Scaling and Masking: A New Paradigm of Data Sampling for Image and Video Quality Assessment

Yongxu Liu[1,2*], Yinghui Quan[2,1], Guoyao Xiao[2,1], Aobo Li[3], Jinjian Wu[3]

[1]Hangzhou Institute of Technology, Xidian University, Hangzhou 311200, China
[2]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[3]School of Artificial Intelligence, Xidian University, Xi'an 710071, China
{yongxu.liu, gyxiao}@xidian.edu.cn, {yhquan, jinjian.wu}@mail.xidian.edu.cn, abli@stu.xidian.edu.cn

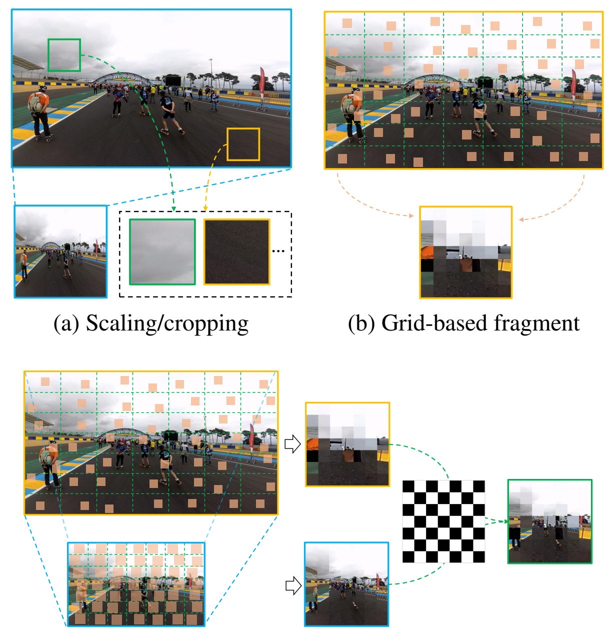
Abstract

Quality assessment of images and videos emphasizes both local details and global semantics, whereas general data sampling methods (e.g., resizing, cropping or grid-based fragment) fail to catch them simultaneously. To address the deficiency, current approaches have to adopt multi-branch models and take as input the multi-resolution data, which burdens the model complexity. In this work, instead of stacking up models, a more elegant data sampling method (named as SAMA, scaling and masking) is explored, which compacts both the local and global content in a regular input size. The basic idea is to scale the data into a pyramid first, and reduce the pyramid into a regular data dimension with a masking strategy. Benefiting from the spatial and temporal redundancy in images and videos, the processed data maintains the multi-scale characteristics with a regular input size, thus can be processed by a single-branch model. We verify the sampling method in image and video quality assessment. Experiments show that our sampling method can improve the performance of current single-branch models significantly, and achieves competitive performance to the multi-branch models without extra model complexity. The source code will be available at https://github.com/Sissuire/SAMA.

Introduction

Image/video quality assessment (I/VQA) is to quantify the perceptual quality/feeling of the given image/video data from the perspective of users (Wu et al. 2019a; Liu et al. 2022b; Wu et al. 2023c,d,b). The best choice of measuring perceptual quality is to conduct a subjective experiment and collect the mean opinion scores (MOS) from a group of subjects (Fang et al. 2022; Jiang et al. 2022). However, the huge cost and inefficiency of the subjective quality assessment cannot satisfy most demands in reality, thus pushing the development of I/VQA methods. As user-generated content (e.g., selfies, short videos) is dramatically overflowing in everyone's mobile phone, every platform (such as TikTok and Facebook), and the whole internet, I/VQA serves as the criterion for quality monitoring, data screening or preference recommendation, and is becoming more active in the society of multimedia and computer vision (Zhu et al. 2020; Wang, Chan, and Loy 2023; Zhang et al. 2023b).

*Corresponding author.



(a) Scaling/cropping (b) Grid-based fragment

(c) The proposed sampling method

Figure 1: An illustration of data sampling methods in quality assessment. Scaling would cause detail loss, while cropping might harm global perception. The proposed method scales the data into a pyramid and masks the pyramid based on data redundancy. The resulting data holds the multi-scale nature with a regular input size.

The intrinsic nature of quality assessment implies that the sampled data prepared for the quality assessment model should well preserve both local details and global appearance, which is not trivial for the fixed input size in deep learning-based models. Before inputting the data into the quality assessment model, a natural process is to interpolate the data into a fixed size (e.g., 224×224). However, the scaling method would introduce the loss of local details, which is inappropriate for I/VQA. Previous work has suggested that preserving the higher resolution of the input data is critical (Hosu et al. 2020). As a consequence, cropping-based sampling is widely used in both IQA and VQA (Ying et al. 2020, 2021). The concern is when facing an image with 1920×1080 or a higher resolution, the cropped patch might be too local to represent the general quality of the image. As shown in Fig. 1(a), the cropped patches could be meaningless (too flat). But more importantly, the local details cannot reflect the global semantics, which is also a key factor in I/VQA (Li et al. 2018; Li, Jiang, and Jiang 2019). Recently, FAST-VQA (Wu et al. 2022, 2023a) extends the patch/cropping-based method and proposes a grid-based fragment sampling. Fragments are sampled at the raw resolution within each grid to preserve the local details (as shown in Fig. 1(b)). The SwinTransformer (Liu et al. 2021, 2022c) is adopted to connect fragments and infer the overall semantics as well as perceptual quality. Still, it is not trivial to put all the global perception onto the model learning with limited fragments, especially for VQA where the number of annotated data is much less than the recognition task.

A natural solution to address the local and global trade-off is to build multi-branch models, where one for the raw fragments to hold the local details, and the other for the scaled input to see the global semantics (Wang et al. 2021; Wu et al. 2023e; Zhao et al. 2023b). However, The multi-branch schemes introduce extra model complexity. MUSIQ (Ke et al. 2021) transforms the input data into a multi-scale image representation, and feeds them all into the model. Although the single-branch model is adopted, the dimension of input data is increased due to the multi-scale sampling, leading to a quadratic growth of the model complexity in attention operation. In summary, both of them introduce much more computational complexity. So, the question is, can it be more elegant and sacrifice less computation to reach a good balance of both local preservation and global perception?

