Subjective and Objective Quality Assessment of Colonoscopy Videos

Guanghui Yue, *Member, IEEE*, Lixin Zhang, Jingfeng Du, Tianwei Zhou, Wei Zhou, *Senior Member, IEEE*, and Weisi Lin, *Fellow, IEEE*

***Abstract*— Captured colonoscopy videos usually suffer from multiple real-world distortions, such as motion blur, low brightness, abnormal exposure, and object occlusion, which impede visual interpretation. However, existing works mainly investigate the impacts of synthesized distortions, which differ from real-world distortions greatly. This research aims to carry out an in-depth study for colonoscopy Video Quality Assessment (VQA). In this study, we advance this topic by establishing both subjective and objective solutions. Firstly, we collect 1,000 colonoscopy videos with typical visual quality degradation conditions in practice and construct a multi-attribute VQA database. The quality of each video is annotated by subjective experiments from five distortion attributes (i.e., temporal-spatial visibility, brightness, specular reflection, stability, and utility), as well as an overall perspective. Secondly, we propose a Distortion Attribute Reasoning Network (DARNet) for automatic VQA. DARNet includes two streams to extract features related to spatial and temporal distortions, respectively. It adaptively aggregates the attribute-related features through a multi-attribute association module to predict the quality score of each distortion attribute. Motivated by the observation that the rating behaviors for all attributes are different, a behavior guided reasoning module is further used to fuse the attribute-aware features, resulting in the overall quality. Experimental results on the constructed database show that our DARNet correlates well with subjective ratings and is superior nine state-of-the-art methods.**

***Index Terms*— Colonoscopy, video quality assessment, subjective and objective quality assessment, deep neural network**

This work was supported in part by the National Natural Science Foundation of China (No. 62371305, No. 62001302 and No. 62103286), in part by Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515030025 and No. 2024A1515030278), in part by Natural Science Foundation of Shenzhen (No. JCYJ20230808105906013 and No. KJZD20230923114615031), in part by Medicine Plus Program of Shenzhen University (No. 2024YG007) and in part by Tencent "Rhinoceros Birds"-Scientific Research Foundation for Young Teachers of Shenzhen University. (Corresponding author: Tianwei Zhou)

G. Yue and L. Zhang are with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen 518054, China, and also with the Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: yueguanghui@szu.edu.cn; cheunglaihip@163.com).

J. Du is with the Department of Gastroenterology and Hepatology, Shenzhen University General Hospital, Shenzhen University, Shenzhen 518060, China (e-mail: djfjms1231@qq.com).

T. Zhou is with the College of Management, Shenzhen University, Shenzhen 518060, China (e-mail: tianwei@szu.edu.cn).

W. Zhou is with the School of Computer Science and Informatics, Cardiff University, UK (e-mail: zhouw26@cardiff.ac.uk).

W. Lin is with School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore (wslin@ntu.edu.sg).

I. Introduction

COLORECTAL cancer ranks as the third most lethal cancer worldwide, posing a significant threat to millions and exerting considerable pressure on healthcare systems [1]–[3]. Colonoscopy allows physicians to directly inspect the characteristics of lesions and remove suspected lesions at an early stage to prevent further deterioration into tumors [4]–[7]. However, owing to the specialized imaging environment, the captured colonoscopy videos usually suffer from multiple distortions, such as motion blur, low brightness, abnormal exposure, and object occlusion [8], [9]. These distortions affect the visual interpretation, potentially increasing the missed diagnosis rate. Thus, effective and reliable Video Quality Assessment (VQA) metrics are highly required to provide quantitative quality feedback, based on which the physicians can adjust their operations to obtain high-quality videos for better diagnosis.

Generally, quality assessment methods can be broadly classified into subjective and objective methods [10]–[12]. The subjective method evaluates video quality through subjective experiments with predefined scoring rules, and is usually used for database construction. A detailed methodology of subjective experiments is presented in Section III-B.2, including task introduction, participant training, and subjective rating. In the literature, the subjective method has been widely adopted to investigate the perceptual quality of videos for entertainment purposes, e.g., in-the-wild videos [13] and user-generated content videos [14], [15]. However, related works on medical images/videos are notably lacking. Among the few attempts, researchers primarily focus on ultrasound images/videos [16], dermoscopy images [17], and retinal images [18], while rarely involving colonoscopy videos. To promote the development of colonoscopy VQA, one simple idea is treating this problem as a distortion classification task [19]. However, this idea cannot inform observers of how much a video has strayed from perfection. Another idea is evaluating the quality of synthetically distorted videos [20]. Unfortunately, it impedes our comprehensive understanding of authentically distorted videos as synthesized distortions differ from authentic distortions greatly. Thus, it is necessary to conduct subjective experiments to detail the factors affecting the quality assessment of authentically distorted colonoscopy videos and investigate the impacting strength of each factor.

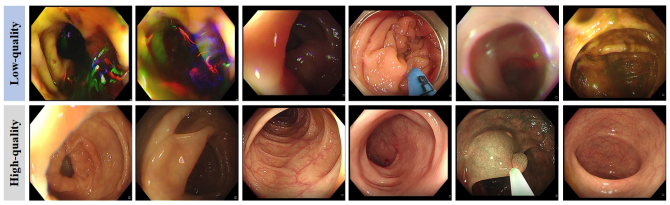

Fig. 1. Illustration of distortion attributes considered in MAC-VQA: the consecutive frames selected from different videos in the first two columns are used to illustrate the distortion attribute of temporal-spatial visibility, and the frames in the last four columns are used to illustrate the distortion attributes of brightness, specular reflection, stability, and utility, from left to right, respectively. Here, "utility" refers to the cleanliness of the colon surface.

During the past years, many objective VQA methods have been reported for natural scene videos [21], [22], in which distortions are usually characterized by compression, blurry-motion, exposure, flicker, etc. To quantify distortions, early researchers primarily decomposed the video into frames and analyzed each frame through handcrafted features by modeling the properties of the human visual system or the rules of natural scene statistics [23]. However, the use of handcrafted features highly requires prior experience to distortion patterns, making the designed VQA methods perform poorly when handling videos with complex distortions. Recently, the rapid development of Deep Neural Networks (DNNs) has brought new solutions for VQA. DNN automatically learns the feature representations of distortions, avoiding the subjectivity and limitations of manually selecting features. Among many attempts, Convolutional Neural Network (CNN) based VQA methods are popular [24]–[26]. More recently, researches [27]–[29] have explored the use of Transformers, which are capable of efficiently modeling long-range dependencies, offering a promising avenue for improving VQA accuracy. Despite this, very few related works involve in the colonoscopy VQA task. Different from natural scene videos, colonoscopy videos need to be evaluated from multiple attributes, e.g., temporal-spatial visibility, brightness, specular reflection, stability, and utility, to ensure their usage in disease diagnosis. As shown in Fig. 1, many distortions affect the quality of colonoscopy videos, such as poor temporal-spatial visibility, low brightness, flicker, and specular reflection, making the VQA task very challenging. Current natural scene video-oriented VQA methods are not suitable for the colonoscopy VQA task as they can only predict the overall quality. Thus, it is necessary to design specific quality assessment methods for colonoscopy videos.

In this study, we comprehensively study the colonoscopy VQA task from both subjective and objective perspectives. Specifically, we collect 1,000 authentically distorted colonoscopy videos and conduct subject experiments to evaluate them from the overall quality as well as five distortion attributes, i.e., temporal-spatial visibility, brightness, specular reflection, stability, and utility. Overall, we collect 90,000 continuous-scale human opinions and construct a Multi-Attribute Colonoscopy VQA (MAC-VQA) database. From the subjective opinions, we find that these distortion attributes have inherent relationships in the VQA task and correlate differently with the overall quality. Based on these findings, we propose a Distortion Attribute Reasoning Network (DARNet), which can not only objectively estimate the overall quality score but also derive the quality scores of distortion attributes. Experimental results on the MAC-VQA database show that our DARNet is competent for the colonoscopy VQA task.

Compared with prior research, our study has the following technical advancements. First, unlike existing works that only evaluate the video quality from a single perspective [24]–[26], our study proposes a comprehensive VQA framework capable of assessing the video quality from multiple perspectives. Such an advantage makes it more suitable for the colonoscopy VQA task. Second, different from prior studies [30], [31] that directly map the features extracted from video frames and sequences to a quality score, our study incorporates a more refined procedure via a well-designed Feature Interaction and Reasoning (FIR) unit to explore and integrate attribute-aware features. This enhancement allows our network to more effectively recognize video distortions. Lastly, motivated by the observation that distortion attributes have inherent relationships in the VQA task, we propose to adaptively mine complementary information within different attribute features using a Multi-Attribute Association (MAA) module. Compared to traditional self-attention block [32], our MAA module can interact with different features, contributing to extracting more discriminative feature representations.

In summary, this study makes the following contributions.

- We construct the first comprehensive colonoscopy VQA database that provides video quality annotations from five distortion attributes (i.e., temporal-spatial visibility, brightness, specular reflection, stability, and utility) as well as an overall perspective. The database will be released at `https://github.com/cheunglaihip/DARNet`.
- We provide a detailed discussion of the constructed database, emphasizing the crucial aspects to consider in the design of VQA methods. Additionally, based on the constructed database, we explore the effectiveness of nine state-of-the-art VQA methods on the colonoscopy VQA task and provide a comprehensive summary of these benchmarks.
- We propose a novel VQA network, termed DARNet. It not only adaptively aggregates the attribute-aware features to predict the quality scores of distortion attributes but also takes into account the subjective rating behaviors to derive the overall quality score. Extensive experiments on the constructed database demonstrate that our method outperforms the state-of-the-art methods.

II. Related Works

A. Video Quality Assessment Databases

In recent decades, the literature has reported numerous VQA databases for natural scene videos based on subjective experiments. Early researches mainly focus on videos with synthetical distortions in consideration of various scenarios, such as video storage and transmission [33]. More recently, scholars have gradually realized the importance of medical image and video quality, which is a prerequisite for achieving

a high detection rate of diseases. For example, Yue *et al*. [9] measured the quality of enhanced colonoscopy images generated by mainstream low-light image enhancement algorithms based on subjective experiments. Lévêque *et al*. [20] evaluated the quality of distorted videos with different transmission errors and synthesized compression artifacts, constructing a telesurgery VQA database. Münzer *et al*. [34] conducted subjective experiments to explore the effects of compression artifacts on both the technical and semantic quality of laparoscopic videos. Usman *et al*. [35] investigated the impact of high efficiency video coder on Wireless Capsule Endoscopy (WCE) videos, and constructed a WCE VQA dataset. Overall, existing works mainly focus on synthetically distorted videos. However, the synthetic distortions differ from authentic distortions greatly, limiting our comprehensive understanding of colonoscopy videos in the clinical environment. Therefore, it becomes critical to construct authentically distorted databases for advancing the colonoscopy VQA task.

