LIVE-ShareChat UGC-VQA Database

Sandeep Mishra, Alan C. Bovik, *Fellow, IEEE,*

***Abstract*—We present a large-scale subjective study of User-Generated-Content Video Quality Assessment (UGC-VQA) on a diverse set of videos from ShareChat. The consistently evolving space of telecommunications combined with the advancements in affordable and reliable consumer capture devices has led to an exponential growth of social media platforms. This effect is particularly remarkable in developing countries like India. Owing to the huge diversity in cultures and languages, the Indian social media platform, ShareChat, provides a safe and culturally oriented space for users to generate and share content in their preferred languages. Such diverse content demands better systems for evaluating the perceived visual quality of these videos in order to provide better user recommendations.**

***Index Terms*—**

I. Introduction

ACCORDING to a recent report by ETGovernment[2], India is expected to surpass 1 billion internet users. Another report by The Economic times

II. Related Work

A. UGC-VQA Databases

The Camera Video Database (CVD2014) was one of the first VQA databases that was relevant for the UGC-VQA scenario. It was followed by the LIVE-Qualcomm Mobile In-Capture Database. Both of these datasets were indeed UGC but modelled only camera-capture distortions on not very diverse content. With significant growth in crowd-sourcing techniques, the authors of [] created the KoNViD-1k VQA database. They sampled 1200 UGC videos from the YFCC100M dataset which were scored on the basis of their perceived video quality by 642 crowd-workers. Another such crowd-sourced VQA database is the LIVE-VQC database with 585 videos, scored by 4776 unique Amazon Mechanical Turk participants. YouTube-UGC Dataset is another similar recently published dataset that contains 1380 video clips rated by more than 8000 human subjects. Another very recent VQA database is the Large-Scale Social Video Quality Database (LSVQ) which contains around 39k videos rated by more than 6000 unique subjects. The differentiating factor for LSVQ is the fact that they also collect human opinions for 117k space-time video patches cropped from the original set of videos.

B. UGC-VQA Models

III. Details of Subjective Study

A. LIVE-ShareChat UGC-VQA Database

The LIVE-ShareChat database contains 600 carefully selected videos from a publicly available set of 20,000 videos on the ShareChat website. The videos were pre-labeled with annotations pertaining to certain issues commonly found in user-generated content such as jitter and blur, abnormal lighting, too much camera movement, etc. We make sure to equally represent each annotated issue in the chosen subset. The dimensions of the available videos depend completely on the camera specifications, the settings chosen during capture, and the editing after. Due to these reasons, the height of the videos varied anywhere between 528 to 5428 pixels, and the width varies between 320 to 2420 pixels. To avoid such huge differences in the dimensions of each video, we specifically choose videos with heights between 1000 to 1500, and widths between 500 to 800 pixels. As can be seen in figure [[**?**]], the peaks of the frequency distribution of each dimension lie in the chosen ranges. Note that for all the chosen videos height is greater than the width, thus making them suitable for viewing in portrait mode which is the preferred mode for social media platforms like ShareChat, Instagram, TikTok, etc. We also filter these videos such that the duration of each video lies anywhere between 10-65 seconds. This allows us to temporally crop the selected videos to clips of 8 seconds making them feasible for a psychometric human study. ITU-T P.913 section 6.5 [[**?**]] highly recommends videos of duration 8-10 seconds, since longer duration videos may have significantly higher quality variations thus making it difficult for the user to provide a global evaluation.

B. Subjective Study Environment

Our large-scale human study was conducted in the Subjective Study room at the Laboratory for Image and Video Engineering at The University of Texas at Austin. Since the majority of social media users browse such videos on mobile devices, we used a Google Pixel 5 with Android 11 operating system to display the videos using an in-house android application. The device has a 6” inch OLED panel with FHD+ resolution supporting a refresh rate of up to 90Hz. We fixed both the brightness and audio of the device to 75% of the maximum to avoid any automatic changes during the study. The device also supports automatic re-scaling of a video to fit the screen, thus removing any requirement on our end to do so. Hence videos are displayed to subjects just as they would have viewed them on their own devices. We also provided an external keyboard and a mouse to facilitate the viewing and rating experience of the subjects.

The study room is both sound and light-proof to mimic an isolated environment. We made sure that the artificial lighting arrangements did not interfere with the viewing conditions by placing them at strategic locations to simulate a living room’s lighting condition. The incident luminance on the mobile screen was measured to be approximately 150 lux. The

device was stationed on a smartphone mount with adjustable viewing angles and a height-adjustable chair was also provided to the subjects for them to comfortably position themselves for a good viewing experience. Subjects were recommended to sit at a distance of three-fourths of their arm's length to mimic typical social media browsing behavior. They were also suggested to avoid any significant changes to their seating and viewing arrangements once their study began. We also instructed them to not alter any device settings and use the provided keyboard and mouse to communicate with the application only when prompted.