In this work, we propose a new paradigm of data sampling for I/VQA, named as SAMA (i.e., scaling and masking), which is highly motivated by the previous grid-based fragment (Wu et al. 2022), multi-scale sampling (Ke et al. 2021) and masking strategy (He et al. 2022). As shown in Fig. 1(c), the core idea is to build a multi-scale pyramid first to hold the multi-granularity image representation from local to global, then mask the pyramid based on spatial/temporal redundancy (corresponding to image/video) into a regular input size. This masking strategy has a similar tuning/rearranging function as Bayer filter (Maschke 2013) does in imaging sensors, filtering different scaled data in a given pattern. The resulting data is scale-interlaced and shares the same size as the regular sampled one, thus can be fed into a single-branch base model. With a baseline model, we verify our proposed sampling method in both image and video quality assessment databases. Further, a branch of relative scale embedding methods are also explored to examine the consistency.

The main contributions can be summarized as follows:

- A novel data sampling method based on scaling and masking is introduced, to address the paradox of multi-granularity perception with the regular dimension.
- A group of relative scale encoding methods are specifically explored for the manipulated data structure.
- The method is verified on both IQA and VQA databases, and experimental results are impressive with almost negligible computation burden.

Related Work

In this section, we give a brief view of both IQA and VQA from the perspective of data sampling.

Blind IQA Blind IQA is to estimate the quality of images without reference information, which is broadly explored in recent years. In the early stage, hand-crafted methods (Saad, Bovik, and Charrier 2012; Wu et al. 2019b) are unaware of the input resolution, and the features are mostly extracted from the raw data. As deep models are becoming prevalent, data sampling seems to be more important. Various methods crop the image into multiple patches and label each patch with the overall MOS (Kim and Lee 2016; Kim, Nguyen, and Lee 2018; Su et al. 2020; Ying et al. 2020). The patch-based sampling causes an incomplete representation (lack of global view), an inaccurate annotation (the overall MOS for each patch), and a costing inference process where multiple patches would be sampled for the inference. To catch the global perception, some work would like to rescale the data first and then crop it into a fixed size (Hosu et al. 2020; Pan et al. 2022a; Song et al. 2022; Zhao et al. 2023a). However, the resize would definitely cause detail loss, which is not friendly to quality assessment. Some other work purely fed a multi-scale image representation (Ke et al. 2021), which significantly increases the computational burden. A similar way is to build the multi-scale module (Wu et al. 2020) or simply multi-branch models to solve the different views of data (Li et al. 2022b; Pan et al. 2022b; Saha, Mishra, and Bovik 2023). Different from existing IQA methods, our proposed sampling method can hold the local and global views simultaneously. Compared with the single-branch models, our input data contains more complete representation, while compared with the multi-branch ones, the proposed method does not require extra model complexity.

Blind VQA Similar to blind IQA above, blind VQA can also be classified into the raw resolution-based, rescaling-based, cropping-based, and multi-branch methods. The raw resolution is generally adopted in traditional methods (Saad, Bovik, and Charrier 2014; Korhonen 2019). Some work adopts pretrained encoders to extract features so that the raw resolution can be preserved (Li, Jiang, and Jiang 2019). However, the fixed encoders cannot be optimized for VQA task, which makes the features unrepresentative. PVQ (Ying et al. 2021) trains the model with video patches, but the same problem would be faced like the patch-based sampling in IQA. The recent FAST-VQA (Wu et al. 2022, 2023a) proposes to sample grid-based fragments and inference the global perception via SwinTransformer. The method partially relieves the local and global trade-off, but burdens the model learning. When the raw resolution increases, the global inference is thought to be harder. Our method is greatly inspired by the fragment sampling, and moves a step further. We introduce a pyramid of fragments to preserve the local and global views, and propose the masking strategy to reduce the pyramid to the same input size as FAST-VQA. Recently, some multi-branch models are proposed, and they generally construct two streams for both the local and global perception (Zhao et al. 2023b; Wu et al. 2023e,f). The performance is promising, but the complexity is also increasing.

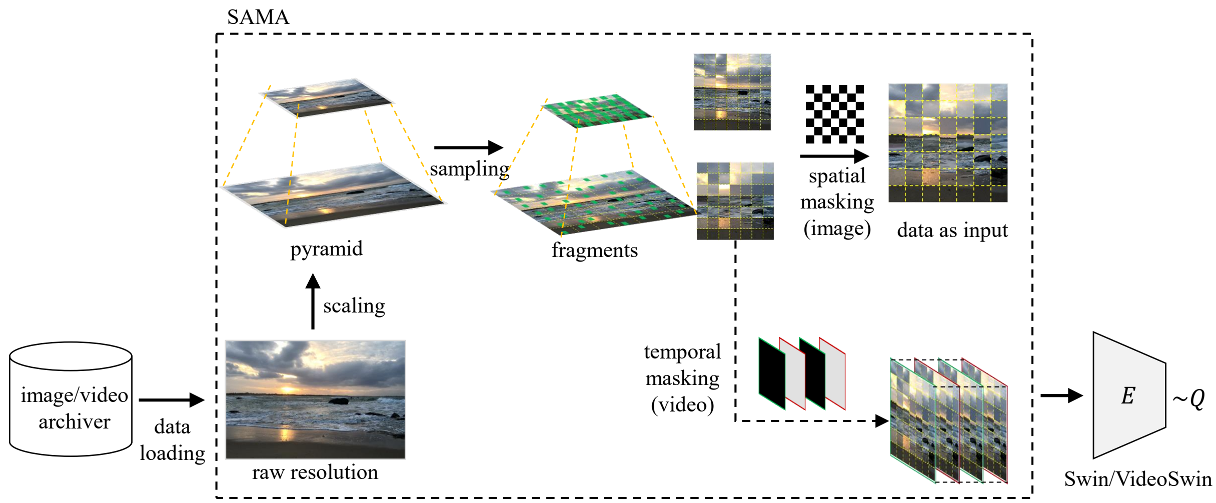


Figure 2: The workflow of SAMA. Image or video data is first scaled into a multi-granularity pyramid via interpolation. Then fragments are sampled in each scale. Afterwards, spatial/temporal masking is constructed to tune the hierarchical fragments into a regular sampling size. The data after SAMA is fed into a base model for quality estimation.