B. Video Quality Assessment Methods

In the early stage, researchers usually divide the video into many frames and evaluate the quality of each frame through popular Image Quality Assessment (IQA) methods. The video quality is estimated by integrating the quality scores of all frames [36]. However, these methods usually perform poorly in complex scenarios as they fail to fully characterize temporal distortions in videos. To address this problem, Saad *et al*. [37] proposed a new VQA method, termed V-BLIINDS, which measures frame-to-frame differences and motion-related distortions using a spatio-temporal natural scene statistics model. Considering that the 3D-DCT domain has the inherent advantage over other 2D transformations in representing spatio-temporal information, Li *et al*. [38] integrated the statistical spatio-temporal features in 3D-DCT domain, resulting in a new VQA method. Korhonen *et al*. [39] proposed an efficient VQA method that only computes high-complexity features on representative frames. Tu *et al*. [40] selected 60 statistical features used in existing VQA methods and fused them using a support vector regressor to estimate the video quality. However, such methods rely heavily on the design and selection of handcrafted features, which may lead to low generalization ability due to knowledge bias.

During recent years, the rapid advancement of DNNs has introduced new solutions to VQA. DNN-based methods automatically learn the quality-aware features by updating network parameters, overcoming the subjectivity and limitations of handcrafted feature selection. Li *et al*. [25] divided the video into frames and fed them into pre-trained CNNs to extract features. After that, a gated recurrent unit and a temporal pooling layer were used to integrate these frame-level features for estimating video quality. Given that merely analyzing video frames is difficult to fully characterize distortions, Sun *et al*. [30] proposed a new VQA network that includes two branches designed specifically for extracting spatial and motion-related features. To reduce computational complexity, only sparse video frames are fed into the upper branch and down-sample video sequences were fed into the lower branch. Li *et al*. [31] proposed a VQA network by transferring the knowledge from two types of source domain, corresponding to spatial and motion distortions in the video, respectively. Wu *et al*. [41] employed a Transformer-based backbone to characterize video distortions and introduced a fast VQA network by using a grid mini-patch sampling strategy. However, existing VQA methods only evaluate the video quality from an overall perspective, and cannot fully match the needs of colonoscopy VQA tasks, i.e., providing both the overall quality score and quality scores for all distortion attributes.

III. Multi-Attribute Colonoscopy VQA Database

A. Video Collection

In this study, we used an Olympus CV290 endoscope to capture 40 long colonoscopy videos of at Shenzhen University General Hospital. These videos were collected from 40 patients, including 24 males and 16 females. The patients range in age from 20 to 69, with a slight asymmetry in each age group. Fig. 2 presents the gender and age distributions of patients involved. The captured videos encompass various scenarios of colon anatomy, such as the rectum, descending colon, ascending colon, cecum, and terminal ileum, along with surgical procedures to remove abnormal tissues. This ensures the diversity of our data, with strong clinical representativeness. Each video has a duration from 15 to 30 minutes, with a frame rate of 24 and resolution of 1920×1080. The main mode of these videos is white light imaging, with a small portion of segments being narrow band imaging due to clinical diagnostic needs. Second, for privacy protection, we remove the patient's information shown on the video, such as name, age, gender, treatment time, etc. Third, we cut the long videos and only preserve video clips that are relevant to disease diagnosis. Approximately 20 to 40 video clips are collected from each patient, without any content repetition. As a result, we collect a total of 1,000 short videos, each with a length of 9 to 11 seconds. Fig. 1 presents some examples.

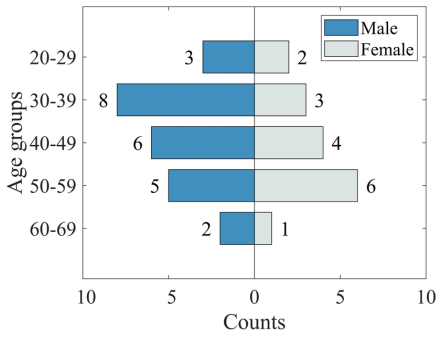

Fig. 2. Gender and age distributions of patients involved in video collection.

B. Subjective Quality Assessment

1) Choice of Quality-Related Attributes: Generally, video quality is often affected by multiple attributes [42], such as temporal-spatial visibility, brightness, and stability. A recent study [43] stated that, apart from these common attributes,

the quality of colonoscopy videos is also affected by specular reflection because of the moist colon surface. Moreover, as shown in Fig. 1, a colon surface is usually occluded by feces and bubbles, potentially leading to diagnosis errors. In view of these, we propose to comprehensively analyze the quality of colonoscopy videos, not only providing the quality scores for five distortion attributes, i.e., temporal-spatial visibility, brightness, specular reflection, stability, and utility, but also reporting the overall quality score. The descriptions of these distortion attributes are given below.

- **Temporal-Spatial Visibility**: This attribute reflects the amount of content information along both the spatial and temporal dimensions. A higher value of temporal-spatial visibility corresponds to more information.
- **Brightness**: Due to uneven lighting and elongated characteristics of intestines, the captured videos are usually with abnormal brightness. Excessively high or low brightness results in poor quality of colonoscopy videos.
- **Specular Reflection**: This attribute affects the quality of colonoscopy videos by the location and area of the specular reflection.
- **Stability**: In colonoscopy, the relative motion between camera and intestine brings blurriness and information loss of the captured videos. The stability reflects the degree of relative motion.
- **Utility**: Colonic contents, such as fluid, bubbles, and feces, occlude the colon surface and affect the video's utility for disease diagnosis. The clearer the colon surface is, the greater the utility of the video.

Notably, the overall quality score is not directly computed from the quality scores of five distortion attributes using predefined formulas during the rating stage. The reason can be explained from three aspects. First, there is currently a lack of systematic research and clinical evidence to clarify how these influencing attributes interact to affect the overall quality of the video. Second, according to the feedback of gastroenterologists, partial attributes have indescribable correlation and may result in a masking effect for the overall quality when appearing simultaneously. For instance, high stability is a necessary condition for high utility, but not a sufficient condition. Third, the distortions usually do not exist throughout the entire video sequence but only in video segments, and the lengths of video segments occupied by different distortions are not the same. Therefore, it is hard to set a fixed weight for each distortion in computing the overall quality scores for different videos.

2) Subjective Ratings: We recruit 15 postgraduates majored in biomedical engineering (20 to 30 years old) to complete the rating task. All subjects have normal or correct-to-normal vision and sign the written consent form. This study is approved by the Medical Ethical Committee of the Shenzhen University Health Science Center (Number: PN-202300016). The subjective experiment is conducted in a lab with environment similar to doctor's office. We set a flexible viewing distance (i.e., one to three times the video height), allowing subjects to find the most comfortable viewing position.

The experiment consists of training and test stages. In the training stage, we first invite two senior gastroenterologists, who have more than ten years of working experience, to help the subjects master the essential skills for colonoscopy video analysis. Then, the experiment administrator provides detailed descriptions of the study's process and objectives, along with 50 sample videos to illustrate the rating system and rules. All these videos are pre-scored by gastroenterologists. Only subjects with more than 75% scoring accuracy to gastroenterologists on these videos are permitted to participate in the formal experiment. In the test stage, subjects evaluate the quality of randomly played videos. To comprehensively evaluate the video quality, they not only rate the overall quality on a continuous scale from 1 to 5 but also to provide five continuous scores from 0 to 3, representing the degrees of temporal-spatial visibility, brightness, specular reflection, stability, and utility, respectively. Through the setting of continuous scores, we can evaluate the video quality in a more accurate manner. For example, one can give a score of 1.5 if he/she hesitates between 1 and 2. Here, a larger rating score corresponds to better perceived quality. The reason why we utilizing different scoring scales for the overall quality and each distortion attribute lies in two aspects, which highly match the clinical experience and needs of gastroenterologists. Specifically, on the one hand, compared to the quality of distortion attributes, the overall quality is a more comprehensive, nuanced, and important indicator to determine whether a video is qualified for diagnosis. As such, its scoring scale should be wider for more detailed and fine-grained outputs. On the other hand, the quality of distortion attributes is used to provide feedback about which part should be adjust to improve video quality. In this point, a basic scoring scale (e.g., [0-3]) narrower than that of the overall quality is enough for each distortion attribute. Such a setting also reduces the scoring burden. To facilitate the experimental process, we design a rating software using MATLAB 2021b, as shown in Fig. 3. Once the rating task for a video is completed, the next video immediately appears. The rating task includes 5 sessions, and the subjects should evaluate 200 videos in each session. The subjects are encouraged to stop to relieve visual fatigue every 20 minutes during the test. All colonoscopy videos are displayed in a random and non-repeating manner on a 27-inch 1920×1080

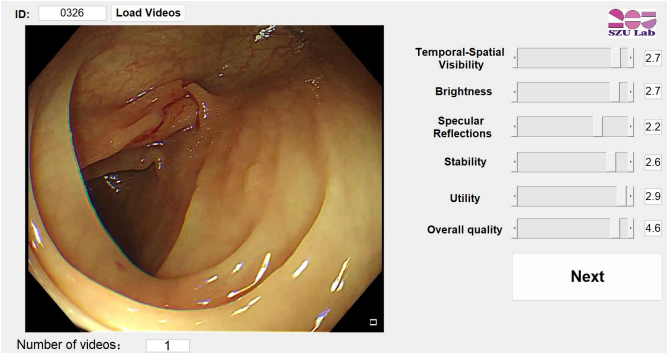

Fig. 3. Graphical user interface of the rating software designed for subjective experiments. The left side of the interface is the video display area, and the right side is the scoring area. Subjects should rate both the overall quality score and quality scores for five distortion attributes. A higher score indicates a lower degree of distortion and better quality.

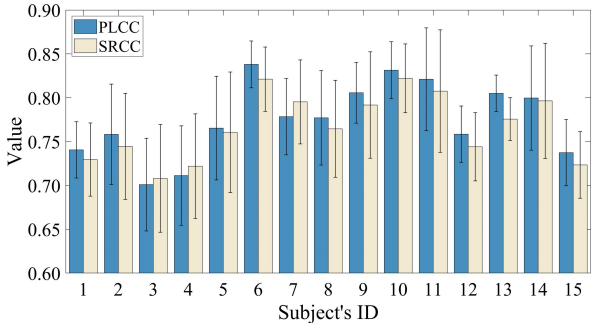

Fig. 4. Illustration of the correlation (in the form of Pearson Linear Correlation Coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC)) between the rated scores of each subject and the average scores of two gastroenterologists. For each correlation metric, we present both mean and standard deviation of values calculated for the overall quality and other five distortion attributes. The descriptions of PLCC and SRCC are given in Section V-A.2.