Upon arrival, each participant was assigned a subject number as an identifier and a predefined playlist of videos is played for them. After each video playback, a rating screen appears with a rating bar for the subject to provide their evaluation of the subjective video quality. The rating bar was a continuous 0-100 scale, based on the SAMVIQ scale suggested in ITU recommendations [[**?**]], which had 5 labels: Bad(0), Poor(25), Fair(50), Good(75), and Excellent(100), where the quantity in the $(\cdot)$ is the absolute score suggested by the label. Only the labels were marked on the rating bar and not the scores. The initial position of the cursor was set to 0 for each rating and the subject was guided to use the wireless mouse to move the cursor to their desired score. Once the subject finalizes the score, they are supposed to press the NEXT button which will record their score in a text file and play the next video for them. The application does not support replaying a video since we want to record only the first instinctual response of the participant, hence they were guided to avoid any possible distractions during video playback.

C. Subjective Testing Protocol

We followed the single stimulus testing protocol in our human study according to the ITU-T recommendations [[**?**]]. As the LIVE-ShareChat dataset contains user-generated content, there is no concept of reference videos. The dataset contains a total of 600 videos with a viewing time of 8 seconds per video, resulting in a total of 4,800 seconds of playback time. Accounting for the rating time we estimated a total of 20 seconds for viewing and rating a single stimulus, resulting in a total of 12,000 seconds or 3.33 hours for the whole dataset. Since this is pretty high pressure for a volunteer, we split the dataset into four unique and uniformly distributed playlists such that each playlist had a total of 150 videos. We also estimated a total of 48 subjects and divided them evenly into 4 groups of 12 each. Each group was assigned two out of four playlists in a round-robin fashion. As a result, each volunteer views two playlists with 150 videos each, resulting in a total viewing time of 6,000 seconds. Since this is well below two hours, we split it into two sessions and play exactly one playlist in a single session. Each subject was required to attend two sessions in order to complete the study while making sure there was a minimum gap of 24 hours between the two sessions to reduce fatigue and bias of any sort. Given that each playlist was viewed by 24 subjects, each video in our dataset gets 24 ratings.

D. Subject Screening and Training

We recruited 48 volunteers from varying academic backgrounds from the student community of The University of Texas at Austin. Some of them are friends/colleagues, and members of the LIVE group while others are completely random. The volunteer pool had little/no experience in video quality evaluation apart from members of the LIVE group. Each subject participated in 2 sessions conducted over different days.

Vision deficiency tests were conducted for each volunteer to ensure low percentage of deficient participants. We conducted the Ishihara Color blindness test and found one color blind participant. While Snellen Eye test ensured everyone had a 20/20 vision with their corrective glasses/contact lenses on, if any. We did not use these tests to, in any way, validate/alter the participation of any volunteer which would render such a psychometric study unrealistic.

In the next step, subjects were introduced to the Subjective study room and the setup inside. They were introduced to the purpose of the study and the nature of the videos they would be viewing. We further instructed them to only rate the perceived visual quality, while completely ignoring the content. Since the videos in the database were originally created for entertainment/learning purposes, it is highly natural that the quality of the content can defocus a subject's attention towards visual aberrations and ultimately hamper their opinion. Thus it was necessary to familiarize them with the type of videos they would be viewing. To enable this, at the beginning of each session, subjects were shown three different videos of different quality and content as training videos. The scores recorded during the training session were not included in the psychometric database.

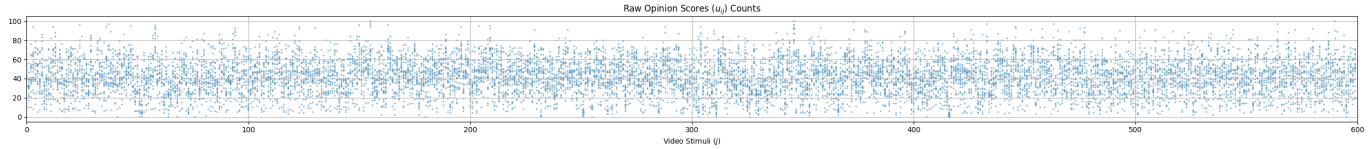
E. Post Study Questionnaire

At the conclusion of each video quality rating session, subjects were asked to fill out a questionnaire. This data was collected to ensure the reliability of the subjective ratings collected during the human study sessions. Within this subsection, we present a summary of answers to those questions and demographic information about the subjects.