Proposed Method

Overall Architecture

The goal of this work is to explore a more elegant data sampling method to achieve the multi-granularity perception in the layer of input data without extra dimension increase. With the data representation, a single-branch model could be easier to capture the local details and global semantics, and achieve similar performance as multi-branch models but with less complexity. To guarantee the multi-granularity, the data would be scaled into the multi-scale pyramid at first. In order to achieve a regular input size, the dimension of the pyramid is expected to be reduced in an appropriate way.

In this work, motivated by the content redundancy in image/video (Bhaskaran and Konstantinides 1997) and masking strategy (He et al. 2022), scales in the pyramid are masked and interlaced into a regular input size. The workflow is shown in Fig. 2. Our proposed SAMA is based on fragment sampling (Wu et al. 2022), and includes three key steps: scaling, sampling, and masking. After loading the raw resolution data, a multi-granularity pyramid is built via scaling. Then, the grid-based fragment sampling is adopted to extract local content for each scale, and forms a pyramid of fragments. Afterwards, we mask and tune the fragments in different scales based on spatial and temporal redundancy, and obtain the sampled data as model input. the sampled data is directly fed into a baseline model (Wu et al. 2022; Liu et al. 2021, 2022c) to predict the perceptual quality.

Grid-based Fragment Sampling

As SAMA is based on the fragment sampling (Wu et al. 2022), we first give a short review in this subsection.

Fragment sampling is explored initially for VQA, but the core idea is also suitable for IQA or other tasks. Given the data x (image or video) with the raw resolution, fragment sampling first segments x into $G_h \times G_w$ grids. Within each grid, a $f_h \times f_w$ patch at raw resolution is sampled. Afterwards, the data x of any size could be sampled as $\hat{x}$ with a fixed size of $S = (G_h \cdot f_h) \times (G_w \cdot f_w)$. We omit the temporal dimension for simplicity here. In the work, the size of the grid is set as 7×7 and each patch is 32×32, thus $\hat{x}$ is with the size of $S = 224 \times 224$. With the fragment sampling, the model input $\hat{x}$ preserves the raw details and a uniform distribution among the spatial content due to the grid partition. Besides, the discontinuity among near patches could be dissolved by Transformer-based architectures.

The fragment sampling preserves the local details of raw data, but puts the burden of global perception on model inference. When the resolution is increasing, the model might be hard to reason the global semantics from the sampled fragments. In this work, we construct a scale pyramid to relieve the dilemma and achieve a good trade-off of both local preservation and global perception.

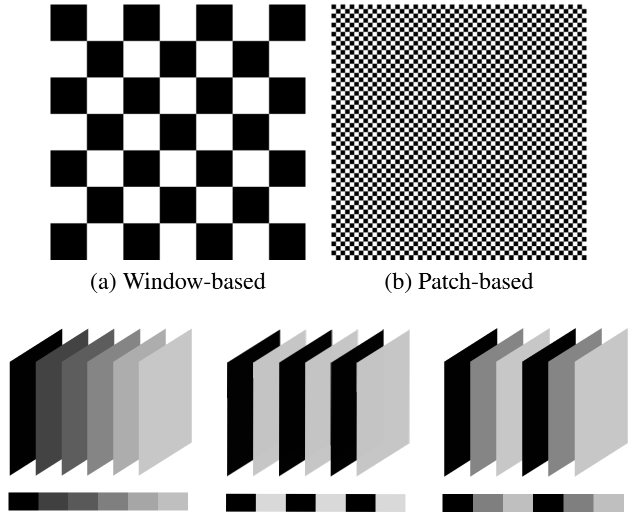
Scaling and Masking in Image

For simplicity, the proposed SAMA is first clarified for image, then it would be extended into video in the following subsection. Considering the great burden of model learning in fragment sampling, we are motivated by MUSIQ ((Ke et al. 2021)), an IQA method with multi-scale image representation, and introduce the multi-granularity pyramid to hold more data information from local to global.

Given the image data x, a pyramid p_x is first built as

$$p_x = \{x_0, x_1, \cdots, x_n\}, \quad (1)$$

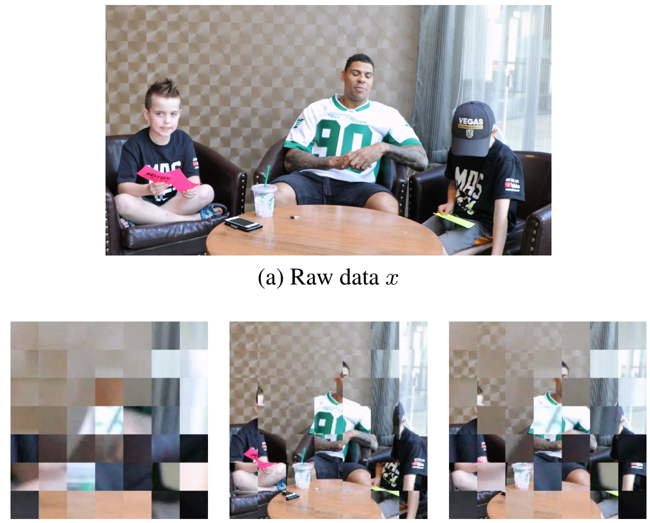
where x_0 is the same as x, and x_i $(i > 0)$ is a downsampled version of x through bilinear interpolation but still keeps the aspect ratio of the image. We denote x_n as the image with the smallest resolution satisfying $\min(h_{x_n}, w_{x_n}) = \min(G_h \cdot f_h, G_w \cdot f_w)$. where h_{x_n}, w_{x_n} is the height and width of x_n. In practice, we uniformly scale the image thus the resolution of x_i in p_x decreases linearly from x_0 to x_n.