HP screen. Overall, a total of 90,000(=1000×15×6) quality annotations are collected, where 1000, 15, and 6 denote the video number, subject number, and quality score number per video, respectively.

Furthermore, to illustrate the reliability of the rated quality scores, we also investigate the scoring quality of each subject. To be specific, we randomly select 100 videos and ask the two senior gastroenterologists to rate them from the perspectives of the overall quality as well as five distortion attributes. After that, we calculate the correlation between the rated scores of each subject and the average scores of two gastroenterologists on these videos. As shown in Fig. 4, all subjects have a strong rating correlation with gastroenterologists, with PLCC and SRCC values greater than 0.7. This indicates that our recruited subjects generally produce qualified and reliable rating outcomes.

C. Subjective Data Processing and Analysis

1) Subjective Data Processing: Generally, differences in task understanding often lead to varied subjective ratings. Therefore, it is imperative to purify the gathered subjective rating data. For this purpose, we filter out invalid data using the outlier rejection method recommended by ITU-R BT.500 [44]. For a given video, we first calculate its mean score $\overline{\mu}_k$ and standard deviation s_k across all subjects by

$$\overline{\mu}_k = \frac{1}{M}\sum_{i=1}^{M} q_{i,k}, \tag{1}$$

$$s_k = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(\overline{\mu}_k - q_{i,k})^2}, \tag{2}$$

where M is the number of subjects who participated in the subjective experiment. $q_{i,k}$ is the rating score of the k-th video by the i-th subject. The confidence interval for the k-th video is defined as [$\overline{\mu}_k$-2s_k, $\overline{\mu}_k$+2s_k]. The rating score is considered an outlier if it falls outside the confidence interval. The data should be excluded when a subject's scores considered outliers are more than 5% and $|\frac{(E_i-L_i)}{(E_i+L_i)}|$ is less than 30%. E_i and L_i are the numbers below the lower boundary and above the upper boundary of the confidence interval, respectively, for the i-th subject. After analysis, none of the subjects is rejected among the 15 subjects. Next, we convert the k-th raw-score $q_{n,k}$ by the n-th subject to Z-score $Z_{n,k}$ as follows:

$$Z_{n,k} = \frac{q_{n,k} - \overline{\mu}_n}{s_n}, \tag{3}$$

where $\overline{\mu}_n$ and s_n are the mean and standard deviation across all qualified rating scores of the n-th subject.

Subsequently, we use a linear mapping function to adjust the Z-Score to the designated rating range, which is [0, 3] for distortion attributes and [1, 5] for the overall quality. The Mean Opinion Score (MOS) M_k for the k-th video is finally calculated as the mean value of scaled Z-score $Z'_{n,k}$:

$$M_k = \frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}} Z'_{n,k}, \tag{4}$$

where $\mathcal{N}$ denotes the number of qualified subjects. Through the above process, we have six MOSs for each colonoscopy video, showing its quality from the perspectives of temporal-spatial visibility, brightness, specular reflection, stability, utility, and overall quality, respectively.

2) Scoring Behavior Analysis: Fig. 5 shows the MOS distributions. It is clear that, these distortion attributes are different in MOS distributions. This inspires us to pay different focuses on these distortion attributes when designing objective VQA methods. To see the inherent relationship within these quality influencing factors, we compute the correlation matrix between

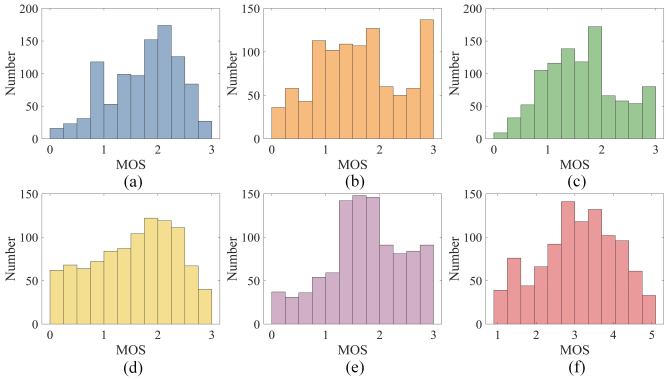

Fig. 5. MOS distributions from the quality dimensions of (a) temporal-spatial visibility, (b) brightness, (c) specular reflection, (d) stability, (e) utility, and (f) overall quality.

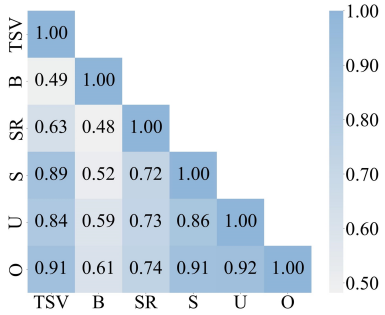

Fig. 6. The correlation matrix (in the form of PLCC) between different quality dimensions rated by all subjects. TSV, B, SR, S, U, and O denote temporal-spatial visibility, brightness, specular reflection, stability, utility, and overall quality, respectively.

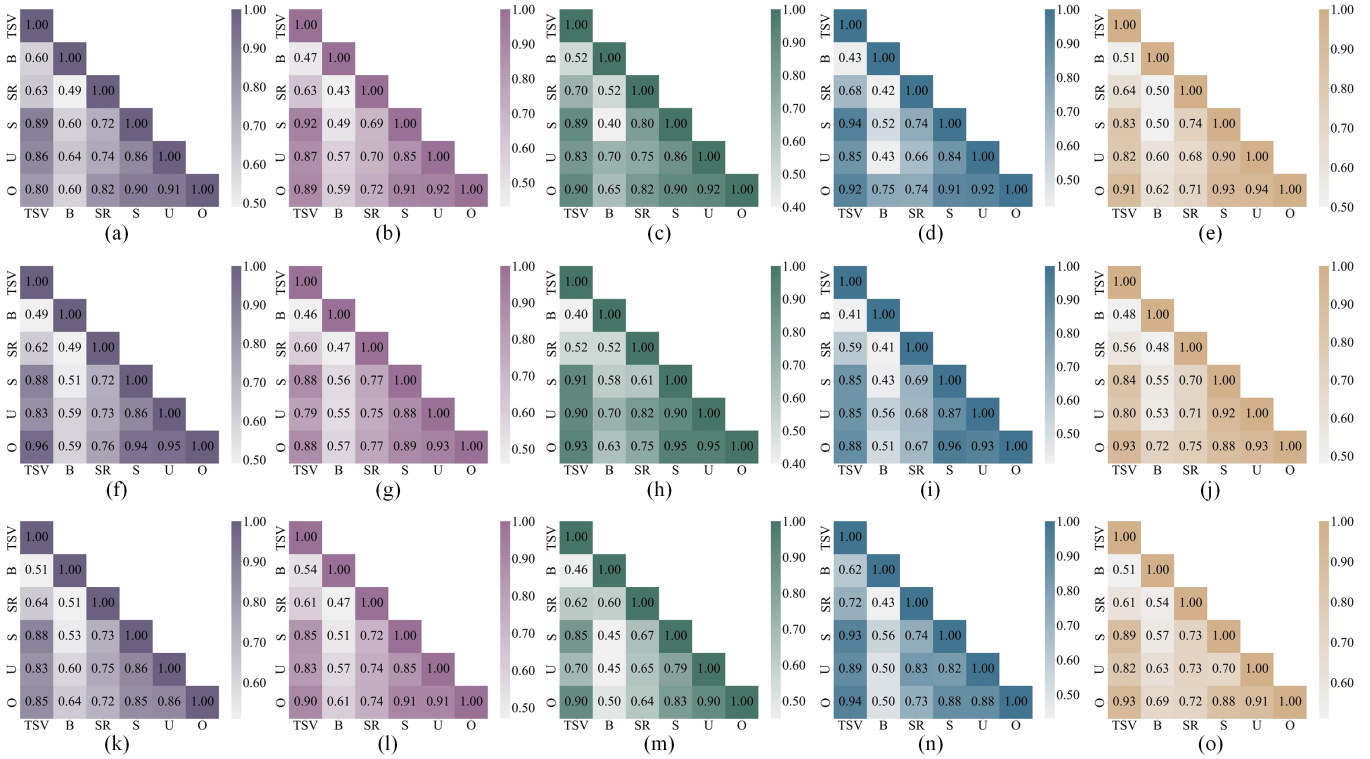

Fig. 7. The correlation matrix (in the form of PLCC) between different quality dimensions rated by each subject. TSV, B, SR, S, U, and O denote temporal-spatial visibility, brightness, specular reflection, stability, utility, and overall quality, respectively.

the MOSs of every two quality dimensions, as shown in Fig. 6. Through the correlation matrix, we have several important findings. First, all distortion attributes relate to the overall quality, with the PLCC value between each attribute and overall quality larger than 0.60. Second, the PLCC values between these distortion attributes and the overall quality are obviously different, indicating that the rating behaviors of subjects vary among different attributes. Third, some distortion attributes, e.g., stability and utility, are closely related. This is because some colonoscopy scenarios, e.g., quick camera movement for lesion search, affect the perceived quality in these two dimensions simultaneously. Such observations can also be found in the correlation matrix calculated from the data of each subjects, as shown in Fig. 7. From the data in Fig. 7, we can also find that the quality annotations are consistent across different subjects.

IV. Proposed Method

A. Motivation

In this study, we propose a novel quality assessment method for colonoscopy videos, named DARNet. It comprehensively evaluates the video quality from five distortion attributes (i.e., temporal-spatial visibility, brightness, specular reflection, stability, and utility) as well as an overall perspective. The technical motivations behind our DARNet are three-fold. 1) Considering that the overall quality is the compounded result of multiple distortion attributes, a multi-task learning framework should be used to simultaneously output the overall quality score and the quality scores of five distortion attributes. Auxiliary tasks, i.e., quality prediction of five distortion attributes, can help improve the performance of the main task, i.e., overall quality prediction. 2) According to the results of subjective experiments, these distortion attributes have certain correlations in the VQA task. In this case, it would be better to interact with the features of different auxiliary tasks to achieve accurate quality predictions. 3) These distortion attributes correlate differently with the overall quality, and the rating behaviors for them vary greatly. Mimicking the rating behaviors is an important clue for accurately predicting the overall quality.

