Video Quality Assessment: A Comprehensive Survey

Qi Zheng, Yibo Fan*, Leilei Huang, Tianyu Zhu, Jiaming Liu, Zhijian Hao, Shuo Xing, Chia-Ju Chen, Xiongkuo Min, Alan C. Bovik†, Zhengzhong Tu†

***Abstract*—Video quality assessment (VQA) is an important processing task, aiming at predicting the quality of videos in a manner highly consistent with human judgments of perceived quality. Traditional VQA models based on natural image and/or video statistics, which are inspired both by models of projected images of the real world and by dual models of the human visual system, deliver only limited prediction performances on real-world user-generated content (UGC), as exemplified in recent large-scale VQA databases containing large numbers of diverse video contents crawled from the web. Fortunately, recent advances in deep neural networks and Large Multimodality Models (LMMs) have enabled significant progress in solving this problem, yielding better results than prior handcrafted models. Numerous deep learning-based VQA models have been developed, with progress in this direction driven by the creation of content-diverse, large-scale human-labeled databases that supply ground truth psychometric video quality data. Here, we present a comprehensive survey of recent progress in the development of VQA algorithms and the benchmarking studies and databases that make them possible. We also analyze open research directions on study design and VQA algorithm architectures.**

***Index Terms*—Video quality assessment, subjective quality study, objective quality study, deep learning, technical evolution.**

I. Introduction

RECENT years have witnessed the rapid development of streaming media technologies and platforms, making video content the dominant form of Internet traffic. Streaming and social media videos play a central role in the daily lives of billions of people, in particular since the evolution of Web 2.0 has spurred an explosion of user-generated content (UGC). Moreover, recent advancements in generative AI technologies have fueled the rapid rise of AI-Generated Content (AIGC), enabling seamless content creation and transformation across various media platforms. As such, it has become increasingly important for video service providers to improve the efficiency of the video transcoding techniques they deploy in the Cloud, while also delivering satisfying visual quality of experience (QoE) to customers. Balancing these goals conditioned on constantly changing network conditions remains a crucial and long-standing challenge of core interest. Perceptually designed video quality assessment models, which act as 'judges' of percieved quality, play an important role in monitoring and measuring visual content quality throughout global communication network and video processing chains, particularly in video coding, enhancement, and reconstruction algorithms.

Video quality studies can be broadly categorized into two groups: objective and subjective methods. **Subjective VQA** involves humans in the loop, who view a number of video contents in a controlled environment, then render judgments of perceptual quality on each content. These raw annotations are further normalized and averaged as mean opinion scores (MOS) or as difference mean opinion scores (DMOS). While subjective studies are time and labor-consuming, they provide valuable psychometric datasets on which to benchmark, develop, compare, and calibrate video quality models. **Objective VQA** methods rely on algorithmic models to predict the perceptual quality of video content. According to the information available from reference signals that are input to an algorithm, objective VQA models can be categorized into full-reference (FR), reduced-reference (RR), or no-reference (NR) models. FR VQA models compare visual differences between high-quality reference videos and distorted counterparts, hence way be viewed as perceptual video fidelity models. RR VQA models are similar to FR methods, but only require significantly reduced amounts of signal information to be extracted from the reference videos to conduct quality prediction. In many practical scenarios, however, the reference signals are inaccessible, *e.g.,* videos uploaded by amateur photographers or videographers to social platforms such as YouTube and TikTok. In such instances, NR VQA models are the only tools available to monitor and analyze such authentically distorted contents without any pristine reference.

Modeling the perception of video quality poses significant challenges. Videos are degraded by a very wide variety of distortions, including noise, blur, ringing, banding, compression, and blockiness, which are very different to model, often occur together and interact, and impact perceptual video quality in ways that are also content-dependent. The complexity of the problem is exacerbated by the commingling of multiple distortions, as well as the presence of complex object and

Qi Zheng, Yibo Fan, Tianyu Zhu, Jiaming Liu, and Zhijian Hao are with Fudan University, Shanghai 200000, China (e-mail: qzheng21@m.fudan.edu.cn; fanyibo@fudan.edu.cn; zhuty22@m.fudan.edu.cn; liujm22@m.fudan.edu.cn; zjhao19@fudan.edu.cn;).

Leilei Huang is with East China Normal University, Shanghai 200000, China (e-mail: llhuang@cee.ecnu.edu.cn).

Xiongkuo Min is with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: minxiongkuo@sjtu.edu.cn).

Chia-Ju Chen and Alan C. Bovik is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: ju40268@utexas.edu;bovik@utexas.edu).

Shuo Xing and Zhengzhong Tu are with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77840 (e-mail: shuoxing@tamu.edu;tzz@tamu.edu). This work was done prior to the employment of Zhengzhong Tu by Texas A&M University, and he was not supported by any grant.

*: Corresponding author. †: Equal advising.

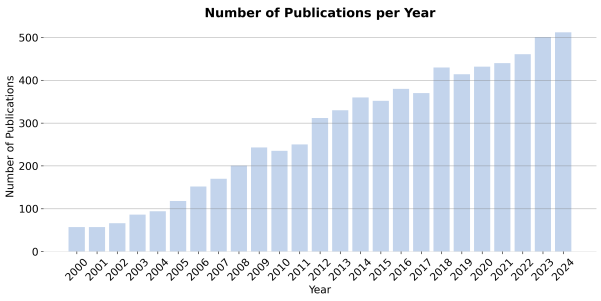


Fig. 1: Number of publications on image and video quality assessment per year (from Google Scholar).

camera motions. Moreover, extracting quality-aware video features and predicting video quality over both space and time must account for temporal human memory effects, making the problem even more challenging.

Conventional VQA models typically rely on computing perceptually relevant visual differences and/or natural statistical regularities. A simple mapping engine, such as a support vector regressor, is often used to learn how to map quality predictions from suitable distortion-aware features. Of course, the feature extraction process is integral to the success of these VQA models. While these approaches have demonstrated significant success and are used in global industry deployments, there remains ample scope for further improvement.