(a) Window-based (b) Patch-based

(c) Progressive (d) Choppy (e) Mixed

Figure 3: The illustration of spatial and temporal masks. (a) and (b) are spatial masks for images, and the last three are temporal masks for videos. Different intensities indicate the different scales.



(a) Raw data x

(b) Fragments on x_0 (c) Fragments on x_n (d) SAMA result

Figure 4: An example of sampling result for image.

With the multi-granularity pyramid, the fragment sampling is adopted to extract content and converts the pyramid with various resolutions into the representation with a fixed size (e.g., 224×224). Thus, a pyramid of fragments is constructed as

$$\hat{p_x} = \{\hat{x_0}, \hat{x_1}, \cdots, \hat{x_n}\}, \tag{2}$$

where $\hat{x_i}$ is the fragment sampled from x_i with the resolution of S. Note that, although each $\hat{x_i}$ shares the same resolution, fragments among different $\hat{x_i}$ are sampled from different scales in the multi-granularity pyramid, thus the fragments $\hat{p_x}$ can be also seen as a pyramid.

The pyramid of fragments dramatically increases the data size from S to $(n+1) \times S$. Some methods directly fed the pyramid into the model, thus increasing the complexity linearly or quadratically. The previous study has shown that the image content is largely redundant and thus can be greatly compressed (Bhaskaran and Konstantinides 1997). Further, the recent masked autoencoder also suggests that even if the image content is masked with a large ratio, the content could still be recognized (He et al. 2022). The work has greatly inspired us to introduce the masking strategy and reduce the dimension of pyramid to a regular input size of S.

In the process of mask designing, we also refer to the Bayer pattern in imaging sensors, where the R/G/B is staggered with a certain arrangement. In this work, we introduce a scale-interlaced masking pattern that tunes the pyramid and transforms it into a regular input. Fig. 3(a) and (b) give the illustration of the spatial masks for images. For simplicity, we only adopt two scales here, thus $n = 1$. In Fig. 3, we use different intensities of luminance to indicate the scales. Thus, in (a) and (b), the black positions correspond to the fragments sampled from x_0 while the white indicates those from x_1. When $n = 1$, the data can be formulated as

$$\tilde{x} = M \cdot \hat{x_0} + (1 - M) \cdot \hat{x_1}, \tag{3}$$

where M is the spatial binary mask, in which 1/0 indicates the raw/scaled fragments.

In Fig. 3, we provide two types of spatial masks. As our work is built on the fragment sampling, we choose the same baseline model, SwinTransformer (Liu et al. 2021), for quality assessment. According to the model architecture, we construct a window-based mask and a patch-based mask for images. For the window-based mask, the size of each black-/white region is equal to the window size (i.e., 32×32) in SwinTransformer. Similarly, the size of each black/white region in the patch-based mask is the same as the embedding patch (i.e., 4×4). Note that the mask can be extended to adapt to more scales. A very familiar way is to imitate Bayer RGGB in imaging sensors or the diamond arrangement in display that staggers multiple scales with a regular size.

Fig. 4 gives an example where (b) shows the fragments sampled from raw resolution, and (c) shows the ones from scaled data. It shows that fragments sampled from x_0 focus more on the local details, and are hard to infer the global semantics. The fragments sampled from x_n contain more global information, but miss the local distortions. With the masking strategy, the proposed SAMA masks and rearranges the multi-scale data into a regular shape, holding both local details and global semantics in a regular data shape.

SAMA for Video

When it comes to videos, the scaling and sampling are similar. The main difference is in the masking. Video data preserves not only the spatial correlation, but rich temporal redundancy. Thus, besides the spatial masks, we can explore more temporal masks for video.

Given a video data v containing T frames, the scaling procedure transforms v into the scale space p_v, and the sampling extracts content from p_v to generate a pyramid of fragments $\hat{p_v}$. In the process of masking, we introduce another three temporal masks based on temporal redundancy. As illustrated in Fig. 3(c)$\sim$(e), a progressive mask indicates that, along the time dimension, the fragments are selected from the finest to the coarsest gradually. Considering the pyramid of fragments $\hat{p_v}$, where each scale actually contains T frames since fragment sampling would not reduce the temporal dimension. What progressive mask does is to select the first two frames of fragments with raw resolution as $\tilde{v}_0$, and select the next two frames of fragments with a coarser resolution from the pyramid as $\tilde{v}_1$, until the coarsest. Thus, we can obtain the new data formed as

$$\tilde{v} = \{\tilde{v}_0, \tilde{v}_1, \cdots, \tilde{v}_{\frac{T}{2}-1}\}, \tag{4}$$

where each $\tilde{v}_i$ contains two frames of scaled fragments from the pyramid. The two-frame setting results from the selection of our baseline model (VideoSwin), where the temporal dimension of embedding patch is 2. As VideoSwin model takes as input 32 frames, thus the input data contains 16 scales of fragments, where the first two frames are the finest resolution, and the last two are the coarsest.

Besides the progressive masking, we also provide a choppy masking and a mixed one. As shown in Fig. 3(d), the data form contains only two scales and they are interlaced in the temporal dimension. Still, each scale also contains two frames of fragments. For a 32-frame input data, the first two frames are the finest scale, and the next two frames are the coarsest. The arrangement would repeat 8 times to get a 32-frame input data. And for the mixed mask, the first 16 frames follow a progressive procedure, and the arrangement would repeat it to obtain 32 frames.

Model Architecture and Implementation Details

Our work serves as the data sampling method. Afterwards, the data is fed into a baseline model. Our implementation is based on FAST-VQA (Wu et al. 2022), and adopts the SwinTransformer/VideoSwin (Liu et al. 2021, 2022c) for image/video tasks. For VQA, we simply substitute the original sampling method in FAST-VQA with SAMA, and keep all the other settings (e.g., loss function, learning rate, batch size, etc.) as default. More details of the implementation can refer to the work (Wu et al. 2022) or our appendix.

Given the input size $S = H \times W$ for images, considering a 4×4 patch embedding, the output of the baseline model, denoted as $z = f(\tilde{x})$, is $\frac{H}{32} \times \frac{W}{32} \times C$, where H, W, and C are the height, width, and channel dimension. Two simple fully-connected (FC) layers are followed (64 hidden nodes) to regress z into quality scores q with the size of $\frac{H}{32} \times \frac{W}{32} \times 1$. The overall quality is obtained by averaging q on the spatial.

