Therefore, the threshold is set to 120 in our study.

D. EXP3: Comparative Study on Public IQA Datasets

It is customary to perform a comparative analysis of IQA models using widely recognized IQA datasets, including LIVE [49], CSIQ [50], TID2013 [51], and KADID-10k [52]. It should be noted that these public datasets include no ground-truth saliency maps of stimuli as required for training our BioSIQNet. Before being able to implement the BioSIQNet model, one practical solution is to generate saliency maps using an SOTA saliency model, i.e., [47], and use them as proxies for ground truth. To compensate for the potential deficiencies of the "generated" ground-truth saliency, we optimize the network architecture of BioSIQNet by leveraging a more powerful ViT-based encoder–decoder mechanism as adopted in MANIQA [9] to predict low and HS as well as IQA. Note that both methods use Swin Transformer as the encoder backbone; the key difference between the transformer-based BioSIQNet and the MANIQA lies in the addition of saliency and the use of a loss function to achieve optimized fusion of IQA and saliency features. As depicted in Fig. 6, we simply replace the VGGNet-based encoding and decoding structure with the ViT-based structure that consists of three parallel networks. Decoder_1 and Decoder_2 are utilized to progressively decode the low saliency enhanced early IQA features and HS enhanced deeper IQA features for the IQA prediction task. As shown in Table IV, the proposed BioSIQNet outperforms SOTA NR-IQA models on all IQA datasets, demonstrating the importance of modeling hierarchical saliency in enhancing IQA prediction. The performance of all models was evaluated by re-running their publicly available code in our experimental environment to ensure fair comparisons under identical settings. Note, differences in implementation details such as data preprocessing, training procedures, or evaluation methodologies may lead to variations in reported performance across different studies. The results of statistical significance testing are shown in Table V, indicating that our proposed model is statistically significantly ($P < 0.05$ at the 95% confidence level) better than any of the other five top-performing

TABLE V

RESULTS OF STATISTICAL SIGNIFICANCE TESTING FOR THE PERFORMANCE COMPARISON OF IQA MODELS. T-L-Q-D-M DENOTES THE TOP-PERFORMING IQA MODELS INCLUDING TRES, LIQE, Q-ALIGN, DOR-IQA, AND MANIQA. "*" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT ($P < 0.05$ AT THE 95% CONFIDENCE LEVEL). "-" MEANS THAT THE DIFFERENCE IS NOT SIGNIFICANT

	LIVE	CSIQ	TID2013	KADID-10k
	T-L-Q-D-M	T-L-Q-D-M	T-L-Q-D-M	T-L-Q-D-M
BioSIQNet	*-*-*-*-*	*-*-*-*-*	*-*-*-*-*	*-*-*-*-*

TABLE VI

CROSS-DATABASE EVALUATION PERFORMANCE COMPARISON

Train on	KADID-10k					
Test on	LIVE		CSIQ		TID2013	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
DOR-IQA [8]	0.853	0.838	0.826	0.820	0.723	0.709
MANIQA [9]	0.864	0.849	0.837	0.822	0.745	0.726
BioSIQNet	**0.883**	**0.872**	**0.841**	**0.833**	**0.757**	**0.749**

TABLE VII

COMPARISON OF ACCURACY AND EFFICIENCY OF CNN-BASED AND TRANSFORMER-BASED BIOSIQNET ON CUID-CUDAS DATASET

Configuration	PLCC	SROCC	GPU memory	Model size
VGG19-based	0.926	0.920	7GB	548MB
ViT-based	**0.932**	**0.925**	12GB	832MB

NR-IQA, TReS [59], LIQE [11], Q-align [65], DOR-IQA [8], and MANIQA [9] in predicting perceived image quality.

To critically evaluate the generalization capability of the IQA models, we perform a cross-dataset evaluation. In this experiment, we only compare our proposed BioSIQNet with the two best-performing SOTA NR-IQA, i.e., DOR-IQA [8] and MANIQA [9]. Each model, including our BioSIQNet, is trained on the KADID-10k dataset (without any additional pretraining on other datasets) and tested on the LIVE and TID2013 datasets, respectively. The results are illustrated in Table VI, showing the superior generalization ability of the proposed BioSIQNet.

E. BioSIQNet Model Configurations and Feature Visualization

In the proposed framework of BioSIQNet, we include two distinct network encoder-decoder configurations, i.e., CNN-based and transformer-based. The choice of network configuration is guided by the complexity of the IQA task. The CNN-based configuration is efficient with spatially localized features with lower computational overhead, but struggles with capturing global relationships. In contrast, the transformer-based configuration captures long-range dependencies but requires more data for generalization, leading to higher computational costs. The transformer-based BioSIQNet, including a transformer encoder (e.g., ViT, Swin), suits complex distortions, while the CNN-based BioSIQNet, including a CNN encoder (e.g., VGG, ResNet), is preferable for real-time applications with limited computational resources.