Fortunately, deep learning has been revitalized by leveraging large datasets of human-labeled images [1], [2], enabling significant progress across high- and low-level computer vision tasks. Neural network architectures trained on vast amounts of data have demonstrated the ability to capture semantic features and universal representations, reducing reliance on manual feature selection and showcasing notable generalization capabilities. Recent advancements in creating large-scale psychometric datasets for images and videos have further fueled the application of deep learning for quality prediction [3]–[6], as shown in Fig. 1. Modern deep VQA models typically employ pre-training on ancillary tasks with abundant data, followed by fine-tuning on more focused video quality databases. While these databases are increasingly comprehensive, their size remains insufficient to train large models due to the high costs of conducting extensive subjective studies. The adoption of deep networks has led to significant performance improvements in quality prediction by capturing diverse distortion phenomena and high-level semantic content while aligning with visual perception. However, the full potential of deep learning for video quality assessment remains unrealized, constrained by the scarcity of large, annotated psychometric datasets and limited understanding of human visual quality perception.

The recent advancement of large models [7]–[10], which are characterized by a significant number of parameters and trained on extensive data, has remarkably promoted the perceptual cognization of machines. Many researchers are exploring various manners of incorporating large language models (LLMs) and large multimodality models (LMMs) for IQA/VQA tasks, such as extending semantic-aware features in the prior knowledge embedded in large models [7], [8], and enhancing the explainability of quality assessment with quality-aware prompts [11], [12], among others.

Towards facilitating a better understanding current progress on modern video quality databases, and the development of deep learning-based models, we have sought to conduct a comprehensive overview of the past and recent advances and the state of the art of video quality modeling. Our key contributions include:

- We review both subjective (database) and objective (algorithm) quality assessment models, providing not only a detailed taxonomy but also a nuanced analysis of their evolution and core methodologies (see Fig. 2 and Fig. 7).
- We introduce readers to the elements of conducting subjective quality assessment studies, and review many of the most popular and representative video quality datasets, which vary by types of video content and targeted use cases.
- We review traditional image and video quality assessment methods, as a set-up for a comprehensive discussion of recent deep learning-based VQA models. Commonly used (perceptual) loss functions are also discussed.
- We compare representative IQA/VQA models on databases of emerging content, providing insights on model design in terms of spatiotemporal information modeling and prior knowledge use in large models.
- We discuss practical applications, some of which are deployed at very large commercial scales, remaining core challenges, and future opportunities for improving and applying deep learning-based VQA models, towards facilitating and inspiring future research efforts and industry deployments in the video streaming and social media fields.

The remainder of the paper is organized as follows. Section II introduces neurostatistical video distortion models and key network layers in deep learning-based VQA models. Section III summarizes subjective video quality assessment methods and widely used datasets. Section IV reviews full-reference and no-reference VQA models, emphasizing deep learning approaches and loss functions. Section V compares the performance of representative VQA models on emerging content databases. Finally, Section VI discusses practical applications and challenges, with Section VII concluding with forward-looking insights.

II. Background

To set the stage for the main topics, we will briefly review a few classical video quality assessment models, including structural similarity based methods and those that use neurostatistical distortion measurements. These established models, which are deep ingrained in principles of visual neuroscience, are quite relevant to processing in early network layers in deep perceptual video quality assessment models.

A. Structural Similarity

Natural photographic images are quite structured, and the human visual system (HVS) is able to efficiently extract structural information as part of the process of seeing. The well-known full-reference image quality assessment (FR-IQA) method called SSIM [13] measures the perceptual similarities of local path luminances, contrasts, and structures. SSIM [13]

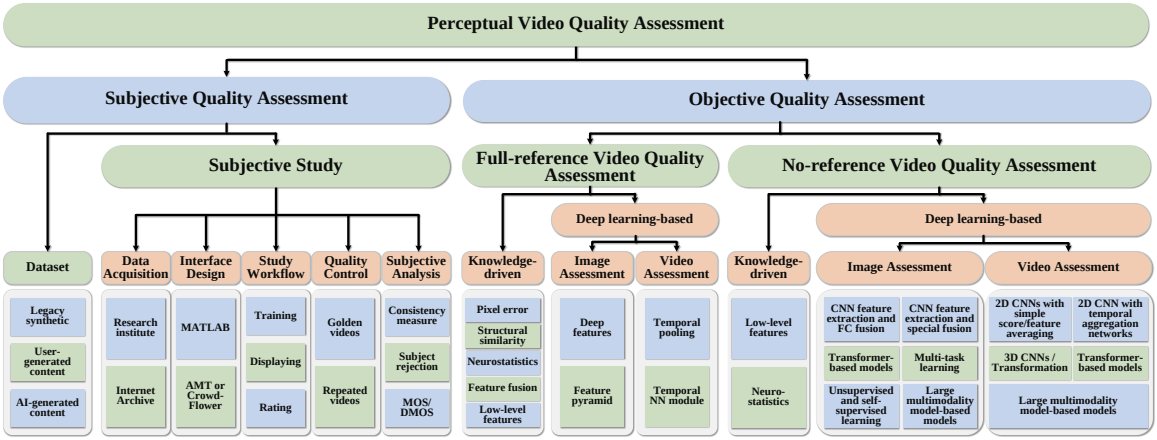


Fig. 2: Taxonomy of existing subjective and objective video quality assessment methods.

has been used in an extraordinary diversity of applications, including traditional FR-IQA [13]–[21], FR-VQA [22]–[24], and deep learning based FR-IQA [25], [26] and FR-VQA [27].