As for videos, the size of prepared data is $S = H \times W \times T$, where T is the temporal dimension. Given a $4 \times 4 \times 2$ embedding patch, the size of z is $\frac{H}{32} \times \frac{W}{32} \times \frac{T}{2} \times C$. Similarly, we also adopt two FC layers for the regression, and obtain the scores q with the size of $\frac{H}{32} \times \frac{W}{32} \times \frac{T}{2} \times 1$. The global average is adopted to get the overall quality.

Relative Scale Encoding

The only concern is whether the model can distinguish the scale-mixed data. To clarify the concern, we explore various relative scale encoding methods in the model to explicitly highlight the scale information. For simplicity, we select the video data with the progressive temporal mask for example.

SAMA-W Instead of directly averaging q, SAMA-W introduces extra FC layers to learn $\frac{T}{2}$ weights in temporal dimension (corresponding to different scales), and compute a weighted average for the final quality.

SAMA-SE In this variation, the SE module (Hu, Shen, and Sun 2018) is introduced in the transformer block. The SE module is set after the attention operation but before the residual connection in each transformer block. Still, we only focus on the temporal dimension where the features are squeezed and excited with a regular SE module.

SAMA-RSB-A The relative scale bias is explored in the transformer block. In the original SwinTransformer, the relative position bias is embedded in the attention module as

$$Attn(Q, K, V) = \mathrm{SoftMax}(QK^T/\sqrt{d} + B)V, \tag{5}$$

where Q, K, V, B, and d are query, key, value, relative position bias, and channel dimension (Liu et al. 2021). In order to tune the scale information adaptively, an additional relative scale bias (denoted as R) is introduced, and the attention is modified as

$$Attn(Q, K, V) = \mathrm{SoftMax}(QK^T/\sqrt{d} + B + R)V. \tag{6}$$

SAMA-RSB-M The relative scale bias R can be also introduced as a multiplication factor instead of the addition term, which is formulated as

$$Attn(Q, K, V) = \mathrm{SoftMax}((QK^T/\sqrt{d} + B) \cdot R)V. \tag{7}$$

Note that these encoding methods are not included in our basic SAMA method. With SAMA, we just explore some variations in model architecture based on data characteristics, and see if both of them are cooperatively enhanced.

Experiments

Benchmark Datasets and Evaluation Protocols

We evaluate our proposed method on both VQA and IQA tasks. For VQA task, four datasets (i.e., LSVQ (Ying et al. 2021), KoNViD (Hosu et al. 2017), LIVE-VQC (Sinno and Bovik 2018) and YouTube-UGC (Wang, Inguva, and Adsumilli 2019)) are adopted to conduct the intra- and inter-dataset validation. For IQA task, we adopt two widely-used datasets (i.e., KonIQ (Hosu et al. 2020), and SPAQ (Fang et al. 2020)). By convention, when trained on LSVQ, we report the performance results on both LSVQ test set, LSVQ 1080p set, KoNViD, and LIVE-VQC. When conducting the intra-dataset validation for both VQA and IQA, the experiment is repeated for 10 rounds and the median performance is adopted to alleviate the randomness. To quantify the performance of different methods, Spearman rank-order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC) are adopted for the criteria. A better model would result in higher SRCC and PLCC values.

Experiments on VQA

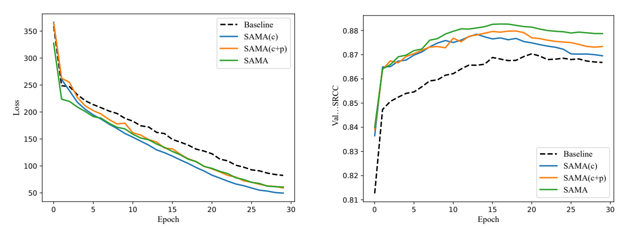
Our proposed sampling method is verified with 9 newly released VQA algorithms, which are TLVQM (Korhonen 2019), VIDEVAL (Tu et al. 2021), VSFA (Li, Jiang,

Trained on LSVQ				LSVQ		LSVQ-1080p		KoNViD		LIVE-VQC	
Type	Method	Year	Publication	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
hand-crafted	TLVQM	2019	T-IP	0.772	0.774	0.589	0.616	0.732	0.724	0.670	0.691
	VIDEVAL	2021	T-IP	0.795	0.783	0.545	0.554	0.751	0.741	0.630	0.640
multi-branch	PVQ (w/o patch)	2021	CVPR	0.814	0.816	0.686	0.708	0.781	0.781	0.747	0.776
	PVQ (w patch)	2021	CVPR	0.827	0.828	0.711	0.739	0.791	0.795	0.770	0.807
	BVQA-2022	2022	T-CSVT	0.852	0.855	0.771	0.782	0.834	0.837	0.816	0.824
	DSD-PRO	2022	T-CSVT	0.875	0.875	0.765	0.793	0.865	0.856	**0.838**	0.849
	ZoomVQA	2023	CVPRW	0.886	0.879	0.799	0.819	0.877	0.875	0.814	0.833
	DOVER	2023	ICCV	**0.888**	**0.889**	**0.795**	**0.830**	**0.884**	**0.883**	0.832	**0.855**
single-branch	VSFA	2019	MM	0.801	0.796	0.675	0.704	0.784	0.795	0.734	0.772
	FAST-VQA	2022	ECCV	0.876	0.877	0.779	0.814	0.859	0.855	0.823	0.844
	Baseline			0.870	0.871	0.766	0.801	0.860	0.855	0.816	0.836
	SAMA (c)			0.879	0.880	**0.785**	**0.825**	0.879	0.877	0.816	0.839
	SAMA (c + p)			0.880	0.881	0.783	0.821	0.879	0.877	0.821	0.838
	SAMA			**0.883**	**0.884**	0.782	0.822	**0.880**	**0.877**	**0.834**	**0.849**
	SAMA (+ spm)			0.878	0.878	0.776	0.814	0.872	0.872	0.818	0.840
	SAMA(+ swm)			0.877	0.878	0.780	0.820	0.873	0.867	0.815	0.836