B. Overall Architecture

Fig. 8 shows the framework of DARNet. As shown, DARNet consists of a Spatio-Temporal Feature Extraction (STFE) unit and an FIR unit. The FIR unit comprises an MAA module and a Behavior Guided Reasoning (BGR) module. For an input video, we first feed it into the STFE unit to obtain a shared feature P_α. Next, P_α is processed by Non-Linear Mapping (NLM) operations separately to generate five sets of features $\mathcal{P}_i$ ($i \in \{1, 2, \cdots, 5\}$). These features are subsequently fed into the MAA module to obtain attribute-aware features $\mathcal{S}_i$. By inputting $\mathcal{S}_i$ into the regression head, we can obtain the score Y_i of the i-th quality dimension. Finally, these attribute-aware features are further interactively fused in the BGR module to reason the overall quality score Y_o.

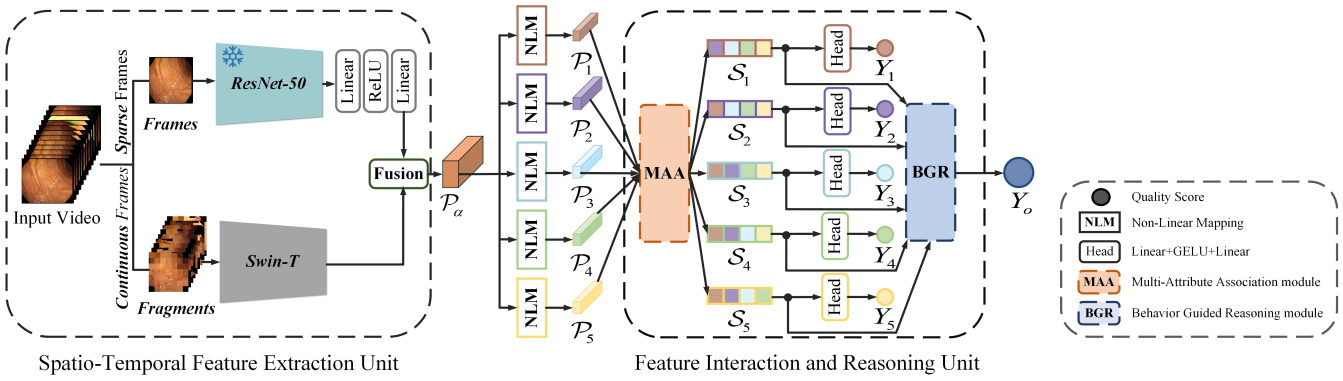

Fig. 8. The framework of DARNet consists of a Spatio-Temporal Feature Extraction (STFE) unit and a Feature Interaction and Reasoning (FIR) unit. The former comprises two streams to extract features for characterizing spatial and temporal distortions. The latter includes a Multi-Attribute Association (MAA) module and a Behavior Guided Reasoning (BGR) module. MAA adaptively aggregates the extracted features $\boldsymbol{\mathcal{P}_i}$ and predicts the quality score $\boldsymbol{Y_i}$ of each attribute. BGR considers the feedback of subjects on the correlation of each distortion attribute to perform the interaction between the attribute-aware features $\boldsymbol{\mathcal{S}_i}$ and derive the overall quality score $\boldsymbol{Y_o}$. To reduce computational burden, we feed 16 frames and 49 fragments into ResNet-50 and Swin-T, respectively. The fragments are spatially stitched to an integrated before feeding into Swin-T.

C. Spatio-Temporal Feature Extraction Unit

As discussed in Section III-B, the colonoscopy videos have both spatial and temporal distortions. In view of these, we use a STFE unit to extract both spatial and temporal features to characterize these distortions.

As illustrated in Fig. 8, the STFE unit includes two parallel branches (i.e., ResNet-50 and Swin-T). The upper branch, i.e., ResNet-50 [45], is used to extract spatial features $\mathcal{F}_s \in \mathbb{R}^{2048\times H\times W}$ from video frames, where H and W are the frame's height and weight, respectively. We split the video into consecutive frames and resize each frame to 224×224. Considering the duration difference between videos, we randomly select T frames from each video and input them into ResNet-50, respectively. Following previous works [31], ResNet-50 is pre-trained on IQA databases and keeps frozen during the training stage. The lower branch, i.e., Swin-T (the abbreviation of Video Swin Transformer Tiny [46]), is used to acquire spatio-temporal features $\mathcal{F}_t \in \mathbb{R}^{T\times768\times H\times W}$ from video sequences. Compared to frames, video feature extraction requires more computational resources because of high dimensionality. To mitigate computational burden, we change the input of Swin-T from video sequences to "fragments", which are generated using the sampling strategy reported in FASTER-VQA [41]. To be specific, for a video, we first cut it into 7×7 non-overlapping grids in the spatial dimension, and randomly crop a 32×32 mini-patch from each spatial grid. Meanwhile, we cut the video into 8 uniform segments in the temporal space, and select 4 consecutive frames from each segment. Notably, to preserve temporal continuity between frames, mini-patches in each spatial grid and temporal segment are aligned to form a mini-cube. The mini-cube is named as "fragment". Finally, we stitch all the mini-cubes spatially to an integrated sample. Here, T is set to 16.

To fuse the features from two branches, we apply two operations on the upper branch. First, we concatenate the features of T selected frames, resulting in $\mathcal{F}'_s \in \mathbb{R}^{T\times2048\times H\times W}$. Second, we process $\mathcal{F}'_s$ with a mapping block, which consists of two linear layers and one ReLU function within them, to map its size to that of $\mathcal{F}_t$. After that, we integrate the resultant feature $\mathcal{F}''_s \in \mathbb{R}^{T\times768\times H\times W}$ and $\mathcal{F}_t$ using a fusion module, resulting in $\mathcal{P}_\alpha \in \mathbb{R}^{N\times C}$. The fusion module has a Fully Connected (FC) layer, a GELU function, and another FC layer. Here, $N = T\times H\times W$. $\mathcal{P}_\alpha$ is further processed by a NLM operation to generate five sets of features $\mathcal{P}_i$ for subsequent quality prediction of multiple distortion attributes:

$$\mathcal{P}_i = \mathcal{M}_{\theta_i}(\mathcal{P}_\alpha), \tag{5}$$

where θ_i denotes the parameters of the NLM operation $\mathcal{M}_{\theta_i}$ for the i-th distortion attribute.

D. Feature Interaction and Reasoning Unit

Motivated by the observations from subjective experiments, we propose a FIR unit to produce accurate quality predictions by making inter-attribute interactions. As shown in Fig. 8, the proposed FIR unit includes an MAA module and a BGR module.

1) Multi-Attribute Association Module: As illustrated by Fig. 6, every two distortion attributes have an inherent relationship to the quality assessment task, with a correlation greater than 0.45. This inspires us to interact with the features of different distortion attributes. For this purpose, we propose an MAA module. It explores complementary information from different features to achieve accurate quality predictions for multiple distortion attributes simultaneously based on the attention mechanism. As illustrated by Fig. 9, the MAA module takes the features $\mathcal{P}_i$ ($i \in \{1, 2, \cdots, 5\}$) as the inputs and outputs the attribute-aware features $\mathcal{S}_i$. Specifically, it has five parallel cross-attention (CA) blocks. Taking the n-th CA block as an example, the feature $\mathcal{P}_n$ is mapped into the *Query* $Q_n \in \mathbb{R}^{N\times C}$ via a linear mapping layer $\mathcal{W}_q^n$. Also, the i-th feature $\mathcal{P}_i$ ($i \neq n$) is mapped into the *Key* $K_{n,i} \in \mathbb{R}^{N\times C}$ and *Value* $V_{n,i} \in \mathbb{R}^{N\times C}$ via linear mapping layers $\mathcal{W}_k^n$ and $\mathcal{W}_v^n$, respectively. The learned positional encoding [32] is added to the P_n and P_i before processing them with the linear mapping layer. The output $\mathcal{O}_{n,i}$ of the attention block is:

$$\mathcal{O}_{n,i} = \sigma(Q_n K_{n,i}^{\mathcal{T}}/\sqrt{d})V_{n,i}, \tag{6}$$

where $\sigma(\cdot)$ is the Softmax function, and $d = C$ is the feature dimension. $^{\mathcal{T}}$ denotes the transpose operation. By respectively using other features to generate the *Key* and *Value*, we can obtain four features from the attention block in total. After that, we add each output with $\mathcal{P}'_n$ and process the resultant feature by Layer Normalization (LN) and Multi-Layer Perception (MLP):

$$\mathcal{F}_{n,i} = MLP(LN(\mathcal{P}'_n + \mathcal{O}_{n,i})), \tag{7}$$

where $\mathcal{P}'_n$ is the addition of $\mathcal{P}_n$ and positional encoding. Next, all $\mathcal{F}_{n,i}$ are concatenated in the channel direction to generate the n-th attribute-aware feature $\mathcal{S}_n \in \mathbb{R}^{N \times C'}$. Finally, we process $\mathcal{S}_n$ with a regression head, which consists of a linear layer, a GELU function, and another linear layer, to generate the quality score Y_n of the n-th distortion attribute. Different from traditional self-attention module [32] that derives the *Query*, *Key*, and *Value* vectors from the same feature, these vectors in our MAA module are derived from the features of two different streams (see Fig. 8), specifically designed for predicting quality scores of distortion attributes. Moreover, our MAA module can integrate multiple features generated by the attention operation, whereas traditional self-attention module cannot. Such settings make our MAA module more suitable for the colonoscopy task.

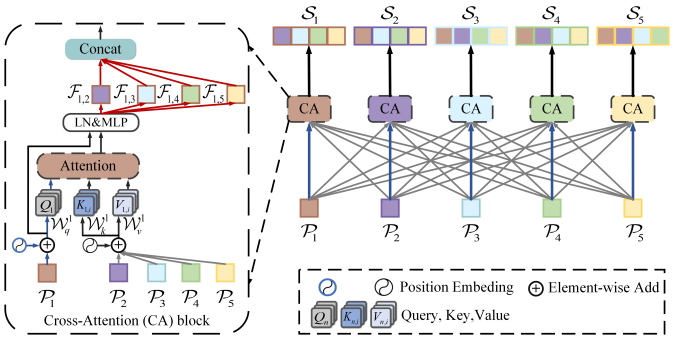

Fig. 9. Framework of the Multi-Attribute Association (MAA) Module.