It is instructive to further consider the construction of the most widely-deployed version of SSIM, called Multi-Scale SSIM or MS-SSIM [13]. MS-SSIM contains three terms that are iterated over five scales, all combined in an exponentially-weighted product, with weights determined by a human study. The five contrast terms of MS-SSIM (just one in SSIM) may be interpreted as center-surround bandpass processes, computed over multiple scales, each rectified then divisively normalized by local bandpass energy. These terms may be viewed as a simple retinal model, with energy-normalized, non-directional center-surround receptive fields resembling the ganglionic responses. They are also similar to non-directional weighting functions learned in the earliest layers of CNNs, but with simple (*e.g.*, ReLu) activations replaced by rectification and local energy normalization, which is a perceptual strategy effectively employed in deep autoencoder based image compression [28], [29]. The structure terms of MS-SSIM, which involve saturation-biased, energy-normalized bandpass signal correlations resemble neural computations of correlations hypothesized to occur in extrastriate cortical area VZ [30], but which also resemble the feature correlations that occur in modern self-attention layers [31]. While the correlations in Transformer architectures are learned, weighted autocorrelations, the correlations in MS-SSIM are only pre-weighted by perceptual optimization of the MS-SSIM exponents. The correlations are computed between bandpass signals and bandpass distorted signals (over scales), hence include autocorrelations of both signal and distortion, as well as cross-terms.

It is also worth noting that SSIM and MS-SSIM are differentiable (and also quasi-convex [32]), and have found extensive use not only for assessing the outcomes of deep image models [33]–[35], but also as network loss functions, enabling perceptual network optimization of any kind of deep image regressor [36]–[38].

B. Neurostatistical Quality Features

High-quality photographic picture and videos reliably exhibit certain statistical regularities that are predictably altered by distortions [39], [40]. Specifically, pictures and videos that have bandpass filtered in space and/or time follow generalized Gaussian distributions [41], [42], and equivalent Gaussian-Scale mixture models [39], [43], [44]. However, when distortions are introduced, the bandpass statistic of pictures and videos tend to predictably deviate from these models. This has led to a plethora of powerful FR and NR picture and video quality prediction models that have been developed in wavelet [40], [45], [46], spatial luminance [47]–[50], chroma [51], [52], space-time [53], discrete cosine transform (DCT) [54], [55], and other bandpass domains [45], [46].

The performances of MS-SSIM, VIF [40], VMAF [43]), and other neuroscience-based models are so good that they are used to monitor and control the scaling and encode quality of most Internet traffic (which is about 80% picture and video data). Indeed, deep models have a hard time beating these models on cinematic-grade streaming data, and are much more compute-intensive. However, on lower-grade, multi-distortion user-generated content (UGC), deep learning-based picture quality prediction models are much better able to map the internal statistical structures of distorted pictures and videos to human percepts of visual quality, especially in the blind image and video quality assessment (BIQA/BVQA) scenarios.

C. Multi-layer Perceptron

We also briefly review the primary types of network layers that constitute modern deep networks. The simplest and earliest is the multilayer perceptron (MLP) [56], is a feedforward network consisting of multiple layers of interconnected nodes, each layer fully connected to the next. These are often referred to as "fully connected networks" (FCNs). When used in deep learning-based VQA models, MLPs usually serve as later regression layers, which serve to map deep features extracted by earlier layers to the final quality score predictions [57]–[66]. MLPs are also used to generate probability vectors that describe the distortions by applying softmax [63], [67].

D. Convolution Layer

Convolutional layers extract features from input pictures and video frames by applying spatial convolution operations,

which involve sliding a small kernel over the input data and computing the dot product between the kernel and each local input patch at each spatial position. Convolutional layers are effective for extracting spatial and/or temporal features over multiple scales from visual data, making them well-suited for video quality assessment tasks. A variety of video quality models apply convolutional neural networks (CNNs or ConvNets) to extract both low-level and high-level semantic features from video frames, to conduct video quality prediction [45], [57], [64]–[66], [68]–[77]. Other approaches employ 3D space-time CNNs to extract spatio-temporal features from blocks of video frames [59], [62], [67]. Some VQA models leverage both 2D and 3D CNNs to separately analyze spatial and temporal quality degradation [4], [58], [60], [61], [78], [79]. These sucecesses are supported by studies that have shown that perceptual distances are well modeled by deep activations within convolutional networks [27], [80]–[82].

E. Recurrent Layer

Videos are spatio-temporal, as are many distortions, hence capturing short- and long-term temporal correspondance is essential for accurate perceptual video quality prediction. Recurrent layers, like long short-term memory (LSTM) blocks [83] and gated recurrent units (GRU) [84] have been widely used to model temporal video dependencies. LSTMs and GRUs may either process multi-frame clips [62] or single frames [65], [72], [73], [76], [85], when predicting video quality scores.

F. Attention Layers

Attention layers [31] have become important components of modern deep learning models. They are able to selectively weight local space-time regions or channels that may be more important for feature representation learning. Typical attention models include convolutional spatial attention mechanism [86] and channel squeeze-and-excitation attention models [87]. Various methods have been proposed for calculating attention weights. For example, some models use Gaussian functions [88] to compute the attention weights, while others use average and maximum pooling [65], standard variance [89], or graph convolutions [65] to obtain the weights. The choice of attention mechanism depends on the nature of the data and the specific requirements of the VQA task at hand.

G. Transformer

Originally proposed for natural language processing tasks [31], Transformers have since been successfully applied to various computer vision tasks [31], [90], [91]. The core components of Transformers are multi-head self-attention mechanisms which can effectively capture global spatial relationships by learning content-dependent adaptive weights. Since large-scale spatial and temporal associations can play an important role in making video quality predictions, Transformer models have also been studied in this context. However, the quadratic computational complexity of self-attention mechanisms presents challenges when processing large amounts of video data. To address this issue, several studies have explored the use of sparsely sampled clip- [92] and/or patch-level [93] features to reduce the number of Transformer tokens. Other have devised less expensive window-based attention mechanisms [91] having only linear complexity with respect to image size [94], [95]. Overall, the Transformer architecture is a promising direction for modeling long-range (memory) dependencies that affect perceived video quality, although further research is needed to explore this potential.

