UNI-IQA: A Unified Approach for Mutual Promotion of Natural and Screen Content Image Quality Assessment

Mengke Song[1,2], Chenglizhao Chen[1,2†], Wenfeng Song[3], and Yuming Fang[4]
[1]Qingdao Institute of Software & School of Computer Science and Technology, China University of Petroleum (East China)
[2]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software
[3]Computer School, Beijing Information Science and Technology University
[4]School of Information Management and Mathematics, Jiangxi University of Finance and Economics

***Abstract*—To date, the image quality assessment (IQA) research field has mainly focused on natural images (NIs)-based IQA and screen content images (SCIs)-based IQA. Usually, these two research branches are quite independent due to the large differences between NIs and SCIs, where NIs, captured by cameras directly, contain pictorial information solely, yet, SCIs, synthesized or GPU-rendered, have pictures and textures. Moreover, the distortion types are also different, and subjective scores of different datasets assigned by participants are usually not well aligned. So, due to the above-mentioned "domain shifts" and "dataset misalignments", our research community has widely believed that it could be very difficult to achieve joint mutual promotions between NIs- and SCIs-based IQA. In this paper, we argue that despite the "differences", there still are some "common characteristics" — our human visual system perceives the "pictures" in both SCIs and NIs almost the same way. Thus, we can still achieve mutual performance promotion if we can appropriately use the "common characteristics" between SCIs and NIs. Our key idea is to devise a "content-aware" data switch, which, from the perspective of input's contents (*i.e.*, pictures or textures), aims at letting the model automatically enhance the commonness and compress the discrepancies between the two tasks. Notice that none of the existing fusion schemes can reach this goal since they are actually content-unaware, degenerating the "mutual interactions" into "mutual interferences". This paper is the first attempt to achieve full end-to-end "mutual interactions" between NIs- and SCIs-based IQA. Using the proposed switch, we are also the first to achieve solid mutual promotions for the two tasks, reaching new SOTA results.**

I. Introduction

Digital images might be sustained to various quality degradations during processing, compression, and transmission, which tends to result in the loss of received visual information of the images and adversely impacts many downstream applications. Thus, the objective of **i**mage **q**uality **a**ssessment (IQA) is to perceive the visual distortions and accurately predict the quality score of the images. According to the image types, the current IQA research can be divided into two categories, *i.e.*, **n**atural **i**mages (NIs)-based IQA (NI-IQA) [1]–[3] and **s**creen **c**ontent **i**mages (SCIs)-based IQA (SCI-IQA) [4], [5].

Previous research [7], [8] usually considers these two categories of IQA tasks as independent research branches due to the distinct differences between NIs and SCIs. while achieving significant improvements over previous iterations, follow a traditional paradigm where the interaction between NIs- and SCIs-based IQA tasks is minimal or non-existent. This lack of interaction stems primarily from challenges related to "domain shifts" and "dataset misalignments", which have been acknowledged as significant hurdles in achieving joint mutual promotions between these two tasks [6], [9]. "Domain shift" refers to the performance degradation of models caused by distribution discrepancies across domains. In this work, it specifically manifests as: 1) Content Differences: Natural Images (NIs) contain only natural scenes (e.g., photographs), while Screen Content Images (SCIs) mix pictorial and textual regions (e.g., text, icons). 2) Distortion Differences: NIs suffer from optical distortions (e.g., blur, noise), whereas SCIs exhibit synthetic/rendering artifacts (e.g., aliasing, color banding). 3) Scoring Differences: Human evaluators prioritize distinct criteria for NIs (e.g., naturalness) and SCIs (e.g., text legibility). These discrepancies lead to mutual interference (rather than ideal mutual promotion) when jointly training NI-IQA and SCI-IQA models.

Fig. 1. Illustration of NIs and SCIs. NIs (A) only contain pictorial information solely, while SCIs (B) have both pictorial regions and textural regions (C).

However, in this paper, we argue that despite the "differences", there still are some "common characteristics[1]" — our human visual system perceives the pictures in both SCIs and NIs almost the same way [6]. Thus, we can still achieve mutual performance promotion if we appropriately use the "common characteristics" between SCIs and NIs. However, in the deep learning era, using the widely-used "multi-task" architecture cannot achieve such mutual performance promotion [10], [11]. And inappropriate joint usage of datasets from these two tasks could degenerate the original intention — pursuing

† Corresponding author: Chenglizhao Chen, cclz123@163.com.

[1]See Sec. IV-J for further explanation.

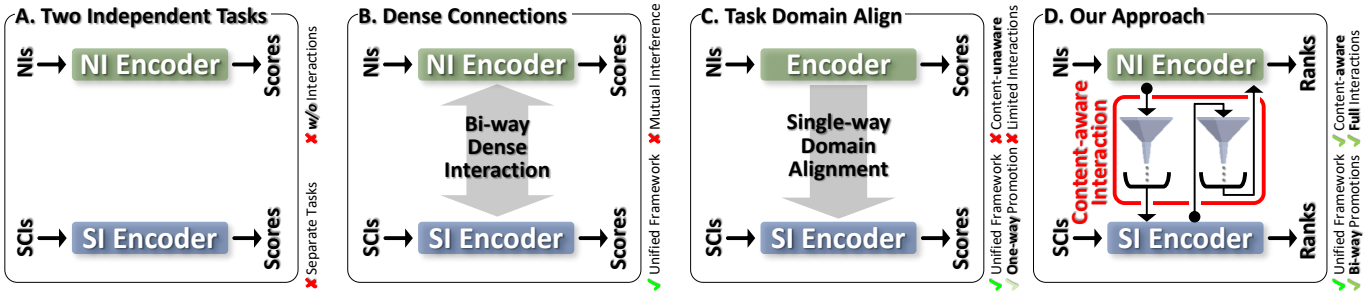

Fig. 2. Most existing methods treated NI-IQA and SCI-IQA as independent tasks (A). However, the most intuitive multi-task methodology could lead the two tasks to interfere (B). Also, the most recent domain alignment-based method [6] (C) only enables single-way interaction with limited performance gain. The key highlight of our approach (D) is the proposed content-aware interaction, which enables the two tasks to interact fully while avoiding "domain shifts" and "dataset misalignment"-induced side effects. We use ✓ and ✗ to denote advantages and disadvantages between our proposed method and the existing methods.

"mutual interactions", into the actual "mutual interferences", resulting in poor results. For a better understanding of why the conventional multi-task rationale is not suitable in IQA, from the data perspective, we shall detail some distinct differences between NIs and SCIs.

First, a significant "domain shift" exists between NIs- and SCIs-based IQA tasks. As shown in Fig. 1-A, NIs are captured by optical cameras, recording real-world scenes with rich natural textures and colors. Their content mainly includes natural elements like landscapes, people, and objects, with continuous and naturally - transitioning scenes. SCIs (Fig. 1-B), on the other hand, contain artificial elements such as text, icons, and graphics (Fig. 1-C). These elements have sharper edges and often regular geometric shapes. In fact, participants could be heavily influenced by the image contents, and the large difference between pictures and textures usually impacts the assigned overall quality scores differently. Existing studies mainly tackle this issue using domain adaptation or domain generalization techniques. For instance, Chen et al. [6] proposed an unsupervised domain adaptation (UDA) framework to transfer knowledge from NIs to SCIs, but its unidirectional transfer can't achieve bidirectional synergy. Ajakan et al. [12] introduced Domain-Adversarial Neural Networks (DANN) to align feature distributions via adversarial training, yet it ignores the content disparity in NI/SCI quality assessment. These methods work well in general domain transfer, but falter in NI/SCI hybrid scenarios as they overlook the impact of hybrid pictorial-textual content, like the dominant role of text clarity in SCIs.

Second, the widely-known datasets misalignment problem, *i.e.*, the subjective scores assigned by participants among different datasets are usually not well aligned. For example, the subjective scores in the KADID-10K set [13], a classic NI set, are ranged between 1 and 5, yet, in a typical SCI set (*e.g.*, SIQAD [14]), the scores are ranged between 0 and 100. And such score range misalignments cannot be solved via typical normalization because the normalization could still result in additional "inter-set" score misalignment [9].

Although inconsistencies in scoring ranges or precision are also prevalent within a single type of dataset, dataset misalignment remains the primary obstacle when unifying natural image (NI) and screen content image (SCI) quality assessment tasks. This is because there are significant differences in data sources, content composition, and distortion types between NIs and SCIs. These differences lead to not only difficulties in aligning the numerical ranges of subjective scores of different datasets, but also differences in the meaning of the scores and the participants' perception of image quality. In addition, current research methods are affected by this, making it difficult to achieve effective integration and mutual promotion of the two types of tasks. The commonly used "multi-task" architecture can turn the "mutual interaction" into "mutual interference" due to dataset misalignment.

Due to the above-mentioned "domain shifts" and "dataset misalignments" issues, our research community has widely believed that it could be very difficult to achieve joint mutual promotions between NIs- and SCIs-based IQA [6]. Thus, current SOTA models [15], [16] proposed in our research community follow the patterns demonstrated in Fig. 2-A, where there are no actual interactions between the two IQA tasks. However, these models lack the ability to capture "mutual interactions" between the tasks and may result in poor performance due to the discrepancies between them. Fig. 2-B has illustrated an intuitive way to achieve "mutual interactions" between the two tasks. Because of the discrepancies between the two tasks, using such plain dense connections fails to account for the distinct characteristics of each domain, leading to sub-optimal performance when attempting to leverage information across these tasks. To improve, Chen et al. [6] have used a learning-to-rank strategy. Learning-to-rank is an approach that calculates relative ranking scores within datasets to align absolute score ranges, addressing inconsistent quality scales. In the work, they adopted an unsupervised domain adaptation to boost the quality score prediction of SCIs (target domain) with the help of NIs (source domain) (see Fig. 2-C). Though this work has realized the importance of utilizing the "common characteristics" between the two tasks to improve performance, the proposed model is just a one-sided promotion (NI→SCI). The interactions between the two tasks are still absent because the proposed one-sided promotion is "content unaware" — not knowing which regions in the input image are textures and which are pictures. This content-unaware characteristic hinders full "mutual interactions" because only the picture-related information in SCIs-based IQA is helpful for the NIs-based IQA.

To improve, we propose a unified framework, UNI-IQA,

to achieve joint mutual promotions between NIs- and SCIs-based IQA (Fig. 2-D). Our key insight is to devise a "content-aware" data switch (CAS), which aims at letting the model automatically enhance the commonness and compress the discrepancies between the two tasks by distinguishing pictorial and textual regions in the input images. Specifically, CAS dynamically adjusts the learning process based on the input content (e.g., pictures or textures), thereby avoiding "mutual interference" between tasks. Notice that the proposed "content-aware" data switch can solve the domain shift problem well. Also, based on the proposed "content-aware" data switch, we further design our model to be a learning-to-rank methodology, avoiding the dataset misalignments, achieving complete "mutual interactions" between NIs- and SCIs-based IQA, and reaching significant performance promotions. To the best knowledge, our UNI-IQA is the first attempt to achieve full end-to-end content-aware "mutual interactions" between NIs- and SCIs-based IQA. In summary, the main contributions of this paper include the following:

- As the first attempt, we have presented a unified framework to integrate the NIs- and SCIs-based IQA task to achieve online end-to-end content-aware joint mutual promotions;
- From the perspective of input's contents (*i.e.*, pictures or textures), we have devised an "content-aware" data switch, which can let the model enhance the commonness and compress the discrepancies between the two tasks;
- We have conducted massive quantitative experiments to verify the effectiveness of our method in bringing extensive performance gain on both 6 NI datasets and 2 SCI datasets; both codes and results will be publicly available at https://github.com/MengkeSong/UNI-IQA.