2) Behavior Guided Reasoning Module: As observed in Section III-C, the correlation coefficients between the overall quality and these distortion attributes are different, with values of 0.91, 0.61, 0.74, 0.91, and 0.92 for temporal-spatial visibility, brightness, specular reflection, stability, and utility, respectively. This indicates that, the subjects have different rating behaviors on different distortion attributes when evaluating the overall quality score of a video. In view of this, we propose a BGR module to estimate the overall quality score by considering the rating behaviors. As shown in Fig. 10, our BGR module is based on the Graph Convolution Network (GCN) [47], [48] and consists of two GCN layers. Specifically, it takes the feature $\mathcal{H}^l$ $(l \in \{0, 1\})$ and the adjacency matrix $\widehat{\mathcal{A}} \in \mathbb{R}^{n \times n}$ as the inputs, and updates $\mathcal{H}^l$ via a GCN layer:

$$\mathcal{H}^{l+1} = h(\widehat{D}^{-\frac{1}{2}}\widehat{\mathcal{A}}\widehat{D}^{-\frac{1}{2}}\mathcal{H}^l\mathcal{W}^l), \tag{8}$$

where $h(\cdot)$ denote the LeakyReLU function for the nonlinear operation. $\widehat{D}$ is the diagonal matrix, and $\widehat{D}_{i,i} = \sum_{j=1}^{n} \widehat{\mathcal{A}}_{i,j}$. To generate $\mathcal{H}^0 \in \mathbb{R}^{n \times P}$ $(P = N \times C')$, we first concatenate all attribute-aware features $(\mathcal{S}_n,\ n \in \{1, 2, \cdots, 5\})$ and then

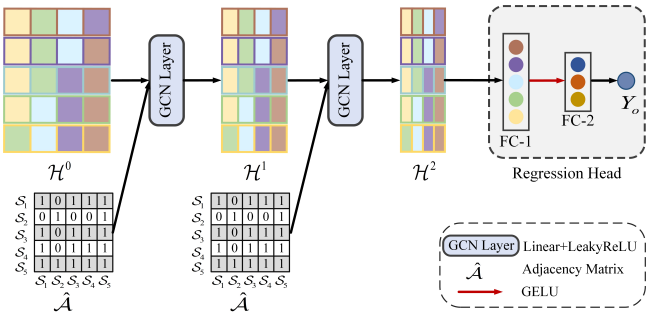

Fig. 10. Framework of the Behavior Guided Reasoning (BGR) Module.

reshape the concatenated feature. $\mathcal{W}^l$ is the l-th layer-specific trainable weight matrix. The adjacency matrix $\widehat{\mathcal{A}}$ is calculated by:

$$\widehat{\mathcal{A}} = \mathcal{A} + \mathcal{I}_{\mathcal{N}}, \tag{9}$$

where $\mathcal{A}$ is the correlation matrix that includes the relationship between different network nodes. $\mathcal{I}_{\mathcal{N}}$ denotes the identity matrix used for self-connections. Different from previous works [49] that initialize the parameters of $\mathcal{A}$ by representing all network nodes with the same distortion attribute, our BGR module sets all distortion attributes as network nodes and gives specific considerations to their relationship based on the behavioral feedback shown in Fig. 6. This setting makes our BGR more interpretable and suitable for the colonoscopy VQA task. Specifically, for the element in $\mathcal{A}$, we can only set its value to 1 when the PLCC value between the attributes in the column and row is larger than 0.55. Finally, we map the output $\mathcal{H}^2$ of the last GCN layer to the overall quality score Y_o by using a regression head.

E. Loss Function

We implement a multi-task learning strategy to optimize our whole network. The total loss function of DARNet is defined as:

$$\mathcal{L}_t = \mathcal{L}(G_o, Y_o) + \sum_{n=1}^{5} \mathcal{L}(G_n, Y_n), \tag{10}$$

where G_o is the MOS of the overall quality, and G_n is the MOS of the n-th distortion attribute of the colonoscopy video. In Eq. (10), the first and second terms are used to supervise the main and auxiliary tasks, respectively. Following previous works [47], [48], [50], we choose the widely used mean square error loss as $\mathcal{L}(\cdot, \cdot)$.

V. Experiments And Results

A. Experimental Setup

1) Data Division: In this study, we randomly divide the constructed MAC-VQA database into training and test sets in a ratio of 8:2. To investigate the performance of a VQA algorithm on individual data, the collected videos from each patient exits in both training and test sets, without content repetition. Furthermore, to avoid performance bias, we conduct the random train-test split procedure 10 times, and the reported results are the average of 10 test outcomes.

TABLE I
RESULTS OF DIFFERENT VQA METHODS ON THE CONSTRUCTED DATABASE.

Methods	Year	Deep	SRCC	PLCC	KRCC	RMSE
TLVQM [39]	2019	✖	0.887±0.015	0.885±0.013	0.702±0.020	0.466±0.025
VIDEVAL [40]	2021	✖	0.888±0.008	0.891±0.006	0.709±0.009	0.461±0.013
VSFA [25]	2019	✓	0.932±0.008	0.934±0.008	0.775±0.015	0.356±0.019
Simple-VQA [30]	2022	✓	0.897±0.012	0.901±0.011	0.718±0.017	0.433±0.018
BVQA [31]	2022	✓	0.928±0.009	0.936±0.008	0.769±0.015	0.354±0.011
FAST-VQA-M [27]	2022	✓	0.898±0.015	0.902±0.013	0.721±0.020	0.433±0.020
FASTER-VQA [41]	2023	✓	0.930±0.011	0.930±0.011	0.771±0.019	0.369±0.021
DOVER [15]	2023	✓	0.933±0.009	0.936±0.009	0.778±0.016	0.352±0.022
Light-VQA [51]	2023	✓	0.836±0.014	0.842±0.012	0.643±0.014	0.550±0.023
DARNet(ours)	2024	✓	**0.940±0.008**	**0.943±0.006**	**0.790±0.013**	**0.332±0.014**

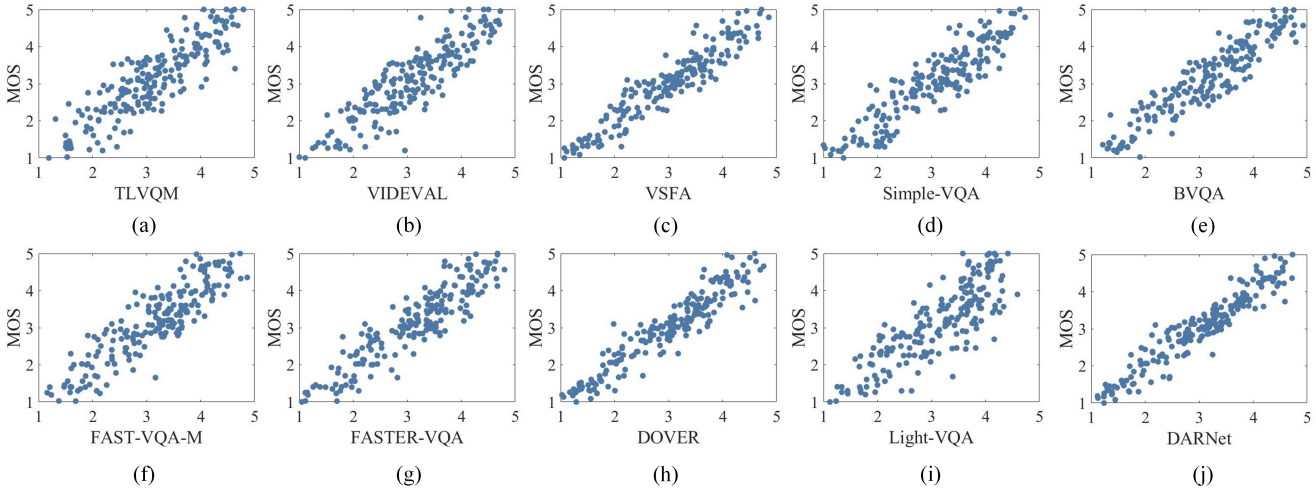

Fig. 11. Scatter plots of nine VQA methods and our proposed DARNet on the constructed database.

2) Evaluation Metrics: To quantitatively evaluate the performance of VQA methods, we select four mainstream evaluation metrics in the VQA field [52], including SRCC, PLCC, Kendall rank-order correlation coefficien (KRCC), and root mean square error (RMSE). Specifically, SRCC and KRCC assess prediction monotonicity, while PLCC and RMSE measure prediction accuracy. Notably, higher values of SRCC, KRCC, and PLCC, along with lower values of RMSE, indicate better performance. A nonlinear fitting function recommended by the video quality experts group [53] is used to map the predicted scores s to subjective scores $f(s)$ before calculating PLCC and RMSE:

$$f(s) = \frac{\eta_1 - \eta_2}{1 + e^{-\frac{s-\eta_3}{\eta_4}}} + \eta_2, \tag{11}$$

where η_i $(i \in \{1, ..., 4\})$ are the fitting parameters.

3) Implementation Details: Our DARNet is implemented within the PyTorch framework. In the training stage, we set the batch size to 8 and apply the AdamW optimizer to train the network by minimizing Eq. (10) for 70 epochs. The initial learning rates for the backbone and the other layers are 1e-5 and 1e-4, respectively, decaying according to a cosine annealing schedule. In the inference stage, for a query video, we use the sampling strategy described in Section IV-B and feed the sampled information into the well-trained model to get the predicted scores. All experiments are conducted on a workstation equipped with an NVIDIA RTX 3090 GPU.

4) Compared Methods: We compare our DARNet with nine state-of-the-art VQA methods, including TLVQM [39], VIDEVAL [40], VSFA [25], Simple-VQA [30], BVQA [31], FAST-VQA-M [27], FASTER-VQA [41], Light-VQA [51], and DOVER [15]. TLVQM and VIDEVAL are traditional handcrafted feature-based methods, while the others are DNN-based methods. We use the official source codes and default settings of these methods to train and test them on our constructed database, with the same data division manner described in Section V-A.1.

B. Comparison on the Whole Database

Table I shows the results of our proposed DARNet and nine competing VQA methods. The best results are marked in bold for convenient comparison.From the data in Table I, we have several findings. First, traditional handcrafted feature-based methods (i.e., TLVQM and VIDEVAL) have an inferior capability for accurately evaluating colonoscopy videos compared to DNN-based methods. A possible reason for this is that, these methods cannot fully characterize the distortions in colonoscopy videos because they rely heavily on the prior experience, e.g., statistic regularity of distortions and human visual system modelling, to characterize distortions in natural scene videos. Notably, the distortions in colonoscopy videos differ from those in natural scene videos greatly, as described in Section III-B. The knowledge employed in TLVQM and

TABLE II
THE PERFORMANCE OF THE COMPETING METHODS ON EACH DISTORTION ATTRIBUTE.