H. Large Models

Large models have shown impressive capabilities on visual understanding tasks, thus it is intuitive to leverage them as perceptual quality assessment tools. Several recent efforts have demonstrated their effectiveness on the quality assessment task. Among them, the large segmentation model SAM [7] extracts highly generic semantic structural information while exhibiting exceptional generalization ability, and has been adopted as a semantic feature extraction module by several IQA models [96], [97]. By training on extensive image-text pair data, the multi-modal model CLIP [8] can build semantic-level relationships between texts and visual information. CLIP can be used as either a semantic-aware feature extraction module [98], [99] or a quality index customized by perception-aware prompts [100]–[102]. Large language models can be extended to multi-modal tasks by incorporating visual inputs via multi-modal models like CLIP, resulting in large multimodal models. A series of recently developed IQA/VQA models successfully leverage LMMs to perform prompt-driven [11], [103] or embedding-based [104] quality evaluation.

III. Subjective Video Quality Assessment

A. Subjective Study Introduction

Subjective VQA by a sufficiently large enough sample of human subjects is the most reliable method to assess perceptual video quality. Subjective VQA studies provide valuable data by obtaining human scores on corpuses of distorted videos. These may be very specific, *e.g.*, directed towards distortions arising from frame rate variations, or they may be extremely diverse, as with UGC content. In any case, they provide ”golden benchmarks” against which to design, assess, and compare the performances of objective VQA models.

Subjective VQA includes the In-lab study and the studies may be broadly divided into two types: those that are conducted under controlled conditions in a laboratory, and those conducted online via crowd-sourcing. The former allows for the assessment of quality under constrained conditions, usually by reliable human raters, while the latter is an efficient and successful way of gathering much larger numbers of human annotations of video quality, but with much less control over viewing devices and environment, and participant network conditions and overall reliability.

*1) **Data Acquisition**:* The creators of early VQA databases obtained open-source videos from public websites, willing academic or industry sources, or by capturing their own content. Because of copyright of issues, obtaining videos that could be openly shared with others was difficult, and early VQA datasets such as LIVE VQA [105] only included about

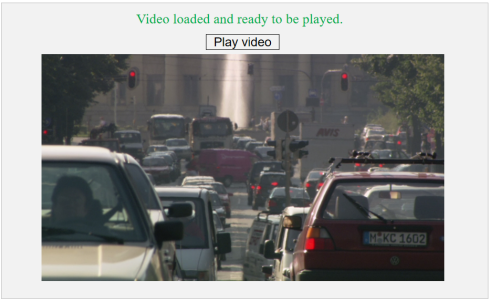


Fig. 3: Example of a visual interface used when playing video.

Fig. 4: Examplar discrete rating scale.

a dozen unique contents. Since the main application in those days was the emergent digital television, these contents were uncompressed, cinematic quality videos captured with high-end professional cameras, then converted to digital format with the utmost care, to guarantee that the reference videos would be as distortion free as possible. Once a suitable set of high-quality videos has been collected, then a set of representative distortions are applied, taking care to create a wide range of perceived seventies. Ideally, the range of distortion exceeds that encountered in the target application, since, without the benefit of very large numbers of human labels (typically only in the tens of thousands in this kind of study), model-building is enhanced by a wider range of less clustered samples. The applied distortions include standard-compliant compression [105], wireless transmission artifacts [106], frame-rate variations [107], flicker [108], scaling distortions [109], bit depth (dynamic range variations) [110], and more.

The other major category of VQA datasets are those containing real, unchanged video content that was generated by consumer-level users and then typically shared on social media platforms. To gather data that is representative of this UGC data, VQA researchers typically acquire much larger numbers of open-source videos from public platforms like the Internet Archive, and YouTube, [4], [111], [112]. If a very large number of videos are collected, these may be statistically sampled based on low-level attributes (such as brightness, colorfulness, temporal activity, etc.) to better match typical social media videos [4], [112]. In the end, the goal is to acquire a large set of videos, typically numbering in the tens of thousands, that have not been altered in any way (including resizing) that were taken by typical real-world casual users. These UGC videos contain very diverse mixtures of distortions.

*2) **Study Interface Design**:* Crowd-sourcing studies using platforms like Amazon Mechanical Turk (https://www.mturk.com/) and CrowdFlower (https://www.crowdflower.com/) are an effective way to gather much larger numbers of human annotations, thereby allowing for the development of data-driven approaches to more general VQA problem scenarios [111], [113]. As shown

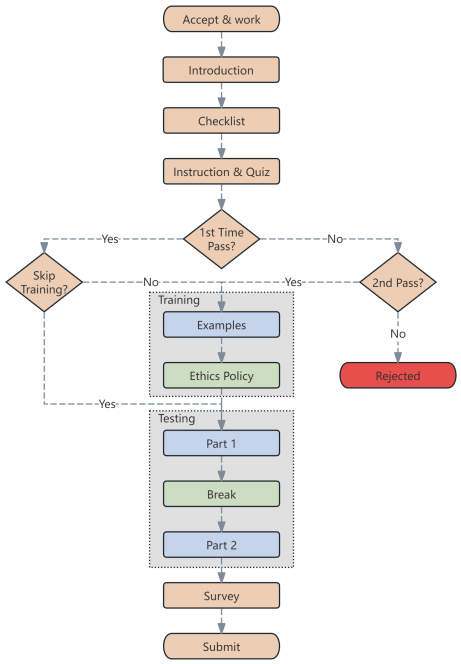


Fig. 5: Flowchart of a online crowd-sourcing study [113].

in Fig. 3 and Fig. 4, each video is played, without display scaling, on a white or gray background. Once viewed, a rating interface appears that prompts subjects to record their quality scores using a sliding bar. While this is happening, next video to be viewed is cached so that it is entirely downloaded before it is played back, to avoid any re-buffering events or stalling that might change a subject's quality judgments. Plugins to speed up the videos are also detected and prohibited.