II. Related Works

A. NR-IQA for NIs

Traditional works [17], [18] of NR-IQA proposed to model handcrafted statistics of natural images and regress parametric deviations to image degradations. However, the handcrafted feature-based approaches could perform well on type-specific distorted datasets while not well in modeling real-world distortions. Thus, learning-based methods [19]–[21] have significantly outperformed previous approaches with regard to real-world distortions by directly extracting disparate features from distorted images. But in the training process of CNN, the images have to be cropped and sampled to a fixed size in a batch, affecting the final predicted image quality scores. Several methods [22]–[24] have attempted to mitigate the distortion from resizing and cropping in CNN-based IQA. Also, some approaches [20], [25] proposed to utilize prior semantic information by using pre-trained models on classification datasets, *i.e.*, ImageNet [26], to handle distortion diversity and content changes. Sun et al. [27] planned to represent each distortion as a graph and proposed a distortion graph representation strategy to distinguish distortion types. However, this GAN-based NR-IQA method cannot handle the restoration task for the free-energy principle-guided NR-IQA methods for the severely destroyed images. To overcome this, Pan et al. [28] proposed a non-adversarial model to handle the distorted image restoration task to compensate for the shortage of the GAN-based NR-IQA methods. Recently, Venkatesh and Soundararajan [29] noted new NR-IQA progress in video scenarios. New deep-learning architectures like Transformer-based ones better capture spatio-temporal features, and temporal attention mechanisms improve dynamic quality evaluation.

B. SCI-IQA for SCIs

Unlike NIs captured by optical cameras only containing natural scenes, SCIs are more complex, consisting of pictorial and textual regions. Therefore, due to the distinct statistics of SCIs, most SCI-IQA methods [30], [31] adopted sharpness of details, screen content statistics, and spatial domain [32] for SCI quality prediction. For example, Fang et al. [33] designed a quality assessment method by combining local and global textual features and luminance features. Since the human visual system (HVS) is highly sensitive to sharp edges, Zheng et al. [34] divided an SCI into sharp and non-sharp edge regions, respectively, such that the hybrid region-based features are extracted for no-reference SCI quality assessment. According to the reality that the HVS is apt to distinguish picture and text, some other methods [35], [36] split the SCI into pictorial regions and textual regions. Li et al. [37] overviewed SCI-IQA. Latest research introduced new deep-learning-based metrics, explored GAN-attention combinations, and improved adaptability in scenarios like digital doc display and video-conferencing screen sharing. Though the above-mentioned methods have achieved steady performance improvements, the major limitation is that they can only handle either NIs or SCIs.

C. Learning to Rank

One bottleneck in IQA is the limited IQA training dataset size. Though there are enough IQA datasets, the quality scales between different IQA datasets are inconsistent, *e.g.*, 0-100 (LIVE [38]) and 0-1 (CSIQ [39]), which cannot be directly combined to as training samples. Further, simply aligning quality scales of different IQA datasets to a unified one [40] is not precise and will bring about the problem of obscure perceptual quality. Thus, Mikhailiuk et al. [41] proposed to use psychometric scaling by rating and ranking preference aggregation methodologies to mix the scores from different datasets and realign them to a common unified scale. Also, several other approaches [9], [10], [42] address NR-IQA as a learning-to-rank problem, where the relative ranking information is used during the training. Specifically, these approaches learn a ranking function from ground-truth rankings by minimizing a ranking loss [43]. This function can then be applied to rank test objects. Gao et al. [42] is the first to exploit the probability that the quality of image A is better than that of image B to generate preference image pairs (PIPs) and combine different handcrafted features to represent image pairs from the IQA dataset. Liu et al. [10] and Zhang et al. [11] inferred discrete ranking information from images of the same content and distortion but at different levels for NR-IQA model pre-training. Ma et al. [44] extracted

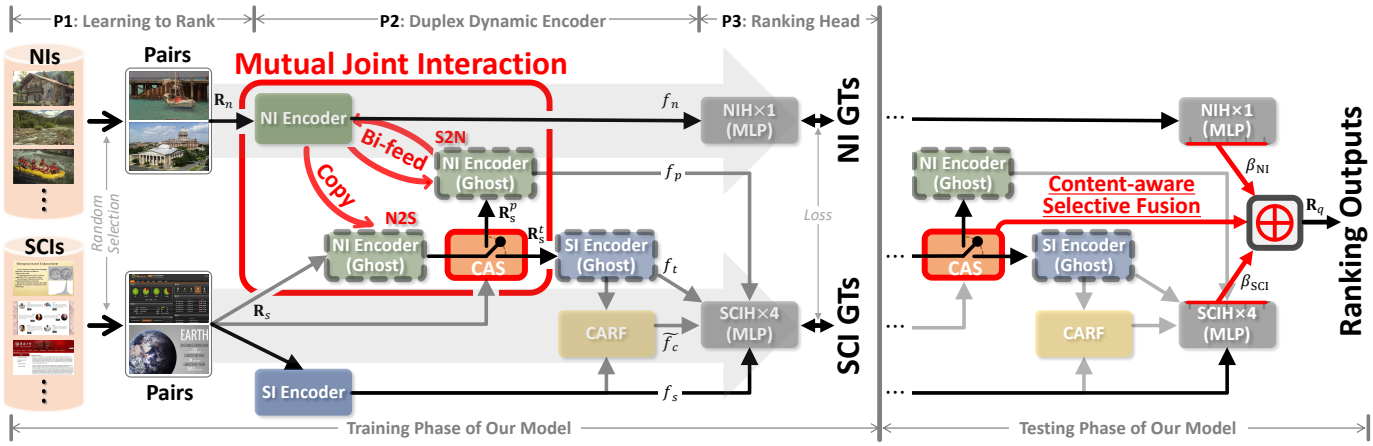

Fig. 3. The pipeline of our approach mainly consists of three major parts. In the training phase, after pre-processing the datasets by learn-to-rank strategy (**P1**), we first feed the paired images to the NI stream (the upper light grey arrow) and the SCI stream (the lower light grey arrow) of the duplex dynamic encoder (**P2**), then obtain two relative ranking scores (**P3**). Specifically, the "**c**ontent-**a**ware" data **s**witch (CAS) aims at letting the model automatically enhance the commonness and compress the discrepancies between the two tasks. The term "Bi-feed" refers to the shared architectures and parameters between NI Encoder and NI Encoder (Ghost), *i.e.*, S2N, and the mutual weight updates. In contrast, "Copy" signifies unidirectional weight transfer from NI Encoder to NI Encoder (Ghost), *i.e.*, N2S. During the testing phase, we employ content-aware selective fusion to generate the final relative ranking score. The symbol "$\oplus$" indicates the weighted addition of the two scores.

binary ranking information from FR-IQA methods to guide the optimization of NR-IQA models. Further, Zhang et al. [9] proposed UNIQUE to address the BIQA issue. Trained on synthetic-distortion databases, models struggle with realistic distortions and vice versa because of the gap between lab-simulated and real-world-captured images. UNIQUE's innovation lies in a unified BIQA model and training method for both distortion types. It samples image pairs from IQA databases, calculates quality-higher probabilities, optimizes a deep neural network with fidelity loss, and enforces a hinge constraint to regularize uncertainty estimation. To alleviate catastrophic forgetting issues, Zhang et al. [45] proposed to employ continuous learning with a learning-to-rank strategy to handle novel distortions of fast development of image processing and computer vision methods for emerging visual applications. The major problem of the above-listed methods is that they merely unified IQA methods without considering SCI-IQA.

III. Proposed Method

Our primary idea is to unify NI-IQA and SCI-IQA methods and prompt the unified framework UNI-IQA to be capable of predicting the quality scores of both NIs and SCIs simultaneously. The overall method pipeline is shown in Fig. 3. Specifically, our method consists of three major parts — **P1**: a learning-to-ranking strategy to randomly select image pairs within each dataset to avoid the dataset misalignments and achieve joint training of all NI and SCI datasets; **P2**: a duplex dynamic encoder to implement mutual joint interaction between NI stream and SCI stream by the "**c**ontent-**a**ware" data **s**witch (CAS), which is the key technical innovation; **P3**: two MLPs used as the ranking heads to predict the relative ranking scores.

A. Learning to Rank for NI-IQA and SCI-IQA

In addressing the dataset misalignment issue, this paper adopts the learning-to-rank strategy to automatically generate score scales for the two tasks to implement a unified end-to-end training protocol. Because it effectively utilizes the relative quality relationships between images rather than just absolute quality scores. This characteristic is particularly important when handling inconsistencies in score ranges across different datasets, such as natural images (NIs) and screen content images (SCIs), which often have differing score ranges and perceptual differences. Compared to traditional normalization or mapping methods [46], [47], learning-to-rank directly learns the ranking relationships between image pairs, allowing better adaptation to cross-dataset training, thus effectively addressing dataset misalignments. Additionally, the learning-to-rank method leverages participants' preference information to capture subtle perceptual differences, avoiding the issues of traditional methods that overlook fine details. Therefore, the learning-to-rank strategy is particularly suitable in the context of this task.

As shown in Fig. 3-**P1**, given an image pair $\{\mathbb{X},\mathbb{Y}\}$, we use the following equation to assign its binary label $p(\mathbb{X},\mathbb{Y})$, indicating if the quality score of $\mathbb{X} \geqslant$ the quality score of $\mathbb{Y}$ (*i.e.*, $p(\mathbb{X},\mathbb{Y})$=1), or the inverse (*i.e.*, $p(\mathbb{X},\mathbb{Y})$=0).

$$p(\mathbb{X},\mathbb{Y}) = \begin{cases} 1 & \text{if } \ \text{score}(\mathbb{X}) \geqslant \text{score}(\mathbb{Y}), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where $\text{score}(\cdot)$ denotes the subjective quality scores assigned by paticipants.

Following the previous [9], we formulate our training set D as:

$$\mathrm{D} = \left[\left[(\mathbb{X}_i^j, \mathbb{Y}_i^j), p_i^j \right]_{i=1}^{s} \right]_{j=1}^{d}, \tag{2}$$

where D contains s randomly sampled image pairs from the d-th dataset (our approach can simultaneously use *multiple* datasets from both NIs- and SCIs-based IQA tasks). Note that we don't use the Gaussian cumulative distributions since some SCI sets are exclusive of variances. This way, those different NI/SCI datasets with varying quality scores are scaled and unified. Thus, they can be trained online in an end-to-end manner.

B. Duplex Dynamic Encoder

As mentioned in Fig. 2, existing researches on IQA tend to regard NI-IQA and SCI-IQA as two individual tasks, and the widely-used architectures have completely overlooked the intercalations between them. Although some recent works (*e.g.*, [6]) have realized the importance of such interaction, there still exist several limitations hindering them from achieving good performances because their methodology is generally "content-unaware" inhibiting the full interactions between the two tasks. To improve, we propose the "duplex dynamic encoder", whose key technical components include NI/SCI streams and a **m**utual **j**oint **i**nteraction (MJI) module. The NI stream is responsible for obtaining feature representations of natural images, and the SCI stream is designed to handle screen content images. The MJI can let the "mutual interactions" become content-aware and achieve full bi-way joint performance promotion. Specifically, the MJI module contains twin ghosts and a "content-aware" data switch, which work together to enhance the commonalities and reduce the discrepancies between the two IQA tasks. Next, we shall respectively detail these technical pieces.