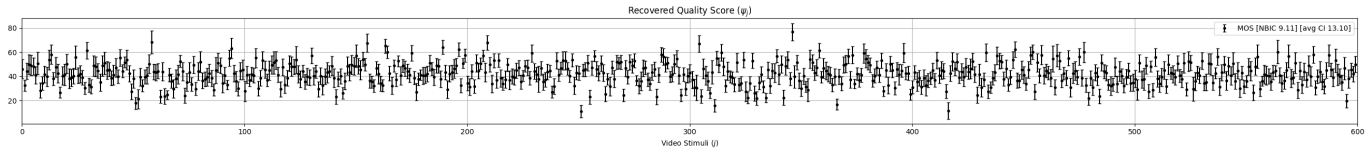
Approximately 85% of the subject population was male and the rest female. The minimum age of the subject pool was 21, while the maximum was 29. The mean, median, and standard deviation of the ages of the participants were found to be 24.75, 24.0, and 2.34. More than 90% of the pool felt that 8 seconds of visual playback were enough to adequately judge the quality of the videos. None of the participants complained about any kind of dizziness during their sessions.

F. Processing of subjective scores

We begin by evaluating the reliability of the recorded opinion scores. To do so, we first compute the inter-subject and intra-subject consistency scores.



(a)



(b)

Fig. 1: Raw opinion scores vs Recovered scores using MLE-MOS estimation

1) Inter-subject consistency: As the name suggests, it is the degree of correlation of the opinion scores amongst different subjects. To calculate the inter-subject consistency we split the scores recorded for every video into two distinct equal groups, and measure the correlation of MOS between these two groups. We repeat this process over 100 trials, and in each trial we the two distinct groups are chosen randomly. The median PLCC (Pearson linear correlation coefficient) over the 100 trials was 0.85 and the median SROCC (Spearman rank order correlation coefficient) over the 100 trials was 0.83.

2) Intra-subject consistency: It is a measure of how consistently an individual subject has scored each video with respect to its MOS. To quantify the intra-subject consistency we compute the PLCC and SROCC between the individual opinion scores and the MOS. The median PLCC was 0.62 and the median SROCC was 0.60. Since we are dealing with a UGC database hence the scores are obviously not as high as one would expect in the case of a synthetically generated database.

	SRCC	PLCC
Inter-Subject Consistency	0.8292	0.8449
Intra-Subject Consistency	0.5925	0.6154

TABLE I: Consistency Scores

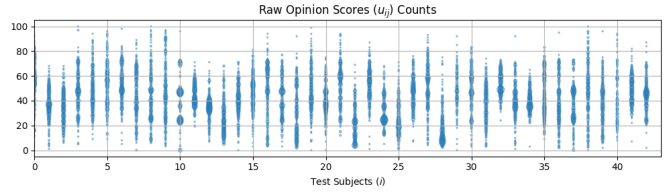


Fig. 2: Visual distribution of raw opinion scores

Given the consistency scores are not very high, we will need to incorporate subject rejection [citehere] via outlier detection while computing the final subjective quality score. To overcome this we employed the method described in [netflix paper citehere] which demonstrates how a maximum likelihood estimate (MLE) method of computing MOS offers advantages over traditional methods. The MLE method is less susceptible to subject corruption and provides tighter confidence intervals.

[netflix paper citehere] models the raw opinion scores of the videos as random variables $\{X_{e,s}\}$ with the following form:

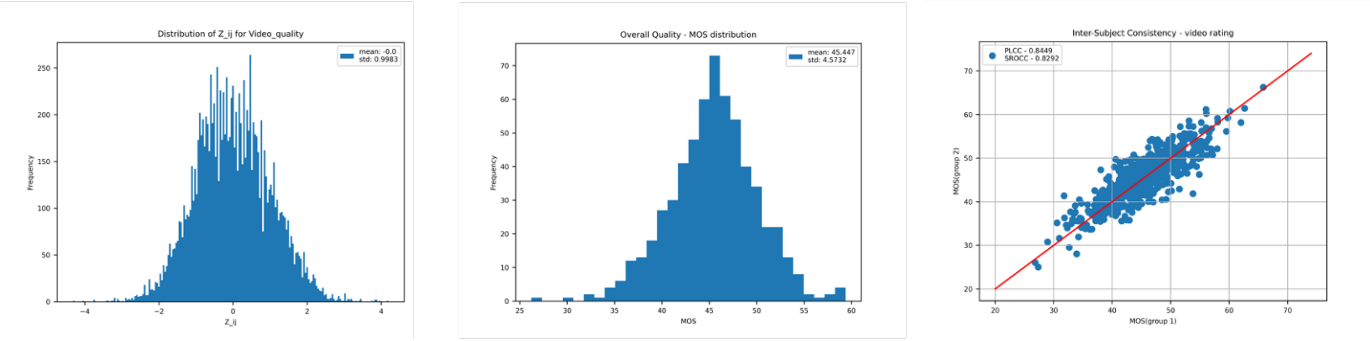
$$\begin{aligned} X_{e,s} &= x_e + B_{e,s} + A_{e,s} \\ B_{e,s} &\sim \mathcal{N}(b_s, v_s^2) \\ A_{e,s} &\sim \mathcal{N}(0, a^2_{c:c(e)=c}) \end{aligned} \tag{1}$$