Several survey questions are asked at the beginning of each session, to collect subject demographics and to record any video and audio device specifications that have not already been measured. Participants whose AMT acceptance rates (known as reliability scores) or equipment specifications do not meet requirements are not allowed to proceed. Before their participation starts, each subject is asked to complete a series of steps to eliminate distractions, ensure audio devices are on (if relevant), adjust seating, and wear corrective lenses.

For studies done in the laboratory, MATLAB has been a popular platform for developing user interfaces for human subjective quality studies. To further ensure uninterrupted playback, all the displayed sequences are pre-loaded into memory before they begin to play, to avoid any latencies due to slow hard disk access of large video files. The videos are viewed by the subjects on a monitor both contemporary to and appropriate for the task at hand, i.e. able to faithfully present the videos at modern resolutions and bit depths reflecting current commercial products, and configurable, if needed, for specific scenarios, such as changing frame rates. In most instances, the display device should present the videos at their native resolutions, and at higher refresh rates than the native frame rates. For subjects taking part in these studies, they are generally naive volunteers, and usually are asked to participate in a short series of vision tests, including a Snellen test of visual acuity, an Ishihara test of color vision, and if

appropriate, a Randot test of 3D vision.

*3) **Study Workflow**:* The general workflows of in-lab and crowd-sourced studies are similar, as depicted in Fig. 5. At the beginning, each subject is given a brief set of instructions regarding the task they will participate in. They are then ordinarily shown a few sample videos exhibiting a range of distortions and qualities that they might encounter in the study.

Methods of video display are generally classified as single, double, or multiple stimulus depending on whether one, two, or several videos are displayed simultaneously or in sequence for each subject to view and quality-rate. The single stimulus method is well suited to a large number of emerging multimedia applications, such as quality monitoring for Video on Demand, IPTV, Internet streaming, etc. One simple reason is that presenting a single stimulus better reflects real viewing. Moreover, it is difficult to compare two videos displayed in space (*e.g.* side by side) or in time (one followed by the other). It also significantly reduces the amount of time needed to conduct the study (given a fixed number of human subjects) as compared to a double stimulus study [105]. Nevertheless, there are situations where a double stimulus study may be called for, *e.g.* if the video distortions are very subtle.

After a short training session to allow the participants to become familiar with the interface, each subject is asked to rate a set of videos. AVQA studies normally use discrete scales such as the Absolute Category Rating (ACR), Degradation Category Rating (DCR), and Comparison Category Rating (CCR) [114] scales to record quality judgments, while others use a continuous rating scale to allow the subjects to record with greater freedom and increased sensitivity. To avoid fatigue, the experiment is usually divided into several sessions of short durations (typically 30-45 minutes) with breaks. Subjects who give very similar quality scores to all videos are disqualified. A wide variety of protocols are used to detect and eliminate the scores of subjects who are distracted, inattentive, or afterwise poorly participating.

*4) **Crowd-sourcing Quality Control**:* When conducting online crowd-sourcing studies, quality control is essential to obtaining reliable labels. While there is a great advantage to access to many more subjects (often orders of magnitude more), the data is much noisier unless precautions are taken to detect and remove poorly connected or dishonest "workers".

One common practice is to include "golden" videos selected from other databases featuring similar content, on which highly reliable subjective scores were previously obtained. These scores are compared with the worker's inputs to determine if there are wide enough divergences. In this way, many unreliable subjects may be prevented from further participation. For more details on these and other methods of improving online crowdsourced data reliability in this context, see [4], [112], [113]. Since some subjects are less serious, distracted, or otherwise poorly focused on their tasks, it is important to apply subjection rejection protocols such as ITU recommendations BT.500 [113], BT.500.11 [128], BT.500.15, or the more accurate and comprehensive SUREAL [129].

Another practice is to select a small random set of videos which are repeated elsewhere in a subject's session as a control. Again, failure of a subject to produce consistent scores on these may lead to their elimination. Some users may have inadequate hardware to receive and display the videos, or poor/low bandwidth connectivity. While the latter may be ameliorated by requiring each video to be completely downloaded before display, the session may be terminated with a polite explanation, since the fault is not the workers. Since Amazon Mechanical Turk includes tools to measure technical aspects of a user's device or connectivity, these are important and practical precautions.

*5) **Subjective Analysis**:* To establish the internal integrity of the final set of collected subjective scores, it is important to examine the consistency of the recorded Mean Opinion Scores (MOS). This is commonly done by randomly dividing the subjects pool into two equal and disjoint sets, then computing the Spearman Rank Correlation Coefficient (SRCC) between the two corresponding sets of MOS.

Repeating this random calculation a reasonable number of times, then taking the average (and/or median) SRCC as a consistency measure, is useful for assessing the difficulty of the task and/or the degree of agreement of the subjects. Values above SRCC=0.85 are generally desirable, while SRCC=0.95 is only occasionally reached (on easier VQA tasks). Additional subject rejection methods, especially useful for large online studies where the subjects can be less reliable, include differences between each subject's MOS and those obtained from a known, reliable set of subjects (perhaps recorded in the laboratory), on a set of "golden" videos [4], [111]. Another method is to present a small number of videos (typically 5) twice within a subject's session, but spaced in time, then determine whether they give sufficiently similar responses on repeated videos as on the first presentations of those contents.

In addition to MOS, difference mean opinion scores (DMOS) [130], [131] are also commonly recorded in single stimulus experiments using hidden reference removal [132]. In each session, a difference score on each video rated by each subject is computed by subtracting the quality score from the corresponding reference quality score. The difference scores of reference videos are zero and are removed. Z-scores [133] are then computed to normalize the difference scores from each session, then are collected over all sessions to form a matrix indexed by subjects and videos. Finally, DMOS is obtained by linearly rescaling Z-scores to the range of $[0, 100]$.

B. Profession Content Datasets

Most early VQA databases were directed towards television and streaming, and contain original, professionally generated studio content which is subsequently distorted under laboratory conditions. These datasets generally comprise 10-20 unique high-quality source videos, processed in a specific manner to simulate one of a few synthetic impairments (*e.g.* MPEG encoding, or packet loss), wherein MOS (or DMOS) are obtained from a relatively small (20-40) group of subjects in a controlled laboratory environment. These datasets are limited in representation of the complex, highly variable characteristics of videos.