1) NI Stream: Our NI encoding stream, the upper light grey arrow in Fig. 3-**P2**, mainly consists of a NI Encoder aiming to obtain feature representations of NIs. Its architecture is the same as [11], which takes NI pairs as input.

2) SCI Stream: The SCI stream, shown in the lower light grey arrow in Fig. 3-**P2**, mainly includes two components, *i.e.*, 1) the SI encoder and its "Ghost", and 2) the **c**ascaded **a**daptive **r**egion **f**usion (CARF). Here we shall detail them respectively.

SI Encoder and Its Ghost. Though the SI encoder and its "Ghost" share the same architecture, *e.g.*, the tailored VGG16 [11] pre-trained on ImageNet, they are updated separately. The SI encoder embeds the "***whole***" input SCI, while the SI encoder (Ghost) is designed to obtain feature representations of the "potential texture regions" ***solely***. To achieve this, we resort to the newly devised "**c**ontent-**a**ware" **s**witch (CAS), which will be detailed later.

Cascaded Adaptive Region Fusion. Using the abovementioned encoders, we can obtain two features, *i.e.*, one from the SI encoder and another from the SI encoder (Ghost). As shown in the figure, features from the SI encoder (Ghost) are "content-aware" because of the usage of CAS, yet the features from the SI encoder, a plain architecture, are "content-unaware". To predict IQA ranks, we propose to fuse these two features using **c**ascaded **a**daptive **r**egion **f**usion (CARF, Fig. 4). Since the SI encoder (Ghost) takes the output of CAS as its input, the corresponding features tend to bias towards the "texture" regions, which cannot present the input SCI image; thus, it is unsuitable to use it solely for the quality assessment. In contrast, the features from the SI encoder are just plain ones, and the SCI stream is not likely to achieve performance gain if only such features are used. Thus, we shall first ***fuse*** the "content-unaware" and "content-aware" features. Then we further apply an additional ***residual*** operation (feed the fused features and the original versions to the "ranking head", (see f_t, $\tilde{f}_c$ and f_s in Fig. 3) to avoid information loss.

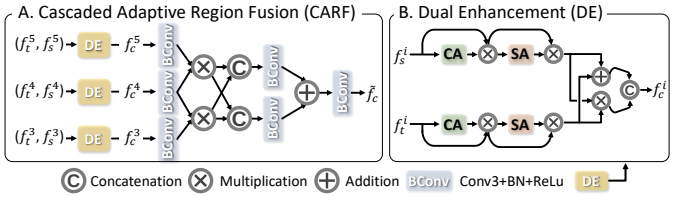

Fig. 4. Detailed architecture of CARF. CA: Channel Attention, SA: Spatial Attention. All upsampling and downsampling are omitted.

Inspired by [48], the proposed fusion strategy (Fig. 4) uses the features $(f_t^i, f_s^i),\ i \in \{3,4,5\}$ of the last three layers of the SI encoder and its Ghost are fed into **d**ual **e**nhancement module (DE) in pair. Then we can obtain the fused feature $\tilde{f}_c$ containing the enhanced picture & textual features, which will be further forwarded to the ranking head (*i.e.*, MLP in Fig. 3-**P3**). The adopted DE is used to mine useful and complementary cues from the whole SCIs and the textual regions.

Rationale of Cascaded Adaptive Region Fusion. The CARF is used to combine the feature representations of the whole image and the texture regions instead of pictural regions because the texture regions are usually more discriminative in terms of SCI-IQA than the pictural regions. In the context of SCI-IQA, the texture regions refer to the regions of an image that contain complex patterns and textures, while the pictural regions refer to the regions that contain smooth or uniform colors, such as the sky or a plain wall. Typically, the texture regions are more informative and contain more high-frequency components than the pictural regions, making them more sensitive to distortion and, thus, more critical for SCI-IQA. Therefore, the features from the SI encoder (Ghost), which are biased towards the texture regions due to the usage of CAS, are more content-aware and can better represent the distortions in the image. While the features from the SI encoder, which are content-unaware, can provide a more global view of the image and capture the overall structure and context. By fusing these two types of features using CARF, we can effectively combine both strengths and obtain a more comprehensive and accurate representation of the image quality.

3) Mutual Joint Interaction: To achieve "mutual interactions" of the NI stream and SCI stream, *i.e.*, NI→SCI and SCI→NI, we devise a **m**utual **j**oint **i**nteraction (MJI) consisting of **t**win **g**hosts (TG) and **c**ontent-**a**ware **s**witch (CAS), which work simultaneously to achieve full bi-way joint performance promotion.

Twin Ghosts. The twin ghosts architecture is composed of two NI encoder ghosts. See in the red box of Fig. 3-**P2**. The primary objective of this design is to: 1) let the two streams can interact with each other, and 2) promote the NI stream by using the *new data* transferred from the SCI stream and filtered by the proposed content-aware data switch. Actually, it's not hard to realize a dense interaction in a bi-stream structure, yet ensuring a significant promotion, especially for the two IQA tasks with distinct differences mentioned above, is really challenging. The reason is that the features transferred from one stream to another are usually only partially helpful. For example, since the SCI stream contains pictures and textures,

only the pictures' features can help promote the NI stream, yet the textures' features could do the opposite. Similarly, features from the NI stream can only benefit those picture regions in the SCI stream. So, the key to achieving the desired "mutual interaction" is to make the interaction process "content-aware" — let the interaction process be exactly clear about the exact feature types in a regional-wise way.

The Ghost design is inspired by parameter sharing and branch decoupling in lightweight networks (e.g., GhostNet [49]), aiming to generate task-specific feature representations by partially sharing parameters from the backbone encoder. In our work, the NI Encoder (Ghost) and SI Encoder (Ghost) inherit the architecture from the main encoder but employ independent parameter updates to focus on task-relevant regional features (e.g., the SI Encoder (Ghost) emphasizes textual regions in SCIs). This design balances shared learning and task-specific preservation, avoiding mutual interference caused by direct parameter sharing. Unlike classical hard parameter sharing in multi-task learning, the Ghost mechanism achieves soft sharing through the dynamic "content-aware" data switch, which activates parameter transfer only when input features align with task objectives, thereby enhancing cross-domain generalization.

The proposed "twin ghosts" serves as the basic platform for achieving this goal, which concludes: 1) the N2S ghost, which provides picture-related information from the NI stream to perceive pictorial regions of an SCI, and 2) the S2N ghost, which selectively learns the useful information embedded in SCI image first, then "bi-feed" the newly learned picture-related knowledge to both streams. Note that both N2S ghost and S2N ghost share the same architecture and parameters as the NI encoder. The major difference is that the N2S ghost adopts a one-way updating scheme, while the S2N ghost uses a bi-way updating scheme. And the "one-way" *vs.* "bi-way[2]" updating illustrations can be seen in the "red arrows" in the Mutual Joint Interaction box of Fig. 3. Thus far, both technical motivations and details of the proposed twin ghosts have been given, and next, we will detail how to achieve "content-aware mutual interaction" via the proposed twin ghosts.

Content-aware Data Switch. To make the "mutual interaction" "content-aware", our idea is to utilize the N2S ghost to provide pixel-wise spatial clues for all potential pictorial regions in the input SCI image. Since the N2S directly "copies" the learned weights of the NI encoder, which has been trained on natural pictures, it can be very sensitive to regions containing pictorial information. Thus, we can directly use the feature responses to perceive which regions in the input SCI image could be pictorial regions. This process to generate picture response map ($\tilde{\mathrm{R}}_s$) can be detailed as:

$$\tilde{\mathrm{R}}_s \;=\; \mathrm{Sigmoid}\Big(\mathrm{Conv1}\big(\mathrm{N2S}(\mathrm{R_s})\big)\Big), \tag{3}$$

where R_s denotes the input SCIs, and $\mathrm{Conv1}$ means 1×1 convolution. In a fully end-to-end fashion, $\tilde{\mathrm{R}}_s$ can explicitly reflect those pictorial regions. We have provided two qualitative examples of $\tilde{\mathrm{R}}_s$ in Fig. 5, where pictorial regions tend to have large feature responses.

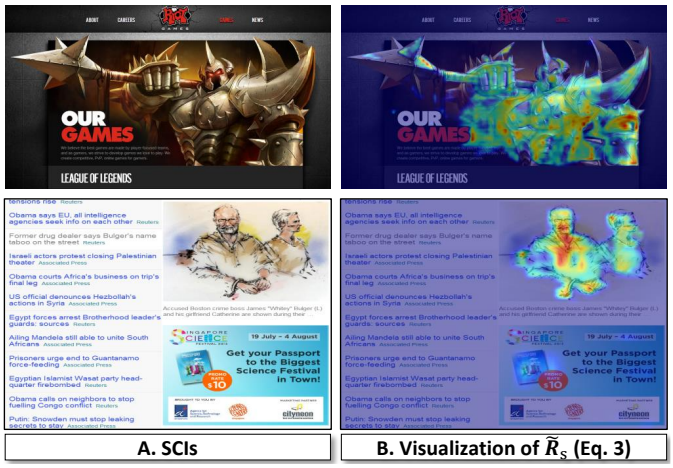

Fig. 5. Visualization of the feature response maps of pictorial regions in the SCIs generated by the proposed "content-aware" data switch (Eq. 4).

Using $\tilde{\mathrm{R}}_s$ as the "content indicator", our **c**ontent-**a**ware **s**witch (CAS) can automatically split its input R_s into two parts, *i.e.*, one part contains pictorials only (R_s^p), and another includes textures only (R_s^t). Notice that this automatical splitting process is the key to our mutual joint interaction to achieve "content-aware", whose technical details have been given as follows:

$$\begin{aligned} \mathrm{R}_s^p &= \Phi(\tilde{\mathrm{R}}_s - \sigma) \odot \mathrm{R}_s, \\ \mathrm{R}_s^t &= \Phi(1 - \tilde{\mathrm{R}}_s - \sigma) \odot \mathrm{R}_s, \end{aligned} \tag{4}$$

where $\odot$ denotes element-wise multiplicative operation, Φ is a sign function that returns 0/1 to elements with values smaller/larger than 0, σ is a pre-defined hard threshold. We have conducted an ablation study on it and selected 0.6 as the optimal choice (please refer to the experiment part). As shown in Fig. 3, the outputs of CAS will be fed to either the NI stream or the SCI stream. Thus, our content-aware mutual joint interaction can avoid negative interferences since the input features (pictures or textures) have been correctly separated by CAS.

Rational of Mutual Joint Interaction and Training. Intuitively, improving the performance of the NI stream can be achieved by either training on more data or using an already trained NL model to train the SCI stream further. However, it is crucial to emphasize the mutual joint interaction and training of the NI/SCI streams, which serve two essential purposes. Firstly, the primary role of the NI stream is to separate SCIs into pictures and textures. Since the NI stream is trained using NIs, it is necessary to retrain the model to learn various image feature information to improve separation accuracy. This is especially important as NIs differ slightly from pictures within SCIs. Secondly, the mutual joint interaction and training of the NI and SCI streams can unify the two tasks, reducing the need for designing complex networks for each separate task. By incorporating this approach, the NI and SCI streams can be trained iteratively with more data, achieving improved performance of both tasks.

[2]Here, "one-way" means the parameters of N2S ghost are simply copied from NI Encoder, where the updating of N2S ghost does not affect the NI Encoder. While "bi-way" denotes S2N ghost and NI Encoder influence and update each other.