Table 1: VQA Performance Comparison when Trained on LSVQ (The default SAMA is with progressive temporal mask, while 'c'/'c+p'/'spm'/'swm' indicates choppy/choppy+progressive/spatial patch-based/spatial window-based mask, respectively.)

and Jiang 2019), Patch-VQ (Ying et al. 2021), BVQA-2022 (Li et al. 2022a), DSD-PRO (Liu et al. 2022a), FAST-VQA (Wu et al. 2022), ZoomVQA (Zhao et al. 2023b), and DOVER (Wu et al. 2023e)). Following the setting in FAST-VQA, models are first trained on LSVQ training set, and the performance is validated on LSVQ test set, LSVQ 1080p set, KoNViD, and LIVE-VQC. The performances of existing methods are directly from the published papers, and the comparison is shown in Tab. 1. We reproduce FAST-VQA as the baseline model in our environment. The default SAMA model is built on the progressive temporal masking. We denote the one with choppy temporal mask as SAMA(c), and the one with mixed mask as SAMA(c+p). We also introduce two more with spatiotemporal masks, where 'spm'/'swm' means spatial patch-based/window-based mask respectively.

As given in Tab. 1, with the proposed sampling, our models surpass the baseline model and achieve good results. Since the only difference between the baseline and our models is the proposed sampling method, the improvement demonstrates the effectiveness of our proposed SAMA method. Further, Fig. 5 gives the training loss and the testing SRCC on LSVQ dataset. With SAMA, our models reach lower training errors, and higher testing SRCC. It is also observed in Tab. 1 that the mixture of spatial and temporal masks may decrease the performance, which is reasonable because the complex pattern might make it harder to learn a good representation, especially in the small-scale dataset. In the table, the best performances in single-branch and multi-branch architectures are highlighted in bold. Note that we weaken the multi-branch models, because our work is focusing the sampling method. Even so, with SAMA, our single-branch models still achieve similar performance to the recent multi-branch models (mixture of sampling or dual-stream decomposition).



(a) Train loss (b) Test SRCC

Figure 5: The illustration of training and testing on LSVQ

Moreover, some relative scale encoding methods are also explored in Tab. 2. From the table, it shows that the additional encoding modules in the architecture do not help much, with little performance improvement. The result may also suggest that the input data from SAMA can be well recognized and inferred by the model architecture, thus additional relative scale encoding modules are redundant. This result is helpful. It indicates that we do not need to change any of the model, and only a simple replacement of the data sampling can improve the performance.

The intra-dataset performance comparison of VQA is also conducted on LIVE-VQC, KoNViD, and YouTube-UGC. As shown in Tab. 3, our default model greatly enhances the baseline model (i.e., the reproduced FAST-VQA). The proposed method is based on the single-branch model, and achieves competitive performance to multi-branch methods with the change of sampling only.

Experiments on IQA

In this subsection, we verify the proposed method on IQA task. We adopt 13 IQA algorithms in the recent three years

Method	LSVQ SRCC	LSVQ PLCC	LSVQ-1080p SRCC	LSVQ-1080p PLCC	KoNViD SRCC	KoNViD PLCC	LIVE-VQC SRCC	LIVE-VQC PLCC
Baseline	0.870	0.871	0.766	0.801	0.860	0.855	0.816	0.836
SAMA	0.883	0.884	0.782	0.822	0.880	0.877	0.834	0.849
SAMA-W	0.882	0.883	0.783	0.822	0.880	0.874	0.833	0.851
SAMA-SE	0.881	0.882	0.785	0.824	0.879	0.878	0.834	0.846
SAMA-RSB-A	0.881	0.883	0.783	0.823	0.877	0.875	0.834	0.849
SAMA-RSB-M	0.883	0.884	0.787	0.825	0.879	0.875	0.835	0.848

Table 2: VQA Performance with Relative Scale Encoding

Method	LIVE-VQC SRCC	LIVE-VQC PLCC	KoNViD SRCC	KoNViD PLCC	YouTube-UGC SRCC	YouTube-UGC PLCC
TLVQM	0.799	0.803	0.773	0.768	0.669	0.659
VIDEVAL	0.752	0.751	0.783	0.780	0.779	0.773
PVQ	0.827	0.837	0.791	0.786	-	-
BVQA-2022	0.834	0.842	0.834	0.836	0.818	0.826
DSD-PRO	**0.875**	0.862	0.861	0.861	0.841	0.836
DOVER	0.860	**0.875**	**0.909**	**0.906**	**0.890**	**0.891**
VSFA	0.773	0.795	0.773	0.775	0.724	0.743
FAST-VQA	0.849	0.862	0.891	0.892	0.855	0.852
Baseline	0.847	0.867	0.881	0.880	0.842	0.840
SAMA	**0.860**	**0.878**	**0.892**	**0.892**	**0.881**	**0.880**

Table 3: VQA Performance within Each Dataset

for the performance comparison, which are HyperIQA (Su et al. 2020), MUSIQ (Ke et al. 2021), IEIT (Song et al. 2022), CVC-IQA (Zhu et al. 2022), VT-IQA (Li et al. 2022b), DACNN (Pan et al. 2022b), SARQUEsarque2022, CONTRIQUE (Madhusudana et al. 2022), VCRNet (Pan et al. 2022a), DEIQT (Qin et al. 2023), Re-IQA (Saha, Mishra, and Bovik 2023), LIQE (Zhang et al. 2023a) and QPT (Zhao et al. 2023a). Our baseline model is based on the naive SwinTransformer. The spatial window-based mask is adopted as our default SAMA method, and the spatial patch-based mask is denoted as 'spm'.