Table I summarizes and classifies several of the most widely-used and successful public VQA databases created

TABLE I: Taxonomy of subjective VQA databases: legacy databases with synthetic distortions against UGC databases with large-scale in-the-wild videos. Legacy databases and UGC databases are indicated by *italic* and orthographic font, respectively.

Year	Name	Unique contents	Total videos	Resolution	Frame Rate	Video Length(s)	Format	Distortion Type	Subjects	Rates per video	Data	Env
2008	*LIVE-VQA* [105]	10	150	786×432	25/50	10	YUV+264	Compression, Transmission	38	29	DMOS+σ	In-lab
2014	*CVD2014* [115]	5	234	720p, 480p	9-30	10-25	AVI	Camera, Capture	210	30	MOS	In-lab
2015	*MCL-V* [114]	12	108	1080p	24-30	6	YUV+264	Compression, Scaling	45	32	MOS	In-lab
2015	*BVI-HFR* [116]	22	88	1080p	15-120	10	YUV420	Temporal downsampling	29	29	MOS	In-lab
2021	*LIVE-YT-HFR* [117]	16	480	2160p, 1080p	24-120	6-8	YUV420	Compression, Temporal downsampling	85	223	MOS, DMOS	In-lab
2022	*BVI-VFI* [118]	36	540	540p	30-120	5	MP4	Frame interpolation	189	57	DMOS	In-lab
2017	KoNViD-1k [119]	1200	1200	540p	24-30	8	MP4	In-the-wild	642	114	MOS+σ	Crowd
2018	LIVE-VQC [111]	585	585	1080p-240p	19-30	10	MP4	In-the-wild	4776	240	MOS	Crowd
2019	YouTube-UGC [5]	1380	1380	4k(HDR)-360p	15-60	20	MKV	In-the-wild	>8k	123	MOS+σ	Crowd
2019	FlickrVid-150k [72]	153841	153841	540p	24-120	5	MPEG-4	In-the-wild	-	5 (89)	MOS	Crowd
2021	LSVQ [4]	38811	38811	Diverse	Diverse	5-12	-	In-the-wild	6300	35	MOS	Crowd
2021	Youku-V1K [120]	1072	1072	1080p	-	10	-	UGC, PGC	-	15+	MOS+σ	Crowd
2021	PUGCQ [121]	10k	10k	$\leq$1080p	-	5	MP4	Professional UGC	50	50	MOS	Crowd
2021	YT-UGC+ [122]	1380	1380	4k(HDR)-360p	15-60	20	MKV	UGC, Compression	>8k	10(labels)	DMOS, MOS+σ	Crowd
2021	LIVE-YT-Gaming [123]	600	600	1080p-360p	30/60	8-9	MP4	UGC gaming	61	3	MOS	Crowd
2022	Tele-VQA [113]	2320	2320	-	-	7	-	In-the-wild	526	34	MOS	Crowd
2023	Maxwell [124]	4543	4543	240p-1080p	-	9	MP4	In-the-wild	35	440	MOS, Ternary choices	In-lab
2023	DIVIDE-3k [125]	3590	3590	240p-1080p	24-30	2-13	MP4	In-the-wild	35	125	MOS+σ	In-lab
2023	TaoLive [126]	418	3762	720p, 1080p	-	8	MP4	UGC, Compression	44	44	MOS	In-lab
2024	KVQ [127]	600	4200	-	-	-	MP4	UGC, Enhancement, Pre-processig, Transcoding	15	15	MOS, Ranking score	In-lab

during the past 15 years. Fig. 6 (a)-(f) shows sample frames taken from six popular legacy databases. The first useful VQA database was the LIVE Video Quality Database [105] published in 2010, which includes 10 reference videos and 150 videos simulating compression and transmission distortions. Other databases containing original and artificially-distorted videos include such as CVD2014 [115], MCL-V [114]. As video dimensions have increased in other ways than spatial resolution, other video quality databases have become available that contain subjectively labeled high motion [134], high frame rates (HFR) [107], [116], high dynamic range (HDR) [135], and frame interpolation [118] videos.

1) ***LIVE-VQA****:* The LIVE Video Quality Database [105] consists of 10 uncompressed reference videos and 150 distorted videos created using four common streaming degradations, such as compression and packet loss. The videos in LIVE-VQA were all in 720P formats and span a wide range of visual qualities. Each video was rated by 38 human subjects in a single stimulus presentation. The subjects scored the video quality on a continuous quality scale.

2) ***CVD2014****:* This database [115] contains 234 videos from 78 cameras without post-processing distortions. Instead of only MOS, it includes quality descriptions from participants, covering dimensions like sharpness, graininess, color balance, darkness, and jerkiness.

3) ***MCL-V****:* The MCL-V dataset [114] includes 12 uncompressed HD video clips with diverse content (*e.g.*, cartoons, faces, action), in YUV420p format, 1080p resolution, and frame rates of 24fps or 30fps. These 6-second clips were distorted using H.264/AVC compression and scaling, resulting in 96 videos. Ratings from 45 subjects using a pairwise comparison method were converted into absolute scores using the Bradley-Terry model.

4) ***BVI-HFR****:* This database [116] examines the impact of high frame rates on video quality. It includes 22 4K, 120fps source sequences, downsampled to 1080p and frame rates of 60fps, 30fps, and 15fps, resulting in 88 videos. Quality ratings were collected from 29 participants in a controlled in-lab study.