C. Ranking Head

As shown in Fig. 3-**P3**, we allocate two ranking heads, including 1) one **NI H**ead (NIH) for the NI stream, which is made up of a two-layer MLP; and 2) four **SCI H**eads (SCIH) for the SCI stream, each made up of a two-layer MLP.

The NIH only takes the output of the NI Encoder as input and predicts one binary ranking prediction $\mathrm{R}_{out}^{\mathrm{NI}}$ (0/1), yet the SCIH takes "four inputs" (f_p, f_t, $\tilde{f}_c$ and f_s) from the last layer of NI Encoder (Ghost), SI Encoder (Ghost), CARF, and SI encoder, respectively, and predicts one binary ranking prediction $\mathrm{R}_{out}^{\mathrm{SCI}}$ by combining the "four outputs". The reason is that the CAS module is used to separate the pictural regions from the SCI and will be used to bi-feed the NI stream. However, this is done assuming that the pictural regions obey the same ranking as the whole SCI, which is not guaranteed. Thus, we dynamically combine these four ranking outputs to avoid the bias of any sub-region, *i.e.*, pictural regions, texture regions, and fused features of CARF.

Instead of directly averaging the four relative ranking predictions, we allocate four learnable parameters ($\alpha_i, i \in \{1,2,3,4\}$ to each input feature via "learnable aggregation", and the combined output $\mathrm{R}_{out}^{\mathrm{SCI}}$ can be formulated as:

$$\begin{aligned}\mathrm{R}_{out}^{\mathrm{SCI}} \;=\; \mathrm{ArgMax}\big(&\alpha_1 \times \mathrm{SCIH}_1(f_p) + \alpha_2 \times \mathrm{SCIH}_2(f_t) + \\ &\alpha_3 \times \mathrm{SCIH}_3(\tilde{f}_c) + \alpha_4 \times \mathrm{SCIH}_4(f_s)\big).\end{aligned} \tag{5}$$

where $\mathrm{SCIH}_i,\ i \in \{1,2,3,4\}$, returns the classificaiton confidences (*i.e.*, probability) of each of the "four inputs"; ArgMax is the argmax function, returning 0 or 1.

D. Loss Function

The total loss functions of the whole training phase can be summarized as follows:

$$\mathrm{L} \;=\; \mathrm{L}_{\mathrm{NI}} + \mathrm{L}_{\mathrm{SCI}}, \tag{6}$$

where both L_{NI} and $\mathrm{L}_{\mathrm{SCI}}$, the loss function of NI stream and SCI stream, adopt the widely-used Fidelity Loss [50]. Notice that each input image pair will be sequentially sent to the network, and the loss is measured only if both images in an image pair have been fed.

The Fidelity Loss is a binary cross-entropy-based function designed to measure the discrepancy between predicted quality ranking probabilities and ground-truth relative labels. It is formulated as:

$$\mathrm{L}_{\text{Fidelity}} = -\frac{1}{N}\sum_{i=1}^{N}\left[p_i \log q_i + (1-p_i)\log(1-q_i)\right], \tag{7}$$

where $p_i \in [0,1]$ is the ground-truth label (1 if the first image has higher quality), $q_i \in [0,1]$ is the predicted ranking probability, and N is the number of image pairs in a batch.

E. Image Quality Assessment Inferencing

The testing phase of our approach has been shown in the right part of Fig. 3, where only one testing image (it could be either SCI or NI) will be sent to the network, flowing through both the NI stream and SCI stream (detailed discussion can be seen in Sec. IV-I). Note that in the testing phase, we only utilize the classification confidences of the two streams' final predictions $\mathrm{R}_{out}^{\mathrm{NI}}$ and $\mathrm{R}_{out}^{\mathrm{SCI}}$. Then, the final ranking score ($\mathrm{R}_{\mathrm{Final}}$) can be computed via content-aware selective fusion, which can be formulated as below:

$$\mathrm{R}_{\mathrm{Final}} \;=\; \beta_{\mathrm{NI}} \times \mathrm{R}_{con}^{\mathrm{NI}} + \frac{\beta_{\mathrm{NI}} \times \beta_{\mathrm{SCI}}}{\beta_{\mathrm{NI}} + \beta_{\mathrm{SCI}}} \times \mathrm{R}_{con}^{\mathrm{SCI}}, \tag{8}$$

where β_{NI} and β_{SCI} respectively denote the pixel-value entropy (in grayscale) of picture regions and texture regions, and the picture and texture split is guided by Eq. 4; $\mathrm{R}_{con}^{\mathrm{NI}}$ and $\mathrm{R}_{con}^{\mathrm{SCI}}$ are the classification confidences of the two streams' outputs $\mathrm{R}_{out}^{\mathrm{NI}}$ and $\mathrm{R}_{out}^{\mathrm{SCI}}$.

After obtaining the classification confidence of $\mathrm{R}_{con}^{\mathrm{Final}}$ from the ranking model, we need to regress the absolute quality score of all images in each dataset to perform a quantitative comparison. In contrast to [6], we compare each image $I_i^n \in D_n$ with all other images in D_n. We sample 50 images from each dataset at a time to calculate the classification confidence R_{con}^{i} using the trained ranking model. Each R_{con}^{i} corresponds to a real absolute quality score (MOS/DMOS). We then determine the corresponding real absolute quality score (MOS/DMOS) of the closest classification confidence R_{con}^{i} to the classification confidence of $\mathrm{R}_{con}^{\mathrm{Final}}$ as the final predicted regression quality score of I_i^n. This process repeats 1,000 times to increase accuracy, and the results are averaged.

TABLE I
SUMMARY OF 6 IQA DATABASES, *i.e.*, LIVE [38], CSIQ [39], KADID(KADID-10K) [13], TID2013 [51], CLIVE(LIVE-CHALLENGE) [52], KONIQ(KONIQ-10K) [53], AND 2 SC-IQA DATABASES, *i.e.*, SIQAD [14] AND SCID [54]. # REF.: THE NUMBER OF REFERENCE IMAGE; # DIST.: THE NUMBER OF DISTORTED IMAGE; # D. TYPE: THE NUMBER OF DISTORTED TYPIES; R. TYPE: RATING TYPES; SYN: SYNTHETIC; AUT: AUTHENTIC.

	DataBases	#Ref.	#Dist.	D. Type	#D. Type	#Rating	R. Type	S. Range
NI Databases	LIVE	29	779	syn.	5	25k	MOS	[0, 100]
	CSIQ	30	866	syn.	6	5k	DMOS	[0, 1]
	KADID-10K	81	10.1k	syn.	25	30.4k	MOS	[1, 5]
	TID2013	25	3,000	syn.	25	524k	MOS	[0, 9]
	CLIVE	-	1,162	aut.	-	350k	MOS	[1, 100]
	KonIQ-10K	-	10,073	aut.	-	1.2m	MOS	[1, 100]
SCI Databases	SIQAD	20	980	syn.	7	288k	DMOS	[0, 100]
	SCID	40	1,800	syn.	9	-	MOS	[0, 100]

IV. EXPERIMENTS

A. Datasets and Metrics

We evaluate the performance of our proposed model extensively on 8 IQA datasets, including 6 NI sets and 2 SCI sets. Among the adopted NI sets, LIVE [38], CSIQ [39], KADID-10K [13], and TID2013 [51] are synthetically distorted, and LIVE-Challenge [52] and KonIQ-10K [53] are authentically distorted. The adopted SCI sets are SIQAD [14] and SCID [54]. The configurations of these databases are presented in Table I. We adopted three evaluation metrics: **s**pearman **r**ank-order **c**orrelation **c**oefficient (SRCC) and **p**earson **l**inear **c**orrelation **c**oefficient (PLCC) and **r**oot **m**ean **s**quare **e**rror (RMSE). The higher the PLCC and SRCC values are, the lower the RMSE value and the better the NI-IQA/SCI-IQA approach.

TABLE II
QUANTITATIVE COMPARISON OF OUR UNI-IQA *v.s.* SOTA NI-IQA METHODS ON SIX WIDELY-USED NI DATASETS. SOME DATA SHOWN IN THE TABLE ARE DIRECTLY BORROWED FROM [55] BECAUSE SOME MODELS' CODES ARE NOT PUBLICLY AVAILABLE. THE TOP-2 ARE MARKED IN RED AND BLUE.

Metrics&Sets / Models	LIVE		CSIQ		KADID-10K		TID2013		CLIVE		KonIQ-10K	
	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑
DBCNN$_{20}$	0.971	0.968	0.959	0.946	0.856	0.851	0.865	0.816	0.869	0.869	0.884	0.875
MetaIQA$_{20}$	0.959	0.960	0.908	0.899	0.775	0.762	0.868	0.856	0.802	0.835	0.856	0.887
P2P-BM$_{20}$	0.958	0.959	0.902	0.899	0.849	0.840	0.856	0.862	0.842	0.844	0.885	0.872
HyperIQA$_{20}$	0.966	0.962	0.942	0.923	0.845	0.852	0.858	0.840	0.882	0.859	0.917	0.906
TIQA$_{21}$	0.965	0.949	0.838	0.825	0.855	0.850	0.858	0.846	0.861	0.845	0.903	0.892
UNIQUE$_{22}$	0.968	0.969	0.927	0.902	0.876	0.878	0.884	0.868	0.890	0.854	0.901	0.896
TReS$_{22}$	0.968	0.969	0.942	0.922	0.858	0.859	0.883	0.863	0.877	0.846	0.928	0.915
Re-IQA$_{23}$	0.970	0.971	0.947	0.946	0.872	0.885	0.804	0.861	0.840	0.854	0.914	0.923
TempQT$_{23}$	0.975	0.977	0.950	0.945	0.878	0.872	0.891	0.877	0.870	0.872	0.903	0.920
LIQE$_{23}$	0.970	0.951	0.939	0.936	0.931	0.930	-	-	0.910	0.904	0.908	0.919
Q-Align$_{23}$	0.840	0.870	0.876	0.845	0.935	0.934	-	-	0.887	0.883	0.935	0.934
IQDLNet$_{23}$	0.974	0.974	0.942	0.940	0.866	0.867	0.882	0.873	0.873	0.869	0.921	0.907
PMLR$_{24}$	0.963	0.970	0.958	0.945	0.898	0.895	0.892	0.877	0.848	0.797	0.901	0.881
ARNIQA$_{24}$	0.966	0.970	0.962	0.947	0.908	0.912	0.880	0.901	0.893	0.870	0.933	0.921
ACVA$_{24}$	0.970	0.976	0.960	0.946	0.912	0.907	-	-	0.871	0.862	0.921	0.909
UNI-IQA (Ours)	0.981 ±0.007	0.979 ±0.003	0.963 ±0.002	0.949 ±0.012	0.883 ±0.015	0.875 ±0.020	0.899 ±0.011	0.878 ±0.006	0.885 ±0.019	0.874 ±0.002	0.937 ±0.004	0.928 ±0.018

TABLE III
QUANTITATIVE COMPARISON OF OUR UNI-IQA *v.s.* SOTA SCI-IQA MODELS ON TWO SCI DATASETS. SOME DATA SHOWN IN THE TABLE ARE DIRECTLY BORROWED FROM [36] BECAUSE SOME MODELS' CODES ARE NOT PUBLICLY AVAILABLE. THE TOP-2 ARE MARKED IN RED AND BLUE.