To fairly compare the predictive power and computational efficiency of CNN-based versus transformer-based BioSIQNet, we evaluate both models on the CUID-CUDAS dataset and analyse the GPU memory consumption and model size under the same training conditions. As shown in Table VII, both models produce high prediction accuracy (i.e., both outperform the SOTA IQA model, MANIQA), with ViT achieving higher performance. However, VGG19 requires only 7GB of GPU memory, whereas ViT consumes 12GB, making VGG19 a more memory-efficient choice for resource-limited environments. Additionally, VGG19's model size is 548MB, whereas ViT's model size is 832MB, further demonstrating the trade-off between computational efficiency and representational power.

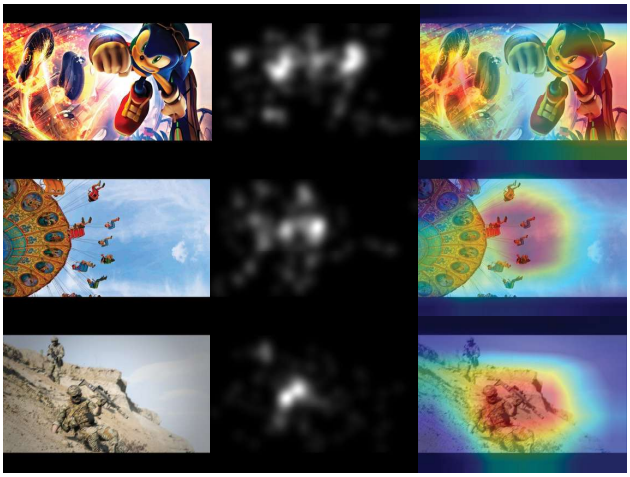

Fig. 7. Illustration of input images, ground-truth saliency maps, and Grad-CAM [66] heatmaps on sample images from the CUID-CUDAS dataset.

TABLE VIII

COMPARISON OF TRAINING EFFICIENCY AND MODEL SIZE ON THE KADID-10K DATASET

Model	Time / epoch	Param count
DOR-IQA	31 mins	107.32M
MANIQA	34 mins	135.62M
BioSIQNet (ours)	38 mins	154.65M

We further analyse our model's interpretability with Grad-CAM [66], a widely adopted post-hoc method that generates visual explanations for CNNs. Grad-CAM highlights the spatial regions most influential in the model's decision-making process. As shown in Fig. 7, our model effectively identifies regions in alignment with human visual attention, as demonstrated by the correspondence between Grad-CAM heatmaps and saliency maps.

F. Computational Complexity Analysis

To further evaluate the practicality of our approach, we conduct an analysis of computational complexity in terms of time consumption and parameter count. We compare the performance of our proposed BioSIQNet with two SOTA models, DOR-IQA and MANIQA, on the KADID-10k dataset. As shown in Table VIII, BioSIQNet incurs a modest increase in training time and parameter scale relative to these SOTA models. However, this increase is justified by the more sophisticated design to better capture the perceptual characteristics of image quality. In practice, this additional complexity represents a deliberate trade-off: by enhancing the modeling of human visual attention, BioSIQNet ultimately achieves superior prediction performance.

V. CONCLUSION

In this article, we have presented a new bioinspired, saliency-based NR IQA (NR-IQA) framework–BioSIQNet. The model integrates hierarchical saliency that represents the visual search mechanism of the visual cortex of the brain into a deep learning architecture for image quality

prediction. The proposed approach leverages the low and high levels of the FOA, where low and HS representations are encoded separately and integrated progressively into the primary IQA network. The saliency and IQA tasks are jointly learned to enhance the representations for the overall task performance. By using a best-of-its-kind dataset–CUID-CUDAS–that includes reliable ground truth of both IQA and saliency, the proposed BioSIQNet model demonstrates the effectiveness of hierarchical saliency integration in NR-IQA. We have also illustrated that our BioSIQNet model is readily extended to various public IQA datasets, showing superior performance in predicting perceived image quality compared to SOTA NR-IQA models.

REFERENCES

[1] W. Liu, R. Cui, Y. Li, and S. Zhang, "Hybrid-input convolutional neural network-based underwater image quality assessment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 1790–1798, Jan. 2025.

[2] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, Jun. 2015.

[3] S. Lao, "Attentions help CNNs see better: Attention-based hybrid image quality assessment network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2022, pp. 1140–1149.

[4] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[5] S. Seo, S. Ki, and M. Kim, "A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2602–2616, Jul. 2021.

[6] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1808–1817.

[7] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3378–3389, Aug. 2012.

[8] H. Wang, Y. Tu, X. Liu, H. Tan, and H. Liu, "Deep ordinal regression framework for no-reference image quality assessment," *IEEE Signal Process. Lett.*, vol. 30, pp. 428–432, 2023.

[9] S. Yang et al., "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1191–1200.

[10] A. Saha, S. Mishra, and A. C. Bovik, "Re-IQA: Unsupervised learning for image quality assessment in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5846–5855.

[11] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14071–14081.

[12] M. Cao, Y. Fan, Y. Zhang, J. Wang, and Y. Yang, "VDTR: Video deblurring with transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 160–171, Jan. 2023.