where $e = 1, 2, 3, ..., 600$ are the indices of the videos in the database and $s = 1, 2, 3, ..., 48$ are the unique human participants. In the above model, x_e represents the quality of the video e as perceived by a hypothetical unbiased and consistent viewer. $B_{e,s}$ are i.i.d gaussian variables representing the human subject s, it is parameterized by a bias (i.e., mean) b_s and inconsistency (i.e., variance) v_s^2. This bias and inconsistency is assumed to remain same constant across all videos viewed by the subject s. $A_{e,s}$ are i.i.d gaussian variables representing a particular video content parameterized by the ambiguity (i.e., variance) a_c^2 of the content c, and $c = 1, 2, ...600$ indexes the unique source sequences in the database. All of the distorted versions of a reference video are presumed to contain the same level of ambiguity, and the video content ambiguity is assumed to be consistent across all users. In this formulation, the parameters $\theta = (x_e, b_s, v_s, a_c)$ denote the variables of the model. To estimate the parameters θ using MLE, the log likelihood function L is defined as :

$$L = \log P(\{x_{e,s}\}|\theta) \tag{2}$$

Using the data obtained from the psychometric study, we derive a solution for $\hat{\theta} = \text{argmax}_\theta L$ using the Belief Propagation algorithm, as shown in [netflix paper citehere]. We discuss more details about the extracted opinion scores in the next section.

Since the MLE model did not detect any outliers, hence as a sanity check, we also use the traditional method of calculating MOS using normalized opinion scores. Within this scope, let b_{ijk} denote the score recorded for video j provided by subject



(a) (b) (c)

Fig. 3: Dataset Distributions

i in session $k = 1, 2$. Let $\delta(i, j)$ be an indicator function, where

$$\delta(i,j) = \begin{cases} 1 & \text{if subject } i \text{ rated video } j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

We need $\delta(i, j)$ since not all videos in the database are rated by every subject. We calculate the normalized opinion scores received across multiple sessions of each subject as

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} m_{ijk}$$

$$\sigma_{ik} = \sqrt{\frac{1}{N_{ik} - 1} \sum_{j=1}^{N_{ik}} (m_{ijk} - \mu_{ik})^2}$$

$$z_{ijk} = \frac{m_{ijk} - \mu_{ik}}{\sigma_{ik}}$$

where z_{ijk} is the normalized opinion score or the Z-score per session and N_{ik} is the number of videos seen by subject i in session k. The Z-scores from all sessions were concatenated to form the matrix $\{z_{ij}\}$ denoting the Z-score assigned by subject i to the videos indexed by j with $j \in \{1, 2, \ldots, 600\}$, where the entries of $\{z_{ij}\}$ are empty at locations (i, j) where $\delta(i, j) = 0$. Assuming z_{ij} to have a standard normal distribution, 99% of the Z-scores were found to lie in $[-5, 6]$. A linear re-scaling was used to map scores to the range $[0, 100]$ as

$$z_{ij}^{'} = \frac{100(z_{ij} + 5)}{11} \tag{4}$$

Finally the Mean Opinion Score (MOS) of each video was calculated by averaging the scores received for that video as

$$MOS_j = \frac{1}{N_j} \sum_{i=1}^{N} z_{ij}^{'} \delta(i, j) \tag{5}$$

where $N_j = \sum_{i=1}^{N} \delta(i, j)$ and $N = 600$. The correlation between the scores obtained by MLE-MOS and traditional methods was 0.996.

The MOS were found to lie in the range [26.19, 65.66], and the mean of standard deviations of the rescaled Z-scores obtained from all subjects across all images was found to be 6.99. The histogram of MOS is shown in Fig. [citeFigHere] indicating a relatively broad MOS variation given the videos belong to in-the-wild scenario.