Metrics&Sets / Methods	SIQAD			SCID		
	PLCC↑	SRCC↑	RMSE↓	PLCC↑	SRCC↑	RMSE↓
PICNN$_{18}$	0.896	0.897	6.790	0.827	0.822	8.013
TFSR$_{18}$	0.862	0.835	7.491	0.802	0.784	8.804
QOD$_{19}$	0.899	0.887	6.516	0.839	0.808	7.824
CBIQA$_{19}$	0.911	0.898	5.893	0.853	0.838	7.393
PQSC$_{20}$	0.916	0.907	5.708	0.918	0.915	5.480
HAMTL$_{21}$	0.909	0.901	5.857	0.874	0.866	6.799
MTD$_{21}$	0.916	0.909	5.711	0.881	0.873	6.703
SAE$_{22}$	0.874	0.854	6.936	0.787	0.756	8.595
DCST$_{22}$	0.926	0.924	5.702	0.913	0.905	6.254
MFSD$_{23}$	0.922	0.914	5.789	0.886	0.882	6.214
EHR$_{23}$	0.931	0.924	5.753	0.919	0.925	5.124
HIDRO$_{24}$	0.934	0.927	5.762	0.915	0.921	5.101
SSL-VQA$_{24}$	0.929	0.921	5.778	0.904	0.916	5.346
UNI-IQA (Ours)	0.941 ±0.005	0.933 ±0.009	5.625 ±0.054	0.922 ±0.012	0.928 ±0.006	5.074 ±0.043

B. Implementation Details

We implement our model by PyTorch in an NVIDIA GeForce RTX 3090 GPU. The sampled paired images are scaled to 256×256×3 and randomly cropped to 224×224×3 as the inputs of our network. The batch size in the training phase is 32, and we adopt the Adam optimizer for optimization. The learning rate is 5e-5 with a weight decay set of 1e-3, and a warm-up training strategy is adopted.

Following [6], we randomly sampled 80% images from each NI/SCI dataset to construct the training sets for the NI stream and SCI stream, and the testing sets are the 20% rests. All quantitative experiments have adopted the 10-fold cross-verification. The model becomes converged after 200 training epochs for eleven hours.

C. Comparisons with the SOTA Models

1) Evaluation on NI Datasets: To prove the effectiveness of our UNI-IQA on NI sets, we have compared our method against 13 most recent SOTA NI-IQA models, *i.e.*, DBCNN [11], TIQA [56], MetaIQA [25], UNIQUE [9], P2P-BM [57], HyperIQA [58], TReS [59], Re-IQA [60], TempQT [61], IQDLNet [62], PMLR [63], ACVA [64] and ARNIQA [65] over six widely-used NI sets. The results have been reported in Table II. Our model outperforms all other competitors by large margins, specifically on LIVE (+0.6%) and CSIQ (+1.5%) sets regarding the PLCC metric, respectively. In realistic distortion-based sets, *i.e.*, the LIVE Challenge, our method still shows competitive results, *e.g.*, 0.870 (ARNIQA) *v.s.* 0.874 (Ours) in terms of the SRCC metric. In a word, this experiment has verified that our model can perform well on the NI datasets. Notice that, in this experiment, our model was trained on both NI and SCI training sets, yet other NI competitors were only trained on NI sets. We are fully aware that the additional gain might be achieved by using additional data, but achieving joint learning on both SCI and NI sets is the key innovation of our approach. And other models trained on both NI and SCI sets could result in significant performance degeneration, and this issue has been fully investigated by [9]. *W.r.t.* the fairness comparison, we have conducted an additional verification in the 3rd part of this experiment.

2) Evaluation on SCI Datasets: We have compared our method against 11 most recent SOTA SCIs-based IQA models, *e.g.*, PICNN [66], TFSR [67], QOD [68], CBIQA [69], SAE [19], HAMTL [70], MTD [71], PQSC [17], DCST [36], MFSD [72], EHR [73], and 2 SOTA screen content video

TABLE IV
FAIR COMPARISONS AMONG THE LEARNING-TO-RANK SOTA MODELS. "-" DENOTES THE MISSING DATA DUE TO THE UNAVAILABLE CODES.

		NI Datasets				SCI Datasets			
	Metrics&Sets	KADID-10K		KonIQ-10K		SIQAD		QACS(HECS)	
	Models	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑
NI Models	UNIQUE$_{21}$ (TID+LIVE+SIQAD)	0.727	0.762	0.765	0.748	0.559	0.530	0.519	0.501
NI Models	TReS$_{22}$ (TID+LIVE+SIQAD)	0.650	0.656	0.751	0.737	0.688	0.699	0.582	0.532
NI Models	**Ours** (TID+LIVE+SIQAD)	**0.755**	**0.787**	**0.803**	**0.795**	**0.827**	**0.752**	**0.610**	**0.639**
SCI Models	UDA$_{21}$ (TID+SIQAD)	-	-	-	-	0.711	0.697	0.765	0.760
SCI Models	**Ours** (TID+SIQAD)	-	-	-	-	**0.875**	**0.843**	**0.786**	**0.781**
SCI Models	UDA$_{21}$ (LIVE+SIQAD)	-	-	-	-	0.702	0.690	0.799	0.764
SCI Models	**Ours** (LIVE+SIQAD)	-	-	-	-	**0.858**	**0.832**	**0.812**	**0.771**

quality assessment models[3] HIDRO [74], SSL-VQA [75] over two widely-used SCI datasets. As shown in Table III, our method yields the best overall performance in terms of PLCC (+1.1%), SRCC (+0.7%), and RMSE (-0.015) on SCID. This experiment has verified that, despite the good performance of our model archived in NI datasets, our model can also perform well on SCI datasets. Notice that, as mentioned above, our model was trained on both NI and SCI training sets, yet other NI competitors were only trained on the SCI sets. For a complete fairness comparison, please see below.

3) "Fair" Comparison on both NI and SCI Sets: Here we have adopted three competitors who have followed the learning-to-rank methodology to handle the datasets misalignment issue. The first two are NIs-based IQA models (UNIQUE [9] and TReS [59]), which have released the codes, and we have retrained them on the TID2013 set and SIQAD set. The last one (UDA [6]) is SCIs-based, trained on the TID2013 & SIQAD sets and LIVE & SIQAD sets, but the codes are not publicly available. Thus, we borrow the results reported in the paper directly. As shown in Table IV, when trained with the same datasets (TID2013 + LIVE + SIQAD), our method sharply outperforms NI-IQA competitors. The main reason is that our approach ensures the mutual promotion of NI-IQA and SCI-IQA tasks. Also, the performance of our method is superior to the SCI-IQA competitor. This experiment can fully demonstrate the advantage of our approach against the existing method in handling dataset misalignment.

D. Component Evaluation

We have conducted an extensive component evaluation to verify the effectiveness of major components used in our approach ("*duplex dynamic encoder*" and "*ranking heads*"), and the quantitative results can be seen in Table V. Mark ⓪ serves as the baseline and consists of three experimental settings (lines 0-2). In lines 0-1, each setting involves either the NI encoder or the SI encoder separately, along with the corresponding ranking head, to train each task independently. In line 2, both the plain NI and SI encoders are utilized, and the outputs of the two streams are combined through a bilinear operation with a single ranking head.

[3]To the best of our knowledge, as of now, there have been no recent works on screen content image quality assessment tasks in 2024. Thus, we comapared our method against two most recent SOTA screen content video quality assessment models.

TABLE V
COMPONENT EVALUATION ON THE MAJOR COMPONENTS.

	Key Components								Datasets			
	Duplex Dynamic Encoder						Head		KADID-10K		SIQAD	
	NIE	S2N	CAS	SCIG	SCIE	CARF	SCIH	NIH	PLCC↑	SRCC↑	PLCC↑	SRCC↑
0	✔	✘	✘	✘	✘	✘	✘	✔	0.840	0.835	0.887	0.882
⓪①	✘	✘	✘	✘	✔	✘	✔	✘	0.837	0.832	0.885	0.879
2	✔	✘	✘	✘	✔	✘	✔	✔	0.844	0.839	0.891	0.885
①3	✔	✔	✔	✘	✔	✘	✔	✔	0.849	0.847	0.896	0.899
② 4	✔	✘	✔	✔	✘	✘	✔	✔	0.854	0.855	0.909	0.904
5	✔	✔	✔	✔	✘	✘	✔	✔	0.860	0.855	0.914	0.908
③6	✔	✔	✔	✔	✔	✘	✔	✔	0.863	0.859	0.919	0.910
⑤7	✔	✔	✔	✔	✔	✔	✘	✔	0.875	0.869	0.930	0.922
④ 8	✔	✔	✔	✔	✔	✔	✔	✔	**0.883**	**0.875**	**0.941**	**0.933**

⓪ **Baseline**
⓪① *Verify* **CAS**
② *Verify* **S2N & SCIG**
③⑤ *Verify* **CARF**
④ *Verify* **SCIH**

NIE: NI Encoder
NIH: NI Head
SCIH: SCI Head
CAS: Content-aware Data Switch
CARF: Cascaded Adaptive Region Fusion
SCIE: SI Encoder
SCIG: SI Encoder (Ghost)
S2N: NI Encoder (Ghost)

Fig. 6. Visualization of pictorial and textural regions in the SCIs generated by the "content-aware" data switch.

The effectiveness of **c**ontent-**a**ware **s**witch (CAS) (Eq. 4) to interact NI stream and SCI stream can be confirmed by comparing ⓪ and ①. As shown, our CAS (merely feeding pictorial-region SCI into S2N Ghost) can achieve persistent performance improvements on all metrics, *e.g.*, the PLCC metric has increased from .837→.849 and .885→.896 in KADID-10K and SIQAD testing set. Fig. 6 provides a visualization of pictorial regions and textural regions in the SCIs generated by the proposed "content-aware" data switch.

As shown by mark ②, compared with the models equipped with SI encoder ghost (SCIG) only (line 2 in Table V) or merely feeding pictorial-region-only SCI into NI encoder ghost (S2N, line 1), a model equipped with both SCIG and S2N (line 3) can achieve significant performance improvements, *e.g.*, boosting the PLCC metric in SIQAD set for about 0.5% (line 2) and 1.1% (line 1), respectively. This experiment suggests that the SCIG can contribute more to the SIQAD set, an SCI set, than S2N. Thus, we can confirm that the proposed CAS can compress those irrelevant picture-related features to harm the SCI stream.

As denoted by mark ③ and line 8, the advantage of our CARF (Fig. 4) can be easily observed, *e.g.*, .859 *vs.* .875

TABLE VI
COMPARISON RESULTS BETWEEN "MUTUAL INTERACTIONS" AND SINGLE-WAY INTERACTIONS BY FIXING EITHER NI OR SCI STREAMS AND ONLY TRAINING ANOTHER IN TERMS OF TWO COLLECTION WAYS OF THE FINAL RESULTS.

Metrics&Sets / Choices	KADID-10K		SIQAD	
	PLCC↑	SRCC↑	PLCC↑	SRCC↑
Fix NI Stream (with SCI Head)	-	-	0.905	0.891
Fix NI Stream (with SCI&NI Heads)	-	-	0.923	0.918
Fix SCI Stream (with NI Head)	0.864	0.858	-	-
Fix SCI Stream (with SCI&NI Heads)	0.872	0.869	-	-
Ours	**0.883**	**0.875**	**0.941**	**0.933**

regarding the SRCC metric in KADID-10K. The reason is that the CARF can fully integrate and enhance the features of the local textual regions and global SCIs. Without CARF, the SCI stream cannot fully take advantage of those valuable features obtained via CAS.