As shown in Tab. 4, the proposed sampling method significantly improves the baseline model, and achieves the best performance on the databases. Note that KonIQ and SPAQ are two different datasets, where KonIQ is mainly the public images while SPAQ is specified to smartphone photography. The improvement on both datasets verifies the proposed sampling method can deal with various image content. More importantly, our baseline is simple enough while existing methods might involves specific pretraining (Madhusudana et al. 2022; Zhao et al. 2023a). The simplicity of SAMA and its promising performance demonstrate the effectiveness of our method. Fig. 6 gives an illustration of how the model estimates quality. In the figure, (b) and (e) are the sampled fragments from the baseline (i.e., grid-based fragment sampling, the left column) and SAMA (the right column), while (c) and (f) are the output responses (i.e., the predicted quality scores on each window, the lower the darker). From the figure, it shows that our SAMA enhances the ability of global perception, thus improving the baseline method.

Complexity and Transferability

SAMA is a data sampling method and does not involves any change in model architecture, thus the model complexity is the same as FAST-VQA (Wu et al. 2022). Indeed, SAMA is thought to cost only more computation for the interpolation when scaling the data into a pyramid. And as for the sampling and masking procedures, the implementation can be simplified and cost the same consumption as FAST-VQA. This is because both of the methods are to sample the same amount of fragments. The only difference is where the data is sampled from – FAST-VQA samples from the raw data, while SAMA samples from the pyramid.

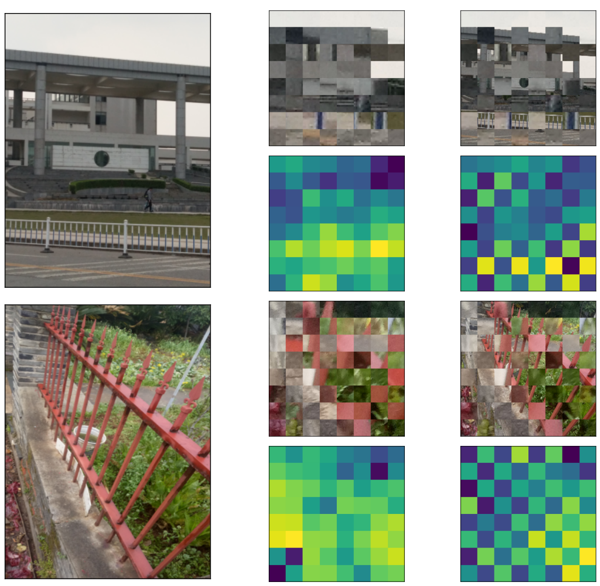
Our SAMA method is expected to cover more tasks and more models. In this work, we verify the method in two resolution-related tasks (i.e., IQA and VQA). Actually, the method should be task-agnostic. These tasks benefiting from multi-scale features are also thought to be the applications of SAMA, and we are expecting to extend SAMA to more tasks. Further, We would also try more model architectures (not only SwinTransformer). The data sampled from SAMA is expected to suit most transformer-based models, which will be examined in our future work.

Conclusion

Various tasks are benefiting from the multi-granularity representation to bold both the local details and the global se-

Method	Year	Publication	KonIQ SRCC	KonIQ PLCC	SPAQ SRCC	SPAQ PLCC
HyperIQA	2020	CVPR	0.906	0.917	0.916	0.919
MUSIQ	2021	ICCV	0.916	0.928	0.917	0.921
IEIT	2022	T-CSVT	0.892	0.916	0.917	0.921
CVC-IQA	2022	T-MM	0.915	0.941	**0.925**	**0.929**
VT-IQA	2022	T-CSVT	0.924	0.938	0.920	0.925
DACNN	2022	T-CSVT	0.901	0.912	0.915	0.921
SARQUE	2022	T-MM	0.901	0.923	0.918	0.922
CONTRIQUE	2022	T-IP	0.894	0.906	0.914	0.919
VCRNet	2022	T-IP	0.894	0.909	-	-
DEIQT	2023	AAAI	0.921	0.934	0.919	0.923
Re-IQA	2023	CVPR	0.923	0.914	0.918	0.925
LIQE	2023	CVPR	0.919	0.908	-	-
QPT	2023	CVPR	0.927	0.941	**0.925**	0.928
Baseline			0.923	0.934	0.908	0.911
SAMA (spm)			0.926	0.940	0.923	0.926
SAMA			**0.930**	**0.942**	**0.925**	**0.929**

Table 4: IQA Performance within Each Dataset



(a) Raw data (b) Baseline (c) SAMA

Figure 6: An illustration of fragments (the first row) and model response (the second row).

mantics, but current data sampling methods cannot balance the local and global paradox simultaneously. The common solution is to increase the input dimension, and thus burden the model complexity. In this work, we have explored a more elegant data sampling method, and put the multi-granularity representation in a regular size with a masking strategy. Experiments on both IQA and VQA have shown that the proposed method is effective with only a single-branch structure. We have also explored various relative scale encoding methods and the results have demonstrated that the proposed SAMA has a good compatibility with the model. We are expecting to apply our method to more basic tasks.

Acknowledgments

This work was supported by Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (GZC20232034), and Postdoctoral Research Grant of Shaanxi Province (2023BSHEDZZ165).

References

Bhaskaran, V.; and Konstantinides, K. 1997. Image and video compression standards: algorithms and architectures.

Fang, Y.; Huang, L.; Yan, J.; Liu, X.; and Liu, Y. 2022. Perceptual quality assessment of omnidirectional images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 580–588.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3677–3686.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hosu, V.; Hahn, F.; Jenadeleh, M.; Lin, H.; Men, H.; Szirányi, T.; Li, S.; and Saupe, D. 2017. The Konstanz natural video database (KoNViD-1k). In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, 1–6. IEEE.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jiang, Q.; Liu, Z.; Gu, K.; Shao, F.; Zhang, X.; Liu, H.; and Lin, W. 2022. Single image super-resolution quality assessment: a real-world dataset, subjective studies, and an objective metric. *IEEE Transactions on Image Processing*, 31: 2279–2294.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Kim, J.; and Lee, S. 2016. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1): 206–220.

Kim, J.; Nguyen, A.-D.; and Lee, S. 2018. Deep CNN-based blind image quality predictor. *IEEE transactions on neural networks and learning systems*, 30(1): 11–24.