G. Analysis and Visualization of the Opinion Scores

IV. Benchmarking Objective NR-VQA Algorithms

We evaluated a number of publicly available No-Reference (NR-VQA) algorithms on the LIVE-ShareChat UGC database to understand the impact of the uniqueness of the database on existing VQA algorithms. We selected six well-known general-purpose NR-VQA models to test: NIQE [citehere], BRISQUE [citehere], VIDEVAL [citehere], RAPIQUE [citehere], VSFA [citehere], and CONTRIQUE [citehere]. NIQE and BRISQUE were originally published as image quality assessment algorithms, to adapt for videos we apply average pooling operation on the quality-aware features, extracted individually for each frame. For the off-the-shelf method NIQE, the predicted frame quality scores were directly pooled, yielding the final video quality scores. For the methods that require supervision before they can be used (BRISQUE, TLVQM, VIDEVAL, RAPIQUE, VSFA, CONTRIQUE), we used a support vector regressor (SVR) with the radial basis function kernel to learn mappings from the pooled quality-aware features to the ground truth MLE-MOS. VIDEVAL carefully curates 60 statistical features that showed high correlation with human opinion scores for the databases available at the time of its inception. VSFA uses a Resnet-50 [35] deep learning backbone to obtain quality-aware features, followed by a single-layer Artificial Neural Network (ANN) and Gated Rectified Unit (GRU) [36] to map features to MLE-MOS. RAPIQUE combined natural scene statistics with deep learning features in an attempt to create a mixture of experts-based model. CONTRIQUE uses contrastive pre-training to learn features associated with image quality. We evaluated the performance of the objective NR-VQA algorithms using the following metrics: Spearman's Rank Order Correlation Coefficient (SROCC), Kendall Rank Correlation Coefficient (KRCC), Pearson's Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE). The metrics SROCC and KRCC measure the monotonicity of the objective model prediction with respect to human scores, while the metrics PLCC and RMSE measure prediction accuracy. As stated earlier for the PLCC and RMSE measures, the predicted

Method	SRCC ↑	KRCC ↑	PLCC ↑	RMSE ↓
NIQE	0.3954	0.2276	0.3288	5.0920
BRISQUE	0.4766	0.3340	0.4922	4.8439
VIDEVAL	0.7104	0.5172	0.7087	3.8681
RAPIQUE (ResNet+SNSS+TNSS)	0.7280	0.5410	0.7392	3.7194
Contrique	0.7154	0.5254	0.7241	3.8120
Contrique+SNSS+TNSS	0.7353	0.5436	0.7403	3.7153
Ours	0.7048	0.5203	0.7150	3.8655
Ours+SNSS+TNSS	**0.7524**	**0.5626**	**0.7599**	**3.5932**

TABLE II: Your caption here

quality scores were passed through a logistic non-linearity function [38] to further linearize the objective predictions and to place them on the same scale as MOS :

$$f(x) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp(-x + \beta_3/|\beta_4|)} \tag{6}$$

We tested the algorithms mentioned above on 1000 random train-test splits using the four metrics. For each split, 80% of the videos were randomly chosen for training and validation, while the remaining 20% constituted the test set. All the algorithms were tested on the test set after pre-training them on the train set generated using the aforementioned train-test split, except NIQE, which doesn't require any pre-training. Since NIQE is an unsupervised model, we evaluated its performance on all 1000 test sets, without any training. We applied five-fold cross-validation to the training and validation sets of BRISQUE, TLVQM, VIDEVAL, RAPIQUE, VSFA and CONTRIQUE to find the optimal parameters of the SVRs they were built on. When testing VSFA, for each of the 1000 splits, the train and validation videos were used to select the best-performing ANN-GRU model weights on the validation set.

A. Performance of NR-VQA Models

Table IV lists the performances of the aforementioned NR-VQA algorithms on the LIVE-ShareChat UGC-VQA database. In addition, we used the 1000 SROCC and PLCC scores produced by the NR VQA models to run one-sided t-tests, using the 95% confidence level, to determine whether one VQA algorithm was statistically superior to another. Each entry in Table V consists of two symbols, where the first symbol corresponds to the t-test done using the SROCC values, and the second symbol corresponds to the t-test done using the PLCC values. We found that NIQE performed poorly, which is unsurprising since it was developed using a set of pristine images available at the time of its development. Over time, the quality and characteristics of cameras and camera processing pipelines have changed drastically. This distinction between the training set of NIQE and the test set contributes largely to its score. However, the performance of the same NIQE features improved when we extracted them and used an SVR to regress from the features to the MLE-MOS in the BRISQUE algorithm. The impact of the set of pristine images used in NIQE is clearly reflected by the gap in performance between NIQE and BRISQUE. The performance of VIDEVAL, RAPIQUE, and CONTRIQUE was definitely a stark improvement over NIQE and BRISQUE. In the case of VIDEVAL, this boost can probably be attributed to the fact that the model uses many hand-tuned hyper-parameters that were selected to optimize the prediction of video quality on general-purpose content. On the other hand, CONTRIQUE being a deep learning model has been trained on a huge dataset of 2M images and performs accordingly. Meanwhile, RAPIQUE which is a hybrid model, performs best by combining the handcrafted perceptual quality features with the high-level semantic features generated through its deep learning module.