Methods	Temporal-Spatial Visibility				Brightness				Specular Reflection				Stability				Utility			
	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE
TLVQM [39]	0.838	0.857	0.651	0.336	0.686	0.714	0.501	0.547	0.781	0.788	0.584	0.413	0.901	0.903	0.725	0.327	0.845	0.835	0.647	0.394
VIDEVAL [40]	0.855	0.873	0.666	0.319	0.680	0.701	0.493	0.598	0.773	0.796	0.579	0.410	0.902	0.907	0.726	0.326	0.833	0.840	0.643	0.395
VSFA [25]	0.904	0.918	0.738	0.259	0.783	0.809	0.591	0.472	0.844	0.860	0.659	**0.338**	**0.934**	**0.937**	**0.778**	**0.268**	0.898	0.903	0.724	0.300
Simple-VQA [30]	0.885	0.904	0.707	0.281	0.710	0.734	0.519	0.546	0.802	0.813	0.609	0.385	0.900	0.911	0.725	0.316	0.852	0.863	0.667	0.352
BVQA [31]	0.899	0.915	0.729	0.261	0.765	0.794	0.576	0.479	**0.844**	**0.863**	**0.659**	0.339	0.927	0.934	0.767	0.275	0.889	0.896	0.714	0.315
FAST-VQA-M [27]	0.859	0.879	0.674	0.309	0.724	0.761	0.532	0.510	0.799	0.815	0.608	0.388	0.901	0.907	0.726	0.327	0.869	0.869	0.682	0.352
FASTER-VQA [41]	0.893	0.910	0.723	0.268	0.770	0.799	0.578	0.474	0.821	0.838	0.631	0.366	0.921	0.926	0.755	0.294	0.889	0.889	0.710	0.325
DOVER [15]	0.908	0.923	0.744	0.252	0.784	0.808	0.593	0.474	0.838	0.854	0.650	0.344	0.925	0.931	0.763	0.281	0.897	0.898	0.723	0.307
Light-VQA [51]	0.853	0.870	0.668	0.324	0.569	0.590	0.397	0.636	0.733	0.736	0.541	0.458	0.873	0.886	0.692	0.361	0.797	0.801	0.605	0.435
DARNet(ours)	**0.910**	**0.924**	**0.744**	**0.252**	**0.798**	**0.821**	**0.607**	**0.459**	0.841	0.855	0.654	0.343	0.928	0.933	0.768	0.276	**0.906**	**0.914**	**0.737**	**0.283**

VIDEVAL is not entirely suitable for the colonoscopy VQA task. Second, among these DNN-based methods, methods that combine 2D-CNN with video Transformer have more advantages in performance than methods only based on video Transformer. For instance, DOVER, which adds a 2D-CNN branch to video Transformer, achieves considerable results, with a SRCC value greater than 0.933. This may be attributed to that, as the complement to video Transformer, 2D-CNN is conducive to capturing frame-level features to better characterize spatial distortions. Third, our proposed DARNet achieves better performance than all competing methods, with relatively smaller standard deviation value at each evaluation metric. This is because DARNet not only combines the merits of 2D-CNN and video Transformer, but also utilizes the multi-task learning and specifically designed modules to improve the quality-aware feature representation. More specifically, considering the inherent relationships between distortion attributes, it utilizes an MAA module to explore complementary information from different features, enhancing the attribute-related feature representations. In addition, it also incorporates the prior knowledge from subjective rating behaviors when fusing features of different distortion attributes, helping the network producing more accurate predictions. Fig. 10 shows the scatter plots of predictions versus subjective ratings. It is obvious that our proposed DARNet can produce a more compact scatter compared to competing VQA methods.

In addition, we investigate the correlation between the quality scores rated by each subject and the associated predictions provided by our DARNet. As shown in Fig. 12, the outcomes of DARNet are consistent with human visual perception. Specifically, the correlation results between our method and all subjects are consistent in 10 trials, with a slight fluctuation less than 0.15 in each quality dimension. For instance, both the mean values of PLCC and SRCC across 10 trials are greater 0.85 in overall quality. This demonstrates that these subjects perform similarly in the quality rating tasks. Moreover, the correlation results on brightness and specular reflection are smaller than those on temporal-spatial visibility, stability, utility, and overall quality. This indicates that, compared to the other distortion attributes, brightness and specular reflection are more challenging in accurate quality prediction. Furthermore, it is also interesting to investigate the performance of our method on each patient's videos. For this purpose, we first divide the test set into 40 groups based on patient ID. Each group contains approximately 4 to 8 videos, collected from one patient. Then, for each group, we calculate the PLCC and SRCC values between the predicted scores and subjective ratings (i.e., the ground-truth scores) of the overall quality. As shown in Fig. 13, the mean values of PLCC and SRCC across 10 trials are greater than 0.80 and 0.75, respectively. Results on other distortion attributes show similar conclusions, and we do not present them due to space limitation. This indicates that our method performs effectively on different patients' data, showing strong generalizability.

C. Comparison on Each Distortion Attribute

In this section, we investigate the effectiveness of VQA methods on evaluating each distortion attribute. Table II shows the results. Due to space limitation, we only present the mean value of each evaluation metric across 10 trials. From Table II, we have the following observations. Firstly, DNN-based methods generally outperform handcrafted feature-based methods (i.e., TLVQM and VIDEVAL), on all distortion attributes. The average SRCC values achieved by these two kinds of methods are 0.889 and 0.847 in temporal-spatial visibility, 0.738 and 0.683 in brightness, 0.815 and 0.777 in specular reflection, 0.913 and 0.902 in stability, 0.875 and 0.839 in utility, respectively. Secondly, the results of each method vary greatly across different distortion attributes. For example, the SRCC values achieved by Light-VQA are 0.733 and 0.873 in specular reflection and stability, respectively. Thirdly, most competing methods produce unsatisfactory performance on brightness, with both SRCC and PLCC values lower than 0.78. Last but not least, our DARNet shows leading advantages on most distortion attributes. Specifically, it achieves the best 12 times and the second best 4 times in a total of 20 comparisons. This indicates that our DARNet is superior to other VQA methods in evaluating the distortion attributes in colonoscopy videos.

D. Ablations Studies

In our DARNet, we employ ResNet-50 as a supplementary feature extractor in the STFE unit to better characterize spatial distortions and use an FIR unit to accurately estimate the quality of colonoscopy videos from different perspectives. This section conducts several ablation experiments to explore the effectiveness of ResNet-50 and FIR unit. All experiments are

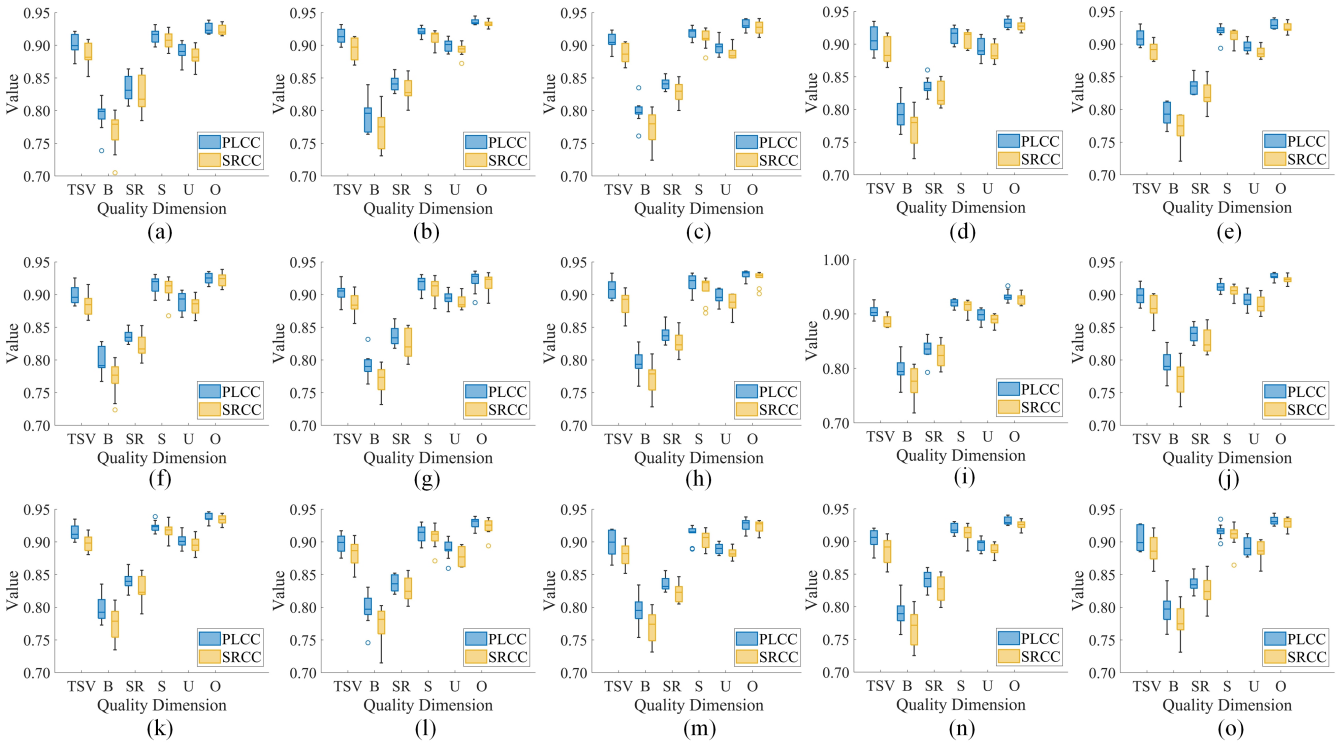

Fig. 12. The PLCC and SRCC scores calculated between the predicted scores of our model and the associated ratings of each subject in each quality dimension. TSV, B, SR, S, U, and O denote temporal-spatial visibility, brightness, specular reflection, stability, utility, and overall quality, respectively.