Korhonen, J. 2019. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12): 5923–5938.

Li, B.; Zhang, W.; Tian, M.; Zhai, G.; and Wang, X. 2022a. 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*, 32(9): 5944–5958.

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

Li, D.; Jiang, T.; Lin, W.; and Jiang, M. 2018. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia*, 21(5): 1221–1234.

Li, L.; Song, T.; Wu, J.; Dong, W.; Qian, J.; and Shi, G. 2022b. Blind image quality index for authentic distortions with local and global deep feature aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8512–8523.

Liu, Y.; Wu, J.; Li, L.; Dong, W.; and Shi, G. 2022a. Quality Assessment of UGC Videos Based on Decomposition and Recomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3): 1043–1054.

Liu, Y.; Wu, J.; Li, L.; Dong, W.; Zhang, J.; and Shi, G. 2022b. Spatiotemporal representation learning for blind video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6): 3500–3513.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022c. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Madhusudana, P. C.; Birkbeck, N.; Wang, Y.; Adsumilli, B.; and Bovik, A. C. 2022. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31: 4149–4161.

Maschke, T. 2013. *Digitale kameratechnik: technik digitaler kameras in theorie und praxis*. Springer-Verlag.

Pan, Z.; Yuan, F.; Lei, J.; Fang, Y.; Shao, X.; and Kwong, S. 2022a. VCRNet: Visual compensation restoration network for no-reference image quality assessment. *IEEE Transactions on Image Processing*, 31: 1613–1627.

Pan, Z.; Zhang, H.; Lei, J.; Fang, Y.; Shao, X.; Ling, N.; and Kwong, S. 2022b. DACNN: Blind image quality assessment via a distortion-aware convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7518–7531.

Qin, G.; Hu, R.; Liu, Y.; Zheng, X.; Liu, H.; Li, X.; and Zhang, Y. 2023. Data-Efficient Image Quality Assessment with Attention-Panel Decoder. *arXiv preprint arXiv:2304.04952*.

Saad, M. A.; Bovik, A. C.; and Charrier, C. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE transactions on Image Processing*, 21(8): 3339–3352.

Saad, M. A.; Bovik, A. C.; and Charrier, C. 2014. Blind prediction of natural video quality. *IEEE Transactions on image Processing*, 23(3): 1352–1365.

Saha, A.; Mishra, S.; and Bovik, A. C. 2023. Re-IQA: Unsupervised Learning for Image Quality Assessment in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5846–5855.

Sinno, Z.; and Bovik, A. C. 2018. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2): 612–627.

Song, T.; Li, L.; Chen, P.; Liu, H.; and Qian, J. 2022. Blind image quality assessment for authentic distortions by intermediary enhancement and iterative training. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7592–7604.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3667–3676.

Tu, Z.; Wang, Y.; Birkbeck, N.; Adsumilli, B.; and Bovik, A. C. 2021. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30: 4449–4464.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2555–2563.

Wang, Y.; Inguva, S.; and Adsumilli, B. 2019. YouTube UGC dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 1–5. IEEE.

Wang, Y.; Ke, J.; Talebi, H.; Yim, J. G.; Birkbeck, N.; Adsumilli, B.; Milanfar, P.; and Yang, F. 2021. Rich features for perceptual quality assessment of UGC videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13435–13444.

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

Wu, H.; Chen, C.; Liao, L.; Hou, J.; Sun, W.; Yan, Q.; Gu, J.; and Lin, W. 2023a. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–17.

Wu, H.; Chen, C.; Liao, L.; Hou, J.; Sun, W.; Yan, Q.; and Lin, W. 2023b. DisCoVQA: Temporal distortion-content transformers for video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Wu, H.; Liao, L.; Chen, C.; Hou, J. H.; Zhang, E.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023c. Exploring opinion-unaware video quality assessment with semantic affinity criterion. In *Processings of International Conference on Multimedia and Expo (ICME)*.

Wu, H.; Liao, L.; Wang, A.; Chen, C.; Hou, J. H.; Zhang, E.; Sun, W.; Yan, Q.; and Lin, W. 2023d. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment. arXiv:2304.14672.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023e. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *IEEE International Conference on Computer Vision*, 1–8.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J. H.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023f. Towards explainable video quality assessment: A database and a language-prompted approach. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*.

Wu, J.; Liu, Y.; Dong, W.; Shi, G.; and Lin, W. 2019a. Quality assessment for video with degradation along salient trajectories. *IEEE Transactions on Multimedia*, 21(11): 2738–2749.

Wu, J.; Ma, J.; Liang, F.; Dong, W.; Shi, G.; and Lin, W. 2020. End-to-end blind image quality prediction with cascaded deep neural network. *IEEE Transactions on image processing*, 29: 7414–7426.

Wu, J.; Zhang, M.; Li, L.; Dong, W.; Shi, G.; and Lin, W. 2019b. No-reference image quality assessment with visual pattern degradation. *Information sciences*, 504: 487–500.

Ying, Z.; Mandal, M.; Ghadiyaram, D.; and Bovik, A. 2021. Patch-VQ:'Patching Up'the video quality problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14019–14029.

Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; and Bovik, A. 2020. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3575–3585.

Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; and Ma, K. 2023a. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14071–14081.

Zhang, Z.; Wu, W.; Sun, W.; Tu, D.; Lu, W.; Min, X.; Chen, Y.; and Zhai, G. 2023b. MD-VQA: Multi-dimensional quality assessment for UGC live videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1746–1755.

Zhao, K.; Yuan, K.; Sun, M.; Li, M.; and Wen, X. 2023a. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22302–22313.

Zhao, K.; Yuan, K.; Sun, M.; and Wen, X. 2023b. Zoom-VQA: Patches, Frames and Clips Integration for Video Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1302–1310.

Zhu, H.; Li, L.; Wu, J.; Dong, W.; and Shi, G. 2020. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14143–14152.

Zhu, Y.; Li, Y.; Sun, W.; Min, X.; Zhai, G.; and Yang, X. 2022. Blind Image Quality Assessment Via Cross-View Consistency. *IEEE Transactions on Multimedia*, 1–14.