Although we noted that RAPIQUE performed the best, the performance of VIDEVAL, and CONTRIQUE did not fall too behind. We must also note that the models, at their core, use different base principles. While VIDEVAL shows the usefulness of the statistical features, CONTRIQUE demonstrates that the same can be achieved solely by deep-learning-based features. RAPIQUE, a hybrid model, further strengthens the fact that the combination of statistical features and deep-learning-based features is better than either one of them.

V. Mixture-of-Expert based NR-VQA Algorithm

To this end, we present a novel NR-VQA algorithm that is designed while keeping the aforementioned learnings in sight. We call our method m-RAPIQUE (modified RAPIQUE), which is a Mixture-of-Experts-based quality assessment algorithm. We extend the concept of RAPIQUE and develop a hybrid model that combines Spatial NSS features, Temporal NSS features, and deep-learning features. Since spatial and temporal NSS features are handcrafted for quality assessment, improving upon them is difficult and hence we borrow them as it is. We do however modify the deep-learning module as the pre-trained model used in RAPIQUE is a simple ResNet-50 trained for image classification task on ImageNet dataset using supervised techniques. RAPIQUE claims that the inherent semantic understanding of the deep-learning model helps them predict perceptual quality closer to human opinions, thus reinstating the fact that semantic information plays a role in perceptual quality assessment. While we agree with this, we believe there are better ways to pre-train such a network in a more effective way than training using a supervised task. Many recent research articles have shown how unsupervised training captures more general information than supervised training, thus helping the model perform better in various correlated tasks instead of specializing in one. With this concept in mind, CONTRIQUE builds a contrastive training environment within which it trains the model to learn the

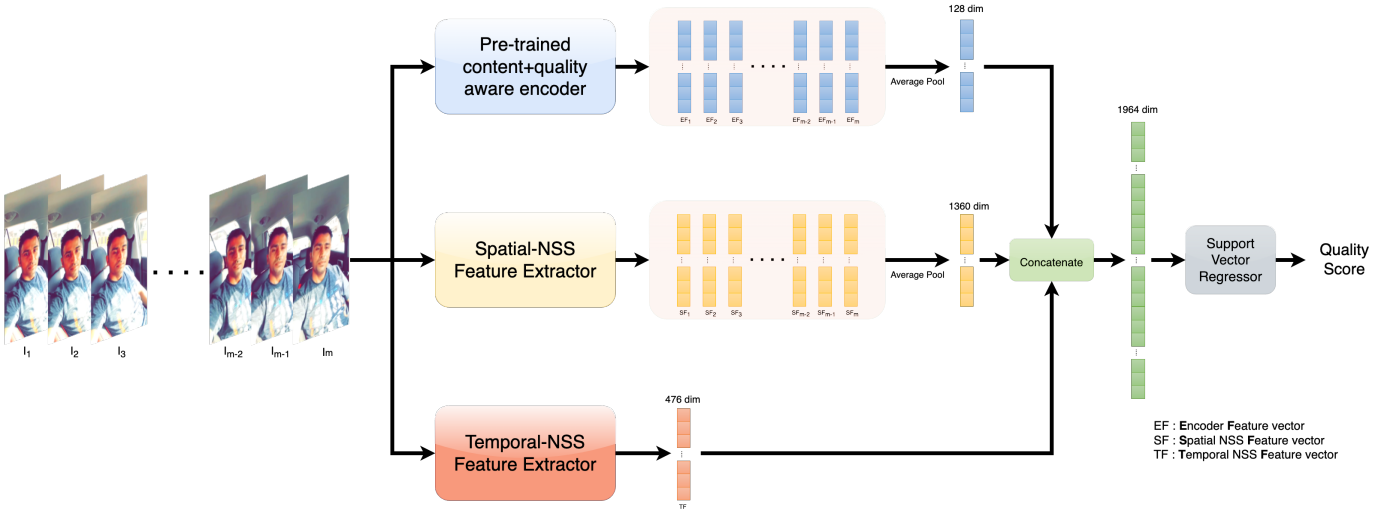


Fig. 4: MoEVA Evaluation Pipeline

distinction between different distortions. CONTRIQUE uses a synthetic dataset generated by applying fixed distortions to a set of pristine images. Any two images with the same distortion are categorized as the same and any two images with different distortion settings are marked as different thus fitting into the contrastive loss. While CONTRIQUE improves the model's understanding of the distortions by forcing it to learn them using a contrastive loss, it also hinders its understanding of the content by using the same loss. It forces the network to generate similar features for two images with the same distortion settings even if their content is different. We develop a novel method to apply contrastive learning to understand distortion behavior under the impact of content.