[13] K.-Y. Lin and G. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 732–741.

[14] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 14143–14152.

[15] J. Shi, P. Gao, and J. Qin, "Transformer-based no-reference image quality assessment via supervised contrastive learning," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 4829–4837.

[16] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.

[17] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "SalientShape: Group saliency in image collections," *Vis. Comput.*, vol. 30, no. 4, pp. 443–453, Apr. 2014.

[18] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1383–1391.

[19] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, nos. 10–12, pp. 1489–1506, Jun. 2000.

[20] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a, 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2012.

[21] X. Gao, F. Gao, D. Tao, and X. Li, "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 2013–2026, Dec. 2013.

[22] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[23] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[24] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[25] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *Proc. 45th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2011, pp. 723–727.

[26] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.

[27] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.

[28] D. Li, T. Jiang, and M. Jiang, "Exploiting high-level semantics for no-reference image quality assessment of realistic blur images," in *Proc. 25th ACM Int. Conf. Multimedia*, Mountain View, CA, USA, Oct. 2017, pp. 378–386.

[29] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "GraphIQA: Learning distortion graph representations for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 25, pp. 2912–2925, 2023.

[30] S. Shi et al., "Region-adaptive deformable network for image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 324–333.

[31] S. Athar and Z. Wang, "A comprehensive performance evaluation of image quality assessment algorithms," *IEEE Access*, vol. 7, pp. 140030–140070, 2019.

[32] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[33] H. Wu et al., "Q-align: Teaching LMMs for visual scoring via discrete text-defined levels," 2023, *arXiv:2312.17090*.

[34] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[35] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, p. 5, Mar. 2009.

[36] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.

[37] M. Kümmerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge, "Understanding low- and high-level contributions to fixation prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4799–4808.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 60, 2017, pp. 84–90.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[40] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[41] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 262–270.

[42] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[43] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder–decoder network for visual saliency prediction," *Neural Netw.*, vol. 129, pp. 261–270, Sep. 2020.

[44] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3264–3274, Jul. 2018.

[45] L. Lévêque et al., "CUID: A new study of perceived image quality and its subjective assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 116–120.
[46] H. Wang, J. Lou, X. Liu, H. Tan, R. Whitaker, and H. Liu, "SSPNet: Predicting visual saliency shifts," *IEEE Trans. Multimedia*, vol. 26, pp. 4938–4949, 2024.
[47] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "TranSalNet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, vol. 494, pp. 455–467, Jul. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222004714
[48] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 281–305, 2012.
[49] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[50] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.
[51] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.
[52] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.
[53] Y. Xia, S. W. Han, and H. J. Kwon, "Image generation and recognition for railway surface defect detection," *Sensors*, vol. 23, no. 10, p. 4793, May 2023.
[54] A. Baitieva, D. Hurych, V. Besnier, and O. Bernard, "Supervised anomaly detection for complex industrial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17754–17762.
[55] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[56] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, Jun. 2016.
[57] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.
[58] J. You and J. Korhonen, "Transformer for image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1389–1393.
[59] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2022, pp. 1220–1230.
[60] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.
[61] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.
[62] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.
[63] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3575–3585.
[64] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3667–3676.
[65] H. Wu et al., "Q-align: Teaching LMMs for visual scoring via discrete text-defined levels," 2023, *arXiv:2312.17090*.
[66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Huasheng Wang received the M.S. degree from Dalian University of Technology, Dalian, China, in 2021, and the Ph.D. degree from Cardiff University, Cardiff, U.K., in 2024.

He is currently an Algorithm Engineer at Taobao, Alibaba, Hangzhou, China. His research interests include image and video quality assessment and saliency prediction.

Yueran Ma received the B.Eng. degree from Beijing Jiaotong University, Beijing, China, in 2016, and the M.S. degree from Southern Methodist University, Dallas, TX, USA, in 2018. He is currently pursuing the Ph.D. degree with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K.

His research interests include image processing, biomedical image processing, image quality assessment, and saliency prediction.

Hongchen Tan received the Ph.D. degree in computational mathematics from Dalian University of Technology, Dalian, China, in 2021.

He is currently an Associate Professor at the Institute of Future Technology, Dalian University of Technology. His research interests include computer vision.

Xiaochang Liu is currently pursuing the bachelor's degree with the School of Mathematics, Sun Yat-sen University, Guangzhou, China.

Her research interests include mathematical modeling and data analytics.

Ying Chen (Senior Member, IEEE) received the B.S. degree in applied mathematics and the M.S. degree in electrical engineering and computer science from Peking University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computing and electrical engineering from Tampere University of Technology (TUT), Tampere, Finland, in 2010.

He is currently leading the Audiovisual Technology Group in Taobao, Alibaba, supporting end-to-end multimedia features and applications within Taobao. His research interests include video coding, image/video restoration and enhancement, image/video quality assessment, and video transmission.

Hantao Liu (Member, IEEE) received the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 2011.

He is currently a Professor at the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. His research interests include intersection of image processing, machine learning, computer vision, applied perception, and medical imaging.