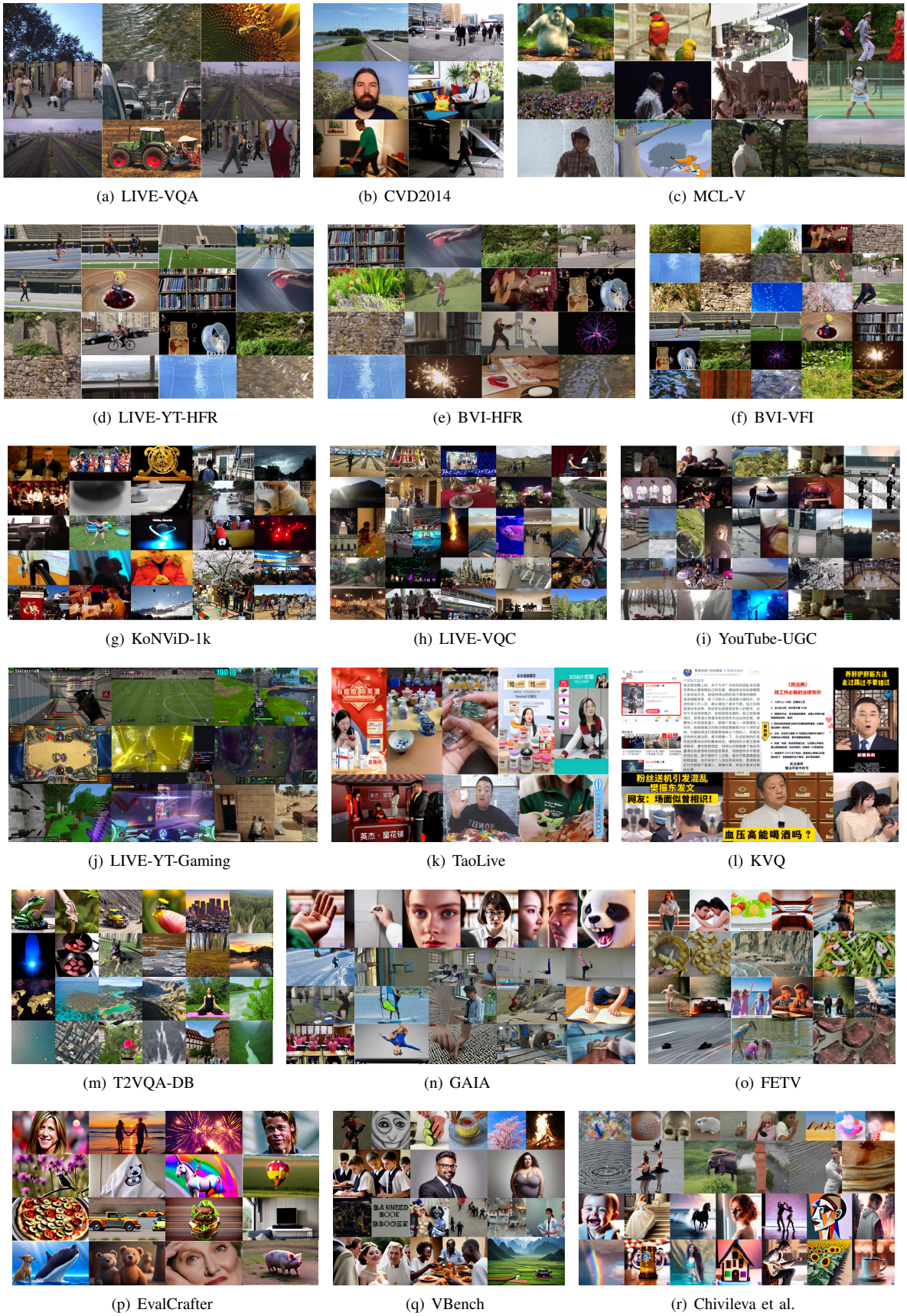
5) ***LIVE-YT-HFR****:* LIVE-YT-HFR [117] studies the combined effects of compression artifacts and frame rate variation, and contains 480 videos from 16 unique contents distorted by combinations of 6 frame rates (24, 30, 60, 82, 98, 120fps) and 5 compression levels (CRF 0 - 63). An in-lab human study was conducted to yield 19,000 human quality ratings obtained from a pool of 85 human subjects.

6) ***BVI-VFI****:* BVI-VFI [118] is proposed to understand how the quality of temporally interpolated content is perceived. It employs five commonly used video frame interpolation algorithms on 36 source videos to generate 540 interpolated videos. More than 10,800 ratings were collected by a large-scale human study involving 189 human participants.

Since increased motion, as occurs in live sports streaming [134], increased frame rates [107], and increased bit depths [135], all operating in conjunction with compression, present special challenges of different distortions, higher bandwidth requirements, and a previous lack of models able to predict these kinds of distortions, new databases are required to allow for the development of VQA targets able to target these expanded classes of distortion scenarios [136]–[138].

C. UGC Datasets

Most of the databases just described are characterized by a small number of unique contents (10-15) rated by a small number of human subjects (< 100), implying limited (but often sufficient) amounts of data. However, in the context of (generally) non-professional, a vast variety of distortions may occur during video acquisition or rendering, resizing, compression, sharing, re-compression, processing, transmission, and reception. Several of this multitude of possible space-time distortions may afflict any video, commingling and interacting to create new, unnamable distortions. To model this plethora of possible distortions, much larger datasets of real-world videos are required to span the space of possible



(a) LIVE-VQA (b) CVD2014 (c) MCL-V

(d) LIVE-YT-HFR (e) BVI-HFR (f) BVI-VFI

(g) KoNViD-1k (h) LIVE-VQC (i) YouTube-UGC

(j) LIVE-YT-Gaming (k) TaoLive (l) KVQ

(m) T2VQA-DB (n) GAIA (o) FETV

(p) EvalCrafter (q) VBench (r) Chivileva et al.

Fig. 6: Samples of video contents selected from legacy datasets ((a)-(f)), UGC datasets ((g)-(l)), and AIGC datasets ((m)-(r)).

required representations. Likewise, much larger amounts of human quality annotations are required to offer the possibility of mapping video measurements to accurate subjective quality predictions. This has led to the creation of large-scale crowdsourced video quality studies, inspired by the success of the LIVE Challenge picture quality database [139]. Table I provides a summary of some of these, and descriptions of several popular and recent UGC VQA databases are introduced as follows. Fig. 6 (g)-(l) shows sample frames taken from six popular UGC databases.