For contrastive learning-based training, we need pairs of images that are labeled either the same or different. We also want to train the network on the images at the original scale same as the test domain. Hence we operate on a patch level instead of a full frame. To create the distinction between the labels and to create a protocol to label pairs accordingly, we lay out the following hypothesis:

- **H1**: Perceptual quality of two nearby patches would have higher chances of being similar than the perceptual quality of two distant patches cropped from the same image, considering the local content within the distant patches to be more different as compared to the content in the nearby patches. We also extend it and assume that the perceptual quality features of two distinct patches taken from two different images are different. Note that this does not enforce any condition on the quality score, different quality features can lead to similar scores since the SVR can be a many-to-one function.

- **H2**: Two differently distorted versions of the same patch ought to have different perceptual quality features. Since the content within the two patches is exactly the same, hence any difference in the perceptual quality should be reflected in the perceptual quality features generated from our network as the SVR is not a one-to-many function.

Contrastive training is a self-supervised training strategy that learns by understanding the connection between the two images in an input pair. This connection is characterized by either a **+ve** sample or a **-ve** sample. A **+ve** sample occurs when the inputs in a pair are labeled as similar/same and hence encourages the model to generate similar features for the two inputs. While a **-ve** sample occurs when the inputs in the pair are labeled as different, thus encouraging the model to generate different features. In the following subsection, we explain our augmentation scheme that aligns with the generation of such pairs while labeling them using the aforementioned hypotheses.

A. Training dataset

Since the 600 videos in the LIVE-ShareChat UGC-VQA database were curated from a larger publicly available set of 20,000 videos, it is safe to assume that the all of the videos not included in the dataset have similar features/characteristics as the ones inlcuded in it. Thus we choose to train our model on the frames extracted from these 20,000 videos excluding the 600 videos in the LIVE-ShareChat UGC-VQA database.

B. Content+Quality Aware Augmentation scheme

To enable the model to learn the distinction between different distortion settings we use an augmentation bank that includes 25 distinct image-specific synthetic distortions, with 5 levels within each distortion. For any source image i^k from the training set, where $k \in 1, 2...K$ and K is the total number of images in the training data, a randomly chosen subset of the augmentations available in the bank are applied to each image resulting in a mini-batch of distorted images. We combine each source image with its distorted versions to form $chunk^k$:

$$chunk^k = [i^k, i^k_1, i^k_2, ..., i^k_n] \tag{7}$$

where i^k_j is the j^{th} distorted version of i^k, and n is the number of augmentations drawn from the bank. We then generate two random crops of $chunk^k$, namely $chunk^k_{c1}$

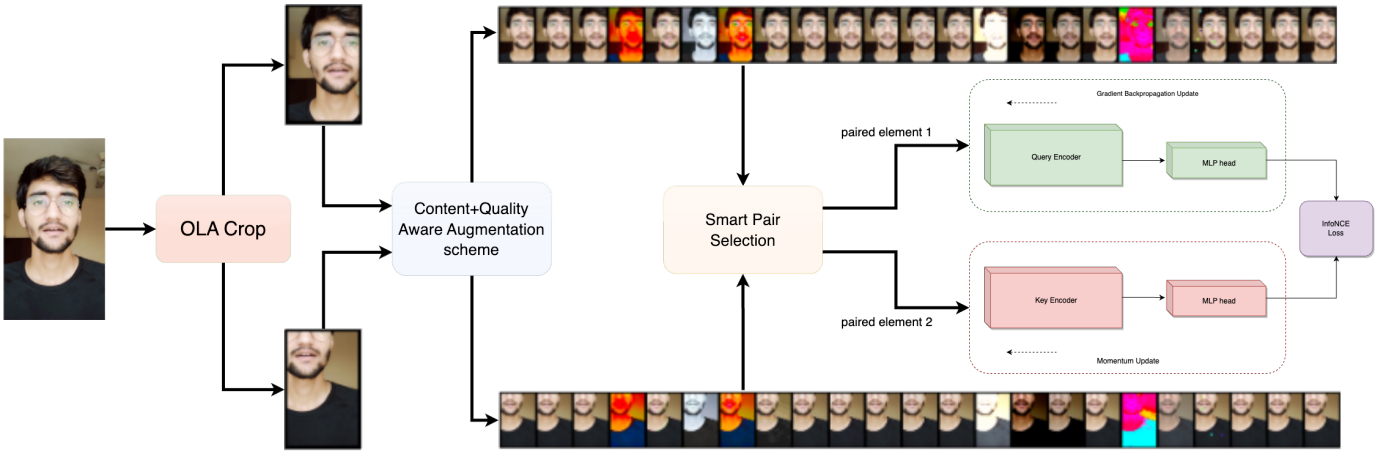


Fig. 5: MoEVA Deep learning module Pre-training scheme

and $chunk^{k}_{c2}$, using an overlap area-based smart cropping mechanism. We choose these crop locations such that the overlapping area (OLA) in the two crops falls within the minimum and the maximum bound of our choosing. We make sure that the crop location is the same over all images in each chunk and different between chunks, resulting in:

$$\begin{aligned} chunk^{k_{c1}} &= [i^{k_{c1}}, {i_1}^{k_{c1}}, {i_2}^{k_{c1}}, ..., {i_n}^{k_{c1}}] \\ chunk^{k_{c2}} &= [i^{k_{c2}}, {i_1}^{k_{c2}}, {i_2}^{k_{c2}}, ..., {i_n}^{k_{c2}}] \end{aligned} \tag{8}$$

after generating the above augmentations we carefully pair and label them as follows:

$$\begin{aligned} [i_m^{k_{c1}}, i_m^{k_{c2}}] &\mapsto similar/same-quality \\ [i_m^{k_{c1}}, i_l^{k_{c2}}] &\mapsto different-quality \\ [i_m^{k_{c1}}, i_l^{k_{c1}}] &\mapsto different-quality \\ [i_m^{k_{c1}}, i_l^{j_{c2}}] &\mapsto different-quality \end{aligned}$$

C. Contrastive Pre-training

We begin by defining two identical encoders 1) Online Encoder (O) and 2) Momentum Encoder (M). Both the encoders are ResNet-50 backbones with an MLP head to generate the final output embedding. We split the pairs we designed in step 1 and pass the first image in the pair through O and the other through M. To calculate the loss between the representation generated by O and M, we use the InfoNCE [**?**] loss function:

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q.k^+/\tau)}{\exp\left(q.k^+/\tau\right) + \sum_{k^-} \exp(q.k^-/\tau)} \tag{9}$$

Here q is the query image, k^+ is a positive sample (similar/same-quality), and k^- are the representations for negative samples (different-quality). τ is a temperature hyper-parameter. This loss is then used to update the weights of O by backpropagation. The weights of M are updated using the weighted sum of the previous weights of M and the new weights of O. Formally denoting the parameters of O as θ_O and parameters of M as θ_M, we update θ_M as:

$$\theta_M \leftarrow m\theta_M + (1-m)\theta_O \tag{10}$$

Here $m \in [0, 1)$, is the momentum coefficient. Once the encoder pre-training has saturated the frozen ResNet-50 encoder weights of the Online encoder O can be used for any downstream task associated with perceptual image quality.

D. VQA Regression

We create the video representative features for each video by average pooling the image representative features generated for each frame using the pre-trained encoder mentioned in the previous subsection. The video representative features are then concatenated with the spatial and temporal NSS features which are then used to train a SVR head to map the collected features to the corresponding MOS.

As we are dealing with videos, it would be more appropriate to apply some kind of temporal pooling instead of a simple average pooling while computing the video representative features from the image representative features. To test this theory we implement temporal pooling using a GRU similar to CONVIQT[] which uses the CONTRIQUE backbone with a GRU. We drop this method as there was no performance boost when compared to simple average pooling. We believe this is largely because temporal quality aspects are already being represented heavily by the temporal NSS features.

E. Experimental results and discussion

We evaluated the our model in the same way as we evaluated the other algorithms. The specialized encoder performs drastically better than RAPIQUE's naive encoder, which is evident from the results. For fair comparison we also evaluate CONTRIQUE, which is basically another such specialized backbone, with the spatial and temporal NSS features. Although CONTRIQUE enjoys a boost when paired with the NSS features, it still falls significantly short when compared to our method.

This is a clear indication that existing UGC-VQA algorithms that have performed competitively well on prior UGC-VQA datasets are clearly suffering from a dataset bias. The uniqueness of the LIVE-ShareChat UGC-VQA database displays the real picture where we can observe that the existing

state-of-the-art algorithms are not sufficient enough to be generalized to all UGC cases.

VI. Conclusion and Future Work

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] 'India set to surpass 1 billion Internet users, 400 bn online spending by 2030: Report' https://government.economictimes.indiatimes.com/news/technology/india-set-to-surpass-1-billion-internet-users-400-bn-online-spending-by-2030-report/96239734, 2022, [Online; accessed 28-December-2022]

PLACE PHOTO HERE

Sandeep Mishra Biography text here.

PLACE PHOTO HERE

Alan C. Bovik Biography text here.