The effectiveness of using "learnable aggregation" head (Eq. 5) has been shown by mark ④, in which four learnable weights are applied to four relative ranking scores. By comparing line 7 (simply averaging the four ranking scores) and line 8 (with learnable weights), the PLCC metric over the SIQAD set has increased by 1.1%.

E. Ablation Study

1) Effectiveness on Mutual Interaction of NI & SCI Sets: To demonstrate the effectiveness of "mutual interaction" between NI and SCI sets, we conducted two single-way interaction settings to verify the mutual promotions between NI and SCI streams. In the first setting, we pre-trained the NI stream on NI datasets and fixed it while training the SCI stream. In the second setting, we pre-trained the SCI stream on SCI datasets and fixed it while training the NI stream. In both experiments, we obtained the final results using our proposed method (eq. 8) and using only the SCI head or the NI head. The experimental results in Table VI show that compared to the "mutual interaction" of NI & SCI streams (line 5), single-way interactions (lines 1-4) where NI/SCI streams were fixed, regardless of the collection ways of the final results, are suboptimal in achieving mutual promotion.

To ensure a fair comparison, we conducted re-training of several NI and SCI models using joint learning on all six NI sets or two SCI sets to confirm the effectiveness of our "mutual interaction". We separately re-trained SOTA open-source NI models such as TRes, GraphIQA, and VCRNet using both original training sets and all six NI sets. For the only open-source SCI model to date, RIQA [4], we also re-trained it using both original training sets and both two SCI sets. The results shown in Table. VII indicate that while training on multiple sets can bring slight performance improvement compared to a small number of original training sets, our mutual interaction based on the learning-to-ranking approach produces the best performance improvement for both NI and SCI sets. For instance, in terms of the PLCC metric, our method achieves a performance gain of 2.5% and 3.3% *vs.* 1.1% and 1.1% (TReS) for the KADID-10K set and SIQAD set, which highlights the validity and rationality of our mutual interaction.

We investigated the impact of the volume of datasets on UNIQUE, TReS, and our re-trained method, which was trained with gradually increasing amounts of synthetically-distorted NI sets, authentically-distorted NI sets, and SCI sets. Table VIII presents the SRCC results. Initially, when using only LIVE, TID, and SIQAD datasets, our UNI-IQA model's performance was inferior to the other two models on both NI testing set KADID, and SCI testing set SIQAD. However, as the amount of data increased, all the compared models showed improved performance. When the number of datasets reached a certain point and continued to increase, the performance growth rate of UNIQUE and TReS slowed down, while our UNI-IQA model continued to increase rapidly. Ultimately, our UNI-IQA outperformed the other compared models after training with all the datasets. These findings highlight the ability of our proposed UNI-IQA model to effectively handle both NI and SCI sets while also demonstrating its potential for handling even larger volumes of training data. Overall, our results suggest that our UNI-IQA is a promising solution for improving IQA.

2) Performance Evaluations Cross Different Databases: We tested the generalization ability of our proposed UNI-IQA in a more difficult cross-database setting. To achieve this, we conducted experiments where both synthetically and authentically distorted NI databases were crossed with SCI databases and vice versa. We trained our model on combined datasets consisting of sampled NI and SCI datasets simultaneously and tested it on a different dataset without any additional finetuning or parameter adaptation. To provide a basis for comparison, we also retrained two top-performing open-source IQA-based methods - VCRNet [28] and GraphIQA [27], as well as two learning-to-rank-based methods - UNIQUE [9] and TReS [59]. As shown in Table IX, UNI-IQA outperformed the two NIQA-based and two learning-to-rank-based models, achieving significantly better SRCC results. This demonstrates that the UNI-IQA model has a strong generalization ability.

3) Effectiveness on σ: We have tested multiple choices regarding σ (Eq. 4), and the exact results can be found in Table X. As shown, the overall performance of our method is moderately sensitive to the choice of σ. Specifically, $\sigma = 0.6$ achieves the best result, and $\sigma = 0.8$ is inferior to $\sigma = 0.5$, *e.g.*, 0.921 *v.s.* 0.929 in PLCC metric regarding SIQAD dataset, showing a larger σ may not always bring performance gain. The main reason lies in two aspects. First, when σ uses a large value, more regions in an SCI are denoted as textures, leading to fewer picture regions to benefit the NI stream. Second, when σ chooses a very small value, more regions are denoted as pictures, shrinking the contribution of CARF to the SCI stream.

We have also conducted an ablation study to confirm the essentiality of entropies β_{NI} and β_{SCI} (eq. 8) that function as weights for the prediction of both NI and SCI streams. Table X presents the results, which indicate that using only one entropy during the testing phase, such as β_{NI} or β_{SCI}, results in suboptimal outcomes compared to our proposed method in (eq. 8). Regarding the PLCC metric, without considering β_{NI}, the performance of the SCI set SIQAD will decrease if we use only β_{SCI}, *i.e.*, 0.932 *v.s.* 0.941. Similarly, without considering β_{SCI}, the performance of the NI set KADID-10K will decrease if we use only β_{NI}, *i.e.*, 0.876 *v.s.* 0.883. Additionally, the

TABLE VII
COMPARISON RESULTS ON RE-TRAINING SEVERAL REPRESENTATIVE NI AND SCI MODELS USING JOINT LEARNING ON ALL SIX NI SETS OR TWO SCI SETS.

	Metrics&Sets	NI Datasets: TID2013		KADID-10K		CLIVE		KonIQ-10K		SCI Datasets: SIQAD		SCID	
	Models	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑
NI models	**TReS** (trained on original sets)	0.883	0.863	0.858	0.859	0.877	0.846	0.928	0.915	**0.908**	**0.895**	**0.784**	**0.821**
NI models	**TReS** (trained on 6 SCI sets+2 SCI sets)	0.892	0.871	0.869	0.867	0.880	0.860	0.932	0.924	**0.919**	**0.907**	**0.885**	**0.902**
NI models	**GraphIQA** (trained on original sets)	0.849	0.852	0.858	0.851	0.862	0.845	0.915	0.911	-	-	-	-
NI models	**GraphIQA** (trained on 6 SCI sets)	0.855	0.869	0.866	0.857	0.873	0.860	0.922	0.916	-	-	-	-
NI models	**VCRNet** (trained on original sets)	0.875	0.846	0.862	0.859	0.865	0.856	0.909	0.894	-	-	-	-
NI models	**VCRNet** (trained on 6 SCI sets)	0.884	0.861	0.870	0.865	0.869	0.862	0.918	0.904	-	-	-	-
SCI models	**RIQA** (trained on original sets)	-	-	-	-	-	-	-	-	0.911	0.900	0.859	0.848
SCI models	**RIQA** (trained on 2 SCI sets)	-	-	-	-	-	-	-	-	0.935	0.927	0.892	0.889
	Ours (trained on 6 NI+2 SCI sets)	**0.899**	**0.878**	**0.883**	**0.875**	**0.885**	**0.874**	**0.937**	**0.928**	**0.941**	**0.933**	**0.922**	**0.928**

TABLE VIII
ABLATION STUDY OF THE VOLUME OF DATASETS ON SRCC METRIC. THE BEST RESULTS ARE MARKED IN **BOLD**.

1: LIVE+TID 2: CSIQ+KADID 3: CLIVE 4: KonIQ-10K 5: SIQAD 6: SCID

Training Sets	1+5		1+2+5		1+2+3+5		1+2+3+4+5		1+2+3+4+5+6	
Testing Sets	KADID	SIQAD	KADID	SIQAD	KADID	SIQAD	KADID	SIQAD	KADID	SIQAD
UNIQUE	0.635	0.698	0.723	0.765	0.799	0.822	0.842	0.858	0.860	0.881
TReS	**0.829**	**0.842**	**0.835**	**0.859**	**0.841**	0.868	0.849	0.879	0.867	0.895
Ours	0.640	0.714	0.682	0.821	0.724	**0.903**	**0.852**	**0.914**	**0.875**	**0.933**

TABLE IX
SRCC RESULTS OF THE CROSS-DATASET EVALUATION. NUMBERS DENOTE THE SINGLE DATASET OR COMBINED DATASETS. THE BEST RESULTS ARE MARKED IN **BOLD**.

	NI Datasets						SCI Datasets	
Training Sets	1+5+6		2+5+6		3+5+6	4+5+6	1+2+3+4+5	1+2+3+4+6
Testing Sets	KADID	CSIQ	LIVE	TID	KonIQ	CLIVE	SCID	SIQAD
GraphIQA	0.635	0.774	0.850	0.618	0.672	**0.815**	-	-
VCRNet	0.644	0.738	0.699	0.623	0.681	0.746	-	-
UNIQUE	0.641	0.769	**0.852**	0.687	0.685	0.725	0.591	0.582
TReS	0.637	0.695	0.725	0.582	0.646	0.705	0.713	0.726
Ours	**0.646**	**0.780**	0.754	**0.698**	**0.690**	0.712	**0.884**	**0.897**

1: LIVE+TID 2: CSIQ+KADID 3: CLIVE 4: KonIQ-10K 5: SIQAD 6: SCID

TABLE X
ABLATION STUDY OF THRESHOLD σ (EQ. 4).

Metrics&Sets	KADID-10K		SIQAD	
Choices	PLCC↑	SRCC↑	PLCC↑	SRCC↑
σ = 0.5	0.872	0.864	0.929	0.918
σ = 0.6 (Ours)	**0.883**	**0.875**	**0.941**	**0.933**
σ = 0.7	0.869	0.862	0.928	0.915
σ = 0.8	0.863	0.857	0.921	0.906
w/o β_{NI}	0.866	0.857	0.932	0.925
w/o β_{SCI}	0.876	0.868	0.921	0.916
β_{NI} + β_{SCI} **(Ours)**	**0.883**	**0.875**	**0.941**	**0.933**

study found that removing the β_{NI} has less of an impact on the SCI set SIQAD than removing the β_{SCI}, and taking away the β_{SCI} has less of an impact on the NI set KADID-10K than taking away the β_{NI}. Therefore, both entropies β_{NI} and β_{SCI} are necessary to obtain optimal results.

4) Effectiveness on Encoder Backbones: In our newly proposed UNI-IQA, we choose tailored SCNN[4] (denoted as T-SCNN) pre-trained on massive distorted images as the backbone of the **N2S** and **S2N**, and pre-trained tailored VGG16 (denoted as T-VGG16) as the backbone of SI encoder and SI encoder (Ghost), respectively. We compare this

[4]Details can be seen in [11].

TABLE XI
ABLATION STUDY OF THE ENCODER BACKBONE CHOICES.

Metrics&Sets	KADID-10K		SIQAD	
Backbone Chioces	PLCC↑	SRCC↑	PLCC↑	SRCC↑
VGG16+ResNet50	0.865	0.862	0.922	0.915
ResNet50+VGG16	0.869	0.865	0.926	0.917
T-SCNN+ResNet50	0.874	0.871	0.932	0.926
T-SCNN+T-VGG16 (Ours)	**0.883**	**0.875**	**0.941**	**0.933**

TABLE XII
ABLATION STUDY OF THE CASCADED ADAPTIVE REGION FUSION (CARF).