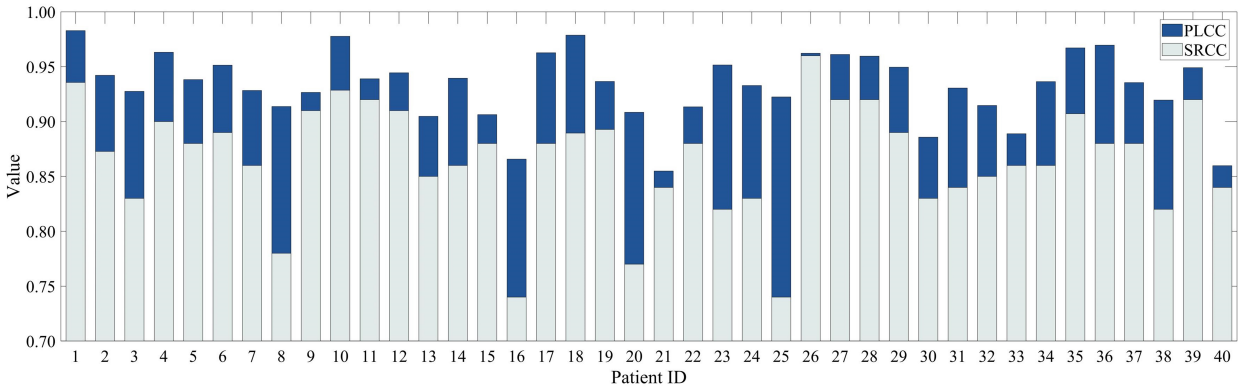

Fig. 13. The PLCC and SRCC scores calculated between the predicted scores and subjective ratings of the overall quality on each patient's videos.

carried out using the same settings described in Section V-A.3. Due to space limitation, only the mean results across 10 trials are present.

1) Effectiveness of the Supplementary Feature Extractor: In our DARNet, ResNet-50 is used as a supplementary feature extractor of Swin-T. Here, we learn a new VQA model by removing ResNet-50 from the STFE unit. In this case, only Swin-T is used to extract spatio-temporal features. The experimental results are tabulated in Table III, in which the best results are marked in bold. From the data, we can find that the removal of ResNet-50 will bright obvious performance drop. For instance, there is approximately 2.6%, 3.2%, 8.0%, 2.7%, 2.0%, and 1.9% of SRCC decrement when evaluating the temporal-spatial visibility, brightness, specular reflection, stability, utility, and overall quality, respectively. This demonstrates that ResNet-50 plays a positive role in accurately evaluating the quality of colonoscopy videos. A possible reason for this is that ResNet-50 helps the network capture more frame-level spatial features, which serve as the supplementary information for Swin-T in understanding spatial distortions.

2) Effectiveness of the FIR Unit: To investigate the effectiveness of the FIR unit, we remove it from DARNet and directly feed the feature $\mathcal{P}_i$ from the STFE unit into a regression head to estimate the quality score of the i-th distortion attribute. To estimate the overall quality score, all the features (i.e., $\mathcal{P}_1$, $\mathcal{P}_1$, $\cdots$, $\mathcal{P}_5$) are concatenated and input into another regression head. As tabulated in Table III, compared to the standard DARNet, the network without the FIR unit shows a decrement of 4.3%, 4.2%, and 4.7% in terms of PLCC, SRCC, and KRCC, respectively, when evaluating specular

TABLE III
ABLATION EXPERIMENT RESULTS OF RESNET-50 AND FIR UNIT.

	ResNet-50	FIR	SRCC	PLCC	KRCC	RMSE
Temporal Spatial Visibility	✖	✖	0.855	0.871	0.670	0.323
	✓	✖	0.875	0.896	0.697	0.291
	✖	✓	0.884	0.901	0.709	0.285
	✓	✓	**0.910**	**0.924**	**0.744**	**0.252**
Brightness	✖	✖	0.654	0.679	0.469	0.590
	✓	✖	0.771	0.798	0.580	0.485
	✖	✓	0.766	0.785	0.575	0.498
	✓	✓	**0.798**	**0.821**	**0.607**	**0.459**
Specular Reflection	✖	✖	0.713	0.715	0.518	0.462
	✓	✖	0.799	0.812	0.607	0.386
	✖	✓	0.761	0.773	0.569	0.419
	✓	✓	**0.841**	**0.855**	**0.654**	**0.343**
Stability	✖	✖	0.882	0.887	0.703	0.355
	✓	✖	0.908	0.915	0.740	0.310
	✖	✓	0.901	0.908	0.727	0.321
	✓	✓	**0.928**	**0.933**	**0.768**	**0.276**
Utility	✖	✖	0.885	0.879	0.704	0.334
	✓	✖	0.889	0.894	0.711	0.314
	✖	✓	0.886	0.892	0.707	0.316
	✓	✓	**0.906**	**0.914**	**0.737**	**0.283**
Overall	✖	✖	0.913	0.913	0.744	0.408
	✓	✖	0.925	0.927	0.763	0.374
	✖	✓	0.921	0.925	0.757	0.381
	✓	✓	**0.940**	**0.943**	**0.790**	**0.332**

TABLE IV
ABLATION EXPERIMENT RESULTS OF THE MAA AND BGR MODULES.

	BGR	MAA	SRCC	PLCC	KRCC	RMSE
Temporal Spatial Visibility	✖	✖	0.875	0.896	0.697	0.291
	✓	✖	0.878	0.896	0.700	0.291
	✖	✓	0.909	0.922	0.742	0.255
	✓	✓	**0.910**	**0.924**	**0.744**	**0.252**
Brightness	✖	✖	0.771	0.798	0.580	0.485
	✓	✖	0.776	0.802	0.584	0.481
	✖	✓	0.786	0.810	0.596	0.470
	✓	✓	**0.798**	**0.821**	**0.607**	**0.459**
Specular Reflection	✖	✖	0.799	0.812	0.607	0.386
	✓	✖	0.797	0.810	0.604	0.388
	✖	✓	0.837	0.850	0.650	0.349
	✓	✓	**0.841**	**0.855**	**0.654**	**0.343**
Stability	✖	✖	0.908	0.915	0.740	0.310
	✓	✖	0.908	0.916	0.736	0.309
	✖	✓	0.926	0.931	0.736	0.280
	✓	✓	**0.928**	**0.933**	**0.768**	**0.276**
Utility	✖	✖	0.889	0.894	0.711	0.314
	✓	✖	0.892	0.895	0.715	0.312
	✖	✓	0.904	0.912	0.734	0.288
	✓	✓	**0.906**	**0.914**	**0.737**	**0.283**
Overall	✖	✖	0.925	0.927	0.763	0.374
	✓	✖	0.926	0.929	0.766	0.368
	✖	✓	0.939	0.942	0.787	0.335
	✓	✓	**0.940**	**0.943**	**0.790**	**0.332**

reflection. This indicates that the proposed FIR unit contributes to achieving accurate quality predictions. In the FIR unit, we utilize some regression heads to help the network learn attribute-aware features under the supervision of quality scores rated for distortion attributes. Here, we further investigate the performance of DARNet when removing these regression heads. In this case, the network no longer uses the multi-task learning strategy and only preserves the supervision for overall quality. Experimental results show that, under such a setting, our method achieves a SRCC value of 0.933, a PLCC value of 0.936, a KRCC value of 0.777, and a RMSE value of 0.352 in evaluating overall quality. Compared with the results (see the last row of Table III) of our standard method, there is an approximate performance drop of 0.7%, 0.7%, 1.3%, and 2.0% in these four evaluation metrics, respectively. A possible reason for this is that, the network's ability to extract quality-aware features has decreased, as no supervisory information is provided and transmitted along the regression head. Overall, through the above results, we can conclude that the usage of attribute regression heads contributes to achieving good performance in our VQA task.

3) Effectiveness of the MAA Module: We further explore the effectiveness of the MAA module in the FIR unit through ablation experiments. Specifically, we remove the MAA module and directly map the feature $\mathcal{P}_i$ to the quality score of the i-th distortion attribute using a regression head, resulting in a new VQA model. As shown in Table IV, the MAA module is conducive to achieving accurate performance. For instance, its usage brings an increment of 2.6% and 3.4% in terms of PLCC and SRCC, respectively, when evaluating temporal-spatial visibility. This may be attributed to that the MAA module helps the network adaptively learn and integrate complementary information from distortion attributes using a CA mechanism. Here, we further investigate the performance of our method when removing the CA operation between these attributes that have a weak correlation. To be specific, only two attributes with a correlation score over the threshold $\mathcal{T}$ are considered to use a cross-attention operation in the MAA module. We set $\mathcal{T}$ to 0.55. Taking the attribute of brightness as an example, according to the results (see Fig. 6) of subjective experiments, it has correlation scores of 0.49, 0.48, 0.52, and 0.59 with temporal-spatial visibility, specular reflection, stability, and utility, respectively. In view of this, we only use the CA operation to integrate the features of brightness and utility. As shown in Table. V, the results of the VQA model (named Manu-$\mathcal{M}$) learned by integrating features from manually selected correlated attributes are similar to that of our standard VQA model (named Auto-$\mathcal{M}$ to distinguish) that adaptively integrates features from all attributes. This indicates that the usage of the CA mechanism in our MAA module helps the network effectively select useful information in an adaptive manner, without any experience-supported manual selection process. In other words, benefiting from the CA mechanism, our network become more flexible and intelligent.

4) Effectiveness of the BGR Module: We carry out more ablation experiments to explore the effectiveness of the BGR module in the FIR unit. Specifically, we remove the BGR module and directly fuse all attribute-aware features (i.e., $\mathcal{S}_1$, $\mathcal{S}_2$, $\cdots$, $\mathcal{S}_5$) using the linear mapping operations to predict the overall quality score. As shown in Table IV, the newly learned VQA model is inferior to our standard DARNet. For instance, removing BGR module leads to a performance decrement of 1.1% and 1.2% in terms of PLCC and SRCC, respectively, when evaluating brightness. This may be attributed to that this module considers the correlation between distortion attributes, which helps the network learn more complementary information from them. In summary, the BGR module plays a positive role in achieving accurate prediction.

TABLE V
RESULTS COMPARISON UNDER DIFFERENT CONFIGURATIONS OF THE MAA MODULE.

Methods	Temporal-Spatial Visibility				Brightness				Specular Reflection			
	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE
Manu-$\mathcal{M}$	**0.912**	**0.925**	**0.747**	**0.250**	0.797	0.819	0.605	0.461	0.840	0.853	0.652	0.345
Auto-$\mathcal{M}$	0.910	0.924	0.744	0.252	**0.798**	**0.821**	**0.607**	**0.459**	**0.841**	**0.855**	**0.654**	**0.343**
Methods	Stability				Utility				Overall			
	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE
Manu-$\mathcal{M}$	**0.930**	**0.934**	**0.770**	**0.274**	0.904	0.910	0.736	0.290	0.939	0.942	0.787	0.335
Auto-$\mathcal{M}$	0.928	0.933	0.768	0.276	**0.906**	**0.914**	**0.737**	**0.283**	**0.940**	**0.943**	**0.790**	**0.332**

E. Discussions

Colonoscopy is a primary choice for screening colorectal diseases. However, captured colonoscopy videos are usually with inadequate quality due to factors caused by poor environmental conditions and moving cameras. Low-quality video not only prevents reliable diagnosis by gastroenterologists, but also affects the performance of computer-aided diagnosis systems. Therefore, it is necessary to design an effective VQA method to monitor the quality of captured colonoscopy videos. Unfortunately, we have witnessed intense discussions on proposing VQA methods for natural scene videos, yet rare efforts for colonoscopy videos.