1) ***KoNViD-1k****:* KoNViD-1k [119] contains 1,200 public-domain videos showcasing diverse "in the wild" distortions sampled from YFCC100m [140]. A crowdsourced study gathered over 73,000 ratings using a five-category quality scale.

2) ***FlickrVid-150k****:* FlickrVid-150k [72] includes 153,841 videos with coarse quality ratings and another 1,596 videos with finer labels. Videos were re-encoded for uniformity but lost some diversity and authenticity.

3) ***LSVQ****:* LSVQ [4] is the largest VQA dataset with 38,811 diverse videos and 116,433 cropped spatial, temporal, and spatiotemporal patches. It includes about 5.5 million quality ratings from 6,284 subjects using a single-stimulus, continuous rating scale protocol.

4) ***LIVE-VQC****:* LIVE-VQC [111] consists of 585 authentic videos captured by 80 inexpert videographers using 43 different models of 101 personal camera devices, thereby obtaining wide ranges of complex, authentic distortions. Over 205,000 opinion scores were collected from 4,776 participants.

5) ***YouTube-UGC****:* YouTube UGC [5] features 1,500 videos across different resolutions and categories, including features like High Dynamic Range (HDR). Each video was rated by over 100 participants.

6) ***Youku-V1K****:* Youku-V1K [120] includes 1,072 videos from youku.com, annotated with over 22,000 ratings using a discrete scale to study content with intentional editing effects like blur of unimportant or background objects.

7) ***PUGCQ****:* PUGCQ [121] contains 10,000 videos of professionally-generated content labeled with attributes like 'face', 'noise', and 'blur'. The dataset aims to maximize content diversity using features from pre-trained models.

8) ***YT-UGC+****:* YT-UGC+ [122] refines YouTube-UGC categories with 610 detailed labels and includes three transcoded variants: Video-On-Demand (VOD), Video-On-Demand with Lower Bit-rate (VODLB), and Constant Bit Rate (CBR). Videos are categorized into three DMOS levels.

9) ***LIVE-YT-Gaming****:* Gaming videos differ from other UGC content due to distinct visuals and specific distortions like improper graphics settings, frame drops, temporal lags, and low-quality recording software. LIVE-YT-Gaming [123] is the first database to focus on real UGC gaming content, with 600 clips from 59 games across diverse resolutions and frame rates. It includes 18,600 ratings from 61 participants.

10) ***Tele-VQA****:* Tele-VQA [113] comprises 2,320 videos, including 1,129 virtual meeting recordings and 1191 YouTube videos, with 78,880 subjective ratings from 526 participants, evaluating both video and audio quality.

11) ***Maxwell****:* Maxwell [124] contains 4,543 videos with over two million opinions annotated by 35 human participants. It collects multidimensional quality ratings covering technical and aesthetic factors to guide model design. Different from previous datasets, these explanation-level human opinions in Maxwell highlight correlations between various quality factors of the human vision system.

12) ***DIVIDE-3k****:* DIVIDE-3k [125] features 3,590 videos sourced from YFCC-100M [148], Kinetics-400 [149], and LSVQ [4], annotated with 450,000 opinions, including scores for aesthetic, technical, and overall quality.

13) ***TaoLive****:* TaoLive [126] explores both in-capture distortions and live-streaming distortions using 418 UGC videos from Taobao [150] live streaming platform, compressed into 3,762 variants with 44 participants rating the videos.

14) ***KVQ****:* KVQ [127] studies short-form videos on Kwai [151], featuring 4,200 videos processed via enhancement, pre-processing, and transcoding. It uses a mixed scoring manner of MOS and ranking labels from 15 participants.

D. AIGC Datasets

The emergence of large text-to-image (T2I) and text-to-video (T2V) models has introduced challenges in assessing the perceptual quality of AI-generated content (AIGC), including text alignment, naturalness, rendering, and temporal consistency. Addressing these issues necessitates subjective studies to benchmark and calibrate objective AIGC VQA models. Recent years have seen the emergence of multiple datasets that model the subjective quality of AI-generated images, focusing on human preferences [125], [152], [153], perceptual quality [154]–[156], text alignment [157], [158], and compression-aware quality [159]. However, there are still few AIGC VQA datasets available, and more work is needed in this direction. Existing AIGC VQA databases are listed in Table II, and sample frames are shown in Fig. 6 (m)-(r).

1) ***Chivileva et al.****:* Chivileva et al. [141] developed a dataset of 1,005 videos generated by five T2V models from 201 carefully selected prompts covering diverse scenarios. 48,240 ratings were collected from 24 subjects on perception and alignment quality.

2) ***EvalCrafter****:* EvalCrafter [142] contains 700 prompts grouped into four metaclasses, used to generate 2,500 videos from five T2V models. It features 16 objective quality assessment tools for four quality aspects and 8,647 subjective ratings from seven subjects on five quality dimensions.

3) ***FETV****:* FETV [143] includes 619 prompts categorized by content, attribute control, and complexity. It evaluates four T2V models across four quality aspects, including spatiotemporal quality and alignment, using assessments from three human subjects.

4) ***VBench****:* VBench [144] benchmarks T2V quality with 16 evaluation dimensions linked to 100 prompts each. It includes a subjective study to validate VBench's alignment with human opinions, whereby human preference annotations were collected on videos generated by four T2V models.

5) ***T2VQA-DB****:* T2VQA-DB [145] features 10,000 videos from nine T2V models based on 1,000 graph-selected prompts. Opinions on text-video alignment and fidelity were collected from 27 subjects.

TABLE II: Taxonomy of subjective AIGC VQA databases.

Year	Name	Total videos	Resolution	Frame Rate	Video Length(s)	Models	Prompts	Subjects	Ratings	Env	Quality Aspects
2023	Chivileva et al. [141]	1,005	-	-	-	5	201	24	48,240	Crowd	