Metrics&Sets	KADID-10K		SIQAD	
Choices	PLCC↑	SRCC↑	PLCC↑	SRCC↑
Addition	0.861	0.851	0.918	0.909
Multiply	0.866	0.853	0.915	0.911
Concat	0.872	0.859	0.919	0.913
Ours (*w/o* out)	0.875	0.865	0.927	0.922
Ours (*w* out)	**0.883**	**0.875**	**0.941**	**0.933**

combination with other choices such as {VGG16+ResNet50}, {ResNet50+VGG16}, and {SCNN + ResNet50}, all of which are pre-trained by ImageNet. As shown in Table XI, our default combination obtains better results. Also, though VGG16 is more lightweight than ResNet50, pre-trained VGG16 on massive distorted images outperforms ResNet50 can also prove that, *e.g.*, 0.875 *v.s.* 0.871 on the SRCC metric regarding KADD-10K set.

F. Effectiveness of CARF

To verify the effectiveness of the proposed **c**ascaded **a**daptive **r**egion **f**usion (CARF, Fig. 4), we have compared it with other simple fusion methods, such as addition, multiplication, and concatenation, and removed the output head of CARF (denoted as w/o out). Results in Table XII show that our CARF outperforms other simple fusion methods by a large margin, *e.g.*, our CARF achieves average performance gains of 2.1% and 2.0% in the SRCC metric of KADID-10K and SIQAD respectively, which demonstrates the effectiveness of CARF to integrate the textual features and the whole SCI features fully. Meanwhile, removing the output head of CARF performs slightly worse (0.865 *v.s.*,0.875). The reason might

TABLE XIII
MODEL SIZE AND RUNNING TIME COMPARISONS.

		← NI-IQA Methods →		← SCI-IQA Methods →	
Methods	Ours	UNIQUE	TReS	UDA	RIQA
Model Size↓	258MB	85MB	582MB	301MB	121MB
FPS↑	7	21	4	9	16

be that the fused features contain more informative cues better to highlight various regions in the whole SCI image scene.

G. Running Time Comparisons

We compared our proposed method's model size and running time with SOTA methods that have released codes or are easily reproducible. For NI-IQA, we considered UNIQUE [9] and TReS [59], while for SCI-IQA, we compared our method with UDA [6] and RIQA [4]. As shown in Table XIII, our method achieved real-time speed with 7 FPS during the inference phase, while its running time and model size was comparable to the other NI/SCI-IQA methods. However, our method outperformed all other models on both NI and SCI datasets. Therefore, our method offers a good balance between speed, size, and accuracy.

H. Discussion of A Unified Approach to Assess the Quality of Natural and Screen Content Images

The need for a unified approach to assessing the quality of natural and screen content images goes beyond the availability of labeled data. The reasons are below.

First, while it is true that there is labeled data available for both NIs and SCIs, the distortion types and subjective scores assigned by participants are often not well aligned between the two domains, making it difficult to compare the performance of existing IQA models directly. Moreover, existing fusion schemes that combine NIs and SCIs-based IQA often fail to achieve mutual performance promotion because they are content-unaware. Second, by developing a content-aware data switch, our approach aims to enhance the commonness between NIs and SCIs while compressing the discrepancies, allowing for full end-to-end "mutual interactions" between the two tasks. This approach improves IQA models' performance on both domains and lays the foundation for future research on developing more advanced techniques that can bridge the gap between NIs and SCIs-based IQA. Third, a unified approach to image quality assessment can simplify the inferencing process and improve efficiency, as we can use a single model for both natural and screen content images instead of requiring separate models for each type of image. This can save time and resources and make image quality assessment more accessible and practical for a wider range of applications.

Therefore, we argue that a unified approach to image quality assessment is reasonable and significant, as it can improve performance, simplify the inferencing process, and make IQA more practical and accessible.

I. Discussion of β_{NI} and β_{SCI} in Inferencing Stage

In the image quality assessment inferencing stage (Eq. 8), we use the entropies of the pictorial and texture regions as weights for the NI and SCI stream predictions, respectively, and compute the final ranking score by summing the output of both streams. This approach is reasonable for two reasons. Firstly, it allows us to use a unified model for both NI and SCI images, eliminating the need for separate models for each image type. This simplifies the inferencing process and makes it more efficient. Secondly, the pictorial regions in an image often contain more discriminative information about the scene category, while the texture regions carry more texture information. Using the content-aware selective fusion approach, we can effectively combine the advantages of both streams and improve the final ranking performance. This approach ensures that the model can leverage the strengths of both pictorial and texture regions to provide a more accurate ranking score.

We have also conducted experiments by directly inputting NIs or SCIs into the corresponding stream in the inferencing stage without interaction with the counterpart stream, and the corresponding performance is calculated separately. Results in Table. have shown that our content-aware selective fusion approach consistently outperforms using the predictions from either stream alone on both the SCI and NI datasets.

J. Why the NIs and SCIs have "common characteristics" since they differ greatly in content and style?

We argue that despite the significant differences in content and style between natural images (NIs) and screen content images (SCIs), there are still some common characteristics in how our human visual system perceives them. We provide three perspectives to answer the question:

First, we agree that the HVS's understanding and feelings towards different content are distinct and involve various physiological brain areas. However, our research does not negate this complexity. Instead, it posits that despite these differences, there is an underlying consistency in how the HVS assesses visual quality. This consistency is influenced by fundamental attributes such as contrast, texture, and edge information, which are prevalent in both SCIs and NIs.

Second, it is indeed true that SCIs and NIs vary greatly in content and style. However, our work explores the intersection where the HVS's assessment of quality converges for these image types. We propose that while the content and style differ, the criteria for quality assessment (like sharpness, color fidelity, and noise) retain a degree of commonality. This commonality is what our model seeks to capture and quantify.

Third, our paper does not claim to fully unravel the complexities of the HVS. Instead, it contributes to the ongoing discourse by highlighting a potential common ground in IQA for SCIs and NIs. We believe this perspective can pave the way for more nuanced and comprehensive models in future research.

K. How is "content-aware" better than "content-unaware"?

It is essential to recognize that while existing fusion schemes, such as the approach by Chen et al. [6], have made strides in addressing the alignment of absolute score ranges through a learning-to-rank strategy, they fall short in achieving

full "mutual interactions" between NIs- and SCIs-based IQA. Specifically, the one-sided promotion from NIs to SCIs is what we call "content unaware", *i.e.*, it does not explicitly consider which regions in the input image correspond to textures and which regions correspond to pictures. This "content-unaware" characteristic limits the ability to achieve full mutual interactions between the two tasks since only the picture-related information in SCIs-based IQA is considered helpful for NIs-based IQA.

In the context of NIs- and SCIs-based IQA, achieving effective multimodal fusion requires the model to have a deep understanding of the nature of the input data. Simply relying on semantic information alone may not be sufficient as different modalities, such as textures and pictures, convey unique and complementary information that can enhance the overall understanding and performance of the model. If the model is unaware of the data type, it may not be able to fully leverage the unique characteristics and specific quality assessment criteria associated with each modality. For instance, in NIs-based IQA, the model may focus on analyzing image sharpness and noise levels, which are crucial for assessing natural images but less relevant for synthetic images. Conversely, in SCIs-based IQA, texture synthesis and rendering quality are important factors that may not be as significant in natural image assessment.

To address this limitation, our proposed UNI-IQA framework introduces a "content-aware" data switch that allows the model to differentiate between pictures and textures. This "content-aware" switch enables the model to have a better understanding of the data and facilitates meaningful interactions between the two modalities. For example, knowing whether the input data is textual or pictural helps the model make informed decisions and leverage the complementary information from both modalities.

L. Limitations

While our study aimed to develop a unified model for assessing the quality of both natural and screen content images, it is important to acknowledge that there are limitations to our approach. 1) One limitation is that we could not directly assess the impact of aspect ratio on quality scores. While we acknowledge that aspect ratio is a well-established phenomenon in the field and can significantly impact quality scores, we were unable to conduct an experiment to assess this effect directly. 2) Another limitation is that our approach involved downsampling the input images to a resolution of $256\times256\times3$. While this allowed us to develop an efficient and effective assessment model, it may not be suitable for assessing images of higher resolution, such as 1920×1080 or 4K. This is because downsampling can result in the loss of important information, impacting the accuracy of the quality score assigned to the image. Future research should aim to address these limitations and develop more comprehensive models for assessing image quality.

V. Conclusion

This paper has presented a unified paradigm to achieve end-to-end mutual joint interactions and performance promotions between NIs- and SCIs-based IQA tasks. To this end, by observing the "common characteristics" of NIs and SCIs, we devise a "content-aware" data switch to make the model automatically enhance the commonness and compress the discrepancies between the two tasks. To alleviate the dataset misalignments and achieve joint training of all NI and SCI datasets, we intergraded the learning-to-rank strategy into the newly designed unified framework by mapping the diverse subjective quality scores into a unified scale. Experimental results on six widely-used NI databases and two SCI databases verified the superiority of our method on both NI- and SCI-IQA tasks compared to other single-task-oriented NI- and SCI-IQA methods. We deem this "content-aware" data switch methodology a promising new take on multi-task learning, potentially of interest to many other research fields. In the future, we plan to explore the unsupervised multi-task learning of image quality assessment, which is more practical in the real world.

Acknowledgments. This research was supported in part by National Natural Science Foundation of China (Grant No. 62172246), the Shandong Provincial Natural Science Foundation of China (Outstanding Young Scientist Program) (Grant No. ZR2024YQ071), and the Fundamental Research Funds for the Central Universities (Grant No. 22CX06037A).

References

[1] Y. Wang, J. G. Yim, N. Birkbeck, J. Ke, H. Talebi, X. Chen, F. Yang, and B. Adsumilli, "Revisiting the efficiency of ugc video quality assessment," in *ICIP*, pp. 3016–3020, 2022.

[2] W. Kim, A.-D. Nguyen, S. Lee, and A. C. Bovik, "Dynamic receptive field generation for full-reference image quality assessment," *IEEE TIP*, vol. 29, pp. 4219–4231, 2020.

[3] H. Zheng, H. Yang, J. Fu, Z.-J. Zha, and J. Luo, "Learning conditional knowledge distillation for degraded-reference image quality assessment," in *ICCV*, pp. 10222–10231, 2021.

[4] X. Jiang, L. Shen, L. Yu, M. Jiang, and G. Feng, "No-reference screen content image quality assessment based on multi-region features," *Neurocomputing*, vol. 386, pp. 30–41, 2020.

[5] S. Wang, K. Gu, X. Zhang, W. Lin, S. Ma, and W. Gao, "Reduced-reference quality assessment of screen content images," *IEEE TCSVT*, vol. 28, no. 1, pp. 1–14, 2018.

[6] B. Chen, H. Li, H. Fan, and S. Wang, "No-reference screen content image quality assessment with unsupervised domain adaptation," *IEEE TIP*, vol. 30, pp. 5463–5476, 2021.

[7] M. Mahmoodpour, A. Amirany, M. H. Moaiyeri, and K. Jafari, "A learning based contrast specific no reference image quality assessment algorithm," in *MVIP*, pp. 1–4, 2022.

[8] J. Xiang, M. Yu, G. Jiang, H. Xu, Y. Song, and Y.-S. Ho, "Pseudo video and refocused images-based blind light field image quality assessment," *IEEE TCSVT*, vol. 31, no. 7, pp. 2575–2590, 2021.

[9] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE TIP*, vol. 30, pp. 3474–3486, 2021.

[10] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *ICCV*, pp. 1040–1049, 2017.

[11] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE TCSVT*, vol. 30, no. 1, pp. 36–47, 2020.