In this study, we advance this topic from two aspects. First, we propose and release the first multi-attribute colonoscopy VQA database. The constructed database includes 1,000 real-world videos, each of which is annotated from five distortion attributes as well as the overall perspective. The results of subjective ratings indicate that these distortion attributes have inherent relationships in the VQA task and correlate differently with the overall quality. This finding provides us insights for designing effective colonoscopy VQA methods. Nine state-of-the-art VQA methods specifically designed for natural scene videos are tested to investigate whether these methods still guarantee the general effectiveness on the colonoscopy data. The results are thoroughly discussed, with a conclusion that we still need to design specific colonoscopy VQA methods. Second, based on the observations from subjective ratings, we propose a specific colonoscopy VQA method, named DARNet. Our DARNet follows a multi-task learning framework, using some auxiliary tasks, i.e., quality predictions of five distortion attributes, to boost the performance of the main task, i.e., overall quality prediction. The key technical contribution of DARNet is a well-designed FIR unit, which explores and integrates attribute-aware features for accurate quality prediction. The FIR unit consists of an MAA module and a BGR module. The former explores complementary information by make interactions between multiple attribute features, while the latter estimates the overall quality by considering the rating behaviors. Experimental results on our constructed database show that our proposed DARNet is competent for the colonoscopy VQA task, with promising results. In addition, as a data-driven VQA method, our DARNet also show great potential in accurately evaluating the quality of nature scene videos. This conclusion is drawn from the results on a natural scene-oriented VQA database, named MaxWell [15]. MaxWell has 4,543 videos (including 3,634 training videos and 909 test videos), and each video is annotated from the perspectives of aesthetics, technology, and overall quality. Compared to the second best method DOVER, our DARNet has a PLCC advantage of 4.8%, 5.1%, and 0.4% on these three quality perspectives, respectively.

As an important technique for selecting high-quality video data, our DARNet has many practical applications. One the one hand, it can provide nuanced and detailed quality scores for videos from the perspectives of five attributes (i.e., temporal-spatial visibility, brightness, specular reflection, stability, and utility) and overall quality, reminding gastroenterologists to adjust their operations to obtain high-quality videos for better disease diagnosis in colonoscopy. On the other hand, considering the emerging demand for learning robust computer-aided diagnosis models by using multi-center data, it can be used as an effective quality examiner to select high-quality colonoscopy videos from different centers.

VI. CONCLUSION

This research conducts a comprehensive study on the colonoscopy VQA field from both subjective and objective perspectives. To cope with the challenge of insufficient public database, we first build a colonoscopy VQA database based on rigorous subjective experiments. Each video is annotated from five distortion attributes and the overall quality. The subjective experiment results show that these five distortion attributes not only have inherent relationships in the VQA task, but also correlate differently with the overall quality. Based on this finding, we propose a novel quality assessment network, termed DARNet, for colonoscopy videos. The proposed DARNet extracts spatio-temporal features using a STFE unit to characterize the complex distortions and adaptively integrates the extracted features using an FIR unit. Thanks to the effective cooperation of these two units, the DARNet provides comprehensive quality assessment for the colonoscopy video, with better performance than nine state-of-the-art VQA methods.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] N. Keum and E. Giovannucci, "Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies," *Nature Reviews Gastroenterology & Hepatology*, vol. 16, no. 12, pp. 713–732, 2019.

[3] X. Guo, C. Yang, Y. Liu, and Y. Yuan, "Learn to threshold: Thresholdnet with confidence-guided manifold mixup for polyp segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 4, pp. 1134–1146, 2020.

[4] X. Liu and Y. Yuan, "A source-free domain adaptive polyp detection framework with style diversification flow," *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1897–1908, 2022.

[5] G. Yue, H. Xiao, T. Zhou, S. Tan, Y. Liu, and W. Yan, "Progressive feature enhancement network for automated colorectal polyp segmentation," *IEEE Transactions on Automation Science and Engineering*, accepted, in press, DOI:10.1109/TASE.2024.3430896, 2024.

[6] S. Jain, R. Atale, A. Gupta, U. Mishra, A. Seal, A. Ojha, J. Kuncewicz, and O. Krejcar, "Coinnet: A convolution-involution network with a novel statistical attention for automatic polyp segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3987–4000, 2023.

[7] G. Yue, P. Wei, Y. Liu, Y. Luo, J. Du, and T. Wang, "Automated endoscopic image classification via deep neural network with class imbalance loss," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[8] Q. Wang, K. Castro, V. N. Desai, W.-C. Cheng, and J. Pfefer, "Comparison of methods for quantitative evaluation of endoscopic distortion," in *Medical Imaging 2015: Physics of Medical Imaging*, vol. 9412. SPIE, 2015, pp. 814–820.

[9] G. Yue, D. Cheng, T. Zhou, J. Hou, W. Liu, L. Xu, T. Wang, and J. Cheng, "Perceptual quality assessment of enhanced colonoscopy images: A benchmark dataset and an objective method," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5549–5561, 2023.

[10] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[11] Y. Guo, M. Hu, X. Min, Y. Wang, M. Dai, G. Zhai, X.-P. Zhang, and X. Yang, "Blind image quality assessment for pathological microscopic image under screen and immersion scenarios," *IEEE Transactions on Medical Imaging*, vol. 42, no. 11, pp. 3295–3306, 2023.

[12] Z. Shen, H. Fu, J. Shen, and L. Shao, "Modeling and enhancing low-quality retinal fundus images," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 996–1006, 2021.

[13] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.

[14] X. Yu, Z. Tu, Z. Ying, A. C. Bovik, N. Birkbeck, Y. Wang, and B. Adsumilli, "Subjective quality assessment of user-generated content gaming videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 74–83.

[15] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 144–20 154.

[16] Q. Chen, X. Min, H. Duan, Y. Zhu, and G. Zhai, "Muiqa: Image quality assessment database and algorithm for medical ultrasound images," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2958–2962.

[17] Y. Lu, F. Xie, Y. Wu, Z. Jiang, and R. Meng, "No reference uneven illumination assessment for dermoscopy images," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 534–538, 2014.

[18] F. Shao, Y. Yang, Q. Jiang, G. Jiang, and Y.-S. Ho, "Automated quality assessment of fundus images via analysis of illumination, naturalness and structure," *IEEE Access*, vol. 6, pp. 806–817, 2017.

[19] Z. A. Khan, A. Beghdadi, M. Kaaniche, and F. A. Cheikh, "Residual networks based distortion classification and ranking for laparoscopic image quality assessment," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 176–180.

[20] L. Lévêque, W. Zhang, C. Cavaro-Ménard, P. Le Callet, and H. Liu, "Study of video quality assessment for telesurgery," *IEEE Access*, vol. 5, pp. 9990–9999, 2017.

[21] Y. Li, S. Meng, X. Zhang, M. Wang, S. Wang, Y. Wang, and S. Ma, "User-generated video quality assessment: A subjective and objective study," *IEEE Transactions on Multimedia*, vol. 25, pp. 154–166, 2021.

[22] Y. Fang, Z. Li, J. Yan, X. Sui, and H. Liu, "Study of spatio-temporal modeling in video quality assessment," *IEEE Transactions on Image Processing*, vol. 32, pp. 2693–2702, 2023.

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

[24] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-vq:'patching up'the video quality problem," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 019–14 029.

[25] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2351–2359.

[26] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "Rirnet: Recurrent-in-recurrent network for video quality assessment," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 834–842.

[27] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," in *European Conference on Computer Vision*. Springer, 2022, pp. 538–554.

[28] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, and W. Lin, "Discovqa: Temporal distortion-content transformers for video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4840–4854, 2023.

[29] S. Jiang, Q. Sang, Z. Hu, and L. Liu, "Self-supervised representation learning for video quality assessment," *IEEE Transactions on Broadcasting*, vol. 69, no. 1, pp. 118–129, 2023.

[30] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 856–865.

[31] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5944–5958, 2022.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[34] B. Münzer, K. Schoeffmann, L. Böszörményi, J. Smulders, and J. J. Jakimowicz, "Investigation of the impact of compression on the perceptional quality of laparoscopic videos," in *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*. IEEE, 2014, pp. 153–158.

[35] M. A. Usman, M. R. Usman, and S. Y. Shin, "Quality assessment for wireless capsule endoscopy videos compressed via hevc: From diagnostic quality to visual perception," *Computers in Biology and Medicine*, vol. 91, pp. 112–134, 2017.

[36] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 491–495.

[37] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[38] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3329–3342, 2016.

[39] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

[40] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.

[41] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, J. Gu, and W. Lin, "Neighbourhood representative sampling for efficient end-to-end video quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 185–15 202, 2023.

[42] X. Guan, F. Li, Z. Huang, and H. Liu, "Study of subjective and objective quality assessment of night-time videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6627–6641, 2022.

[43] S. Ali, F. Zhou, A. Bailey, B. Braden, J. E. East, X. Lu, and J. Rittscher, "A deep learning framework for quality assessment and restoration in video endoscopy," *Medical Image Analysis*, vol. 68, p. 101900, 2021.

[44] R. BT, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, vol. 4, 2002.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[46] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3202–3211.

[47] Y. Huang, L. Li, Y. Yang, Y. Li, and Y. Guo, "Explainable and generalizable blind image quality assessment via semantic attribute reasoning," *IEEE Transactions on Multimedia*, vol. 25, pp. 7672–7685, 2023.

[48] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Learning graph convolutional networks for multi-label recognition and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6969–6983, 2023.

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

[50] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3289–3299.

[51] Y. Dong, X. Liu, Y. Gao, X. Zhou, T. Tan, and G. Zhai, "Light-vqa: A multi-dimensional quality assessment model for low-light video enhancement," *arXiv preprint arXiv:2305.09512*, 2023.

[52] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[53] J. Antkowiak, T. J. Baina, F. V. Baroncini, N. Chateau, F. FranceTelecom, A. C. F. Pessoa, F. S. Colonnese, I. L. Contin, J. Caviedes, and F. Philips, "Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000," *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment march 2000*, 2000.