[12] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-adversarial neural networks," *arXiv preprint arXiv:1412.4446*, 2014.

[13] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *QoMEX*, 2019.

[14] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE TTP*, vol. 24, no. 11, pp. 4408–4421, 2015.

[15] W. Chen, K. Gu, T. Zhao, G. Jiang, and P. L. Callet, "Semi-reference sonar image quality assessment based on task and visual perception," *IEEE TMM*, vol. 23, pp. 1008–1020, 2021.

[16] F. Meng, S. Li, and Y. Chang, "No-reference stereoscopic image quality assessment based on the human visual system," in *ICASSP*, pp. 2100–2104, 2021.

[17] Y. Fang, R. Du, Y. Zuo, W. Wen, and L. Li, "Perceptual quality assessment for screen content images by spatial continuity," *IEEE TCSVT*, vol. 30, no. 11, pp. 4050–4063, 2020.

[18] J. Yang, Z. Bian, J. Liu, B. Jiang, W. Lu, X. Gao, and H. Song, "No-reference quality assessment for screen content images using visual edge model and adaboosting neural network," *IEEE TIP*, vol. 30, pp. 6801–6814, 2021.

[19] J. Yang, Y. Zhao, J. Liu, B. Jiang, Q. Meng, W. Lu, and X. Gao, "No reference quality assessment for screen content images using stacked autoencoders in pictorial and textual regions," *IEEE TCYB*, vol. 52, no. 5, pp. 2798–2810, 2022.

[20] G. Yin, W. Wang, Z. Yuan, C. Han, W. Ji, S. Sun, and C. Wang, "Content-variant reference image quality assessment via knowledge distillation," in *AAAI*, 2022.

[21] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *IEEE TIP*, vol. 31, pp. 4149–4161, 2022.

[22] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *ICCV*, pp. 5128–5137, 2021.

[23] V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *CVPR*, pp. 9375–9383, 2019.

[24] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng, and J. Fan, "Adaptive fractional dilated convolution network for image aesthetics assessment," in *CVPR*, pp. 14114–14123, 2020.

[25] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *CVPR*, pp. 14131–14140, 2020.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, pp. 248–255, 2009.

[27] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "Graphiqa: Learning distortion graph representations for blind image quality assessment," *IEEE TMM*, 2022.

[28] Z. Pan, F. Yuan, J. Lei, Y. Fang, X. Shao, and S. Kwong, "Vcrnet: Visual compensation restoration network for no-reference image quality assessment," *IEEE TIP*, vol. 31, pp. 1613–1627, 2022.

[29] Q. Zheng, Y. Fan, L. Huang, T. Zhu, J. Liu, Z. Hao, S. Xing, C.-J. Chen, X. Min, A. C. Bovik, *et al.*, "Video quality assessment: A comprehensive survey," *arXiv preprint arXiv:2412.04508*, 2024.

[30] H. Zeng, H. Huang, J. Hou, J. Cao, Y. Wang, and K.-K. Ma, "Screen content video quality assessment model using hybrid spatiotemporal features," *IEEE TIP*, vol. 31, pp. 6175–6187, 2022.

[31] S. Cheng, H. Zeng, J. Chen, J. Hou, J. Zhu, and K.-K. Ma, "Screen content video quality assessment: Subjective and objective study," *IEEE TIP*, vol. 29, pp. 8636–8651, 2020.

[32] P. Cheraaqee, Z. Maviz, A. Mansouri, and A. Mahmoudi-Aznaveh, "Quality assessment of screen content images in wavelet domain," *IEEE TCSVT*, vol. 32, no. 2, pp. 566–578, 2022.

[33] Y. Fang, J. Yan, L. Li, J. Wu, and W. Lin, "No reference quality assessment for screen content images with both local and global feature representation," *IEEE TIP*, vol. 27, no. 4, pp. 1600–1610, 2018.

[34] L. Zheng, L. Shen, J. Chen, P. An, and J. Luo, "No-reference quality assessment for screen content images based on hybrid region features fusion," *IEEE TMM*, vol. 21, no. 8, pp. 2057–2070, 2019.

[35] C. Chen, H. Zhao, C. Peng, T. Yu, H. Yang, and H. Qin, "Full reference screen content image quality assessment by fusing multi-level structure similarity," *ACM TMCCA*, vol. 17, no. 94, pp. 1–21, 2021.

[36] C. Zhang, Z. Huang, S. Liu, and J. Xiao, "Dual-channel multi-task cnn for no-reference screen content image quality assessment," *IEEE TCSVT*, vol. 32, no. 8, pp. 5011–5025, 2022.

[37] X. Min, K. Gu, G. Zhai, X. Yang, W. Zhang, P. Le Callet, and C. W. Chen, "Screen content quality assessment: Overview, benchmark, and beyond," *ACM CSUR*, vol. 54, no. 9, pp. 1–36, 2021.

[38] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE TTP*, vol. 15, no. 11, pp. 3440–3451, 2006.

[39] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 11, pp. 1–21, 2010.

[40] A. Kaipio, M. Ponomarenko, and K. Egiazarian, "Merging of mos of large image databases for no-reference image visual quality assessment," in *MMSP*, pp. 1–6, 2020.

[41] A. Mikhailiuk, M. Pérez-Ortiz, D. Yue, W. Suen, and R. K. Mantiuk, "Consolidated dataset and metrics for high-dynamic-range image quality," *IEEE TMM*, vol. 24, pp. 2125–2138, 2022.

[42] F. Gao, D. Tao, X. Gao, and X. Li, "Learning to rank for blind image quality assessment," *IEEE TNNLS*, vol. 26, no. 10, pp. 2275–2290, 2015.

[43] W. Chen, T.-y. Liu, Y. Lan, Z.-m. Ma, and H. Li, "Ranking measures and loss functions in learning to rank," in *NIPS*, pp. 315–323, 2009.

[44] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE TIP*, vol. 26, no. 8, pp. 3951–3964, 2017.

[45] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, "Continual learning for blind image quality assessment," *IEEE TPAMI*, pp. 1–1, 2022.

[46] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Task-specific normalization for continual learning of blind image quality models," *IEEE TIP*, 2024.

[47] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1238–1257, 2021.

[48] M. Song, W. Song, G. Yang, and C. Chen, "Improving rgb-d salient object detection via modality-aware decoder," *IEEE Transactions on Image Processing*, vol. 31, pp. 6124–6138, 2022.

[49] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *CVPR*, pp. 1580–1589, 2020.

[50] M.-F. Tsai, T.-Y. Liu, T. Qin, H.-H. Chen, and W.-Y. Ma, "Frank: A ranking method with fidelity loss," in *SIGIR*, pp. 383–390, 2007.

[51] N. Ponomarenko, L. Jin, O. Ieremeiev, and et al, "Image database tid2013: Peculiarities, results and perspectives," *IEEE TMM*, vol. 30, pp. 57–77, 2015.

[52] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE TIP*, vol. 25, no. 1, pp. 372–387, 2016.

[53] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE TIP*, vol. 29, pp. 4041–4056, 2020.

[54] Z. Ni, L. Ma, H. Zeng, Y. Fu, L. Xing, and K.-K. Ma, "Scid: A database for screen content images quality assessment," in *ISPACS*, pp. 774–779, 2017.

[55] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *WACV*, pp. 3989–3999, 2022.

[56] J. You and J. Korhonen, "Transformer for image quality assessment," in *ICIP*, pp. 1389–1393, 2021.

[57] Z. Ying, H. Niu, P. Gupta, D. Mahajan, and D. Ghadiyaram, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *CVPR*, 2020.

[58] 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 *CVPR*, pp. 3664–3673, 2020.

[59] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *WACV*, pp. 3989–3999, 2022.

[60] A. Saha, S. Mishra, and A. C. Bovik, "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *CVPR*, pp. 5846–5855, 2023.

[61] J. Shi, P. Gao, and A. Smolic, "Blind image quality assessment via transformer predicted error map and perceptual quality token," *IEEE TMM*, pp. 1–11, 2023.

[62] J. Xie, Y. Luo, J. Ling, and G. Yue, "No reference image quality assessment via quality difference learning," in *ICME*, pp. 1301–1306, 2023.

[63] Q. Huang, B. Fang, X. Ai, and T. Nie, "Perceiving multi-layer representations for no-reference image quality assessment," in *ICASSP*, pp. 3945–3949, 2024.

[64] A. Shukla, A. Upadhyay, S. Bhugra, and M. Sharma, "Opinion unaware image quality assessment via adversarial convolutional variational autoencoder," in *WACV*, pp. 2153–2163, January 2024.

[65] L. Agnolucci, L. Galteri, M. Bertini, and A. Del Bimbo, "Arniqa: Learning distortion manifold for image quality assessment," in *WACV*, pp. 189–198, 2024.

[66] J. Chen, L. Shen, L. Zheng, and X. Jiang, "Naturalization module in neural networks for screen content image quality assessment," *IEEE SPL*, vol. 25, no. 11, pp. 1685–1689, 2018.

[67] J. Yang, J. Liu, B. Jiang, and W. Lu, "No reference quality evaluation for screen content images considering texture feature based on sparse representation," *Signal Processing*, vol. 153, pp. 336–347, 2018.
[68] X. Jiang, L. Shen, G. Feng, L. Yu, and P. An, "Deep optimization model for screen content image quality assessment using neural networks," 2019. arXiv:1903.00705.
[69] Y. Bai, M. Yu, Q. Jiang, and Z. Jiang, Gangyi Zhu, "Learning content-specific codebooks for blind quality assessment of screen content images," *Signal Processing*, vol. 161, pp. 248–258, 2019.
[70] R. Gao, Z. Huang, and S. Liu, "Multi-task deep learning for no-reference screen content image quality assessment," in *MMM*, pp. 213–226, 2021.
[71] Y. Bai, Z. Zhu, G. Jiang, and H. Sun, "Blind quality assessment of screen content images via macro-micro modeling of tensor domain dictionary," *IEEE TMM*, vol. 23, pp. 4259–4271, 2021.
[72] T. Tang, C. You, and R. Zhang, "Efficient but effective perceptual quality model of screen content image," in *APSCON*, 2023.
[73] Z. Xu, Y. Yang, Z. Zhang, and W. Zhang, "No reference quality assessment for screen content images based on entire and high-influence regions," in *ICASSP*, pp. 1–5, 2023.
[74] S. Saini, A. Saha, and A. C. Bovik, "Hidro-vqa: High dynamic range oracle for video quality assessment," in *WACV*, pp. 469–479, January 2024.
[75] S. Mitra and R. Soundararajan, "Knowledge guided semi-supervised learning for quality assessment of user generated videos," in *AAAI*, vol. 38, pp. 4251–4260, 2024.

Mengke Song is currently pursuing a Ph.D. in the College of Computer Science and Technology, and and Qingdao Institute of Software at China University of Petroleum (East China). His research interests is computer vision and deep learning,

Chenglizhao Chen is a Professor in the College of Computer Science and Technology, China University of Petroleum (East China). He received his Ph.D. degree from Beihang University in 2017. His research interests include virtual reality, computer vision, deep learning, and pattern recognition.

Wenfeng Song is an Associate Professor in the College of Computer Science and Technology, Beijing Information Science and Technology University. Her research interests include computer graphics, virtual reality, AI+ medicine, large models, and human-computer interaction.

Yuming Fang is a Professor in the School of Information Management, Jiangxi University of Finance and Economics. His research interests include visual big data, video image processing, computer vision, machine learning, intelligent information processing, and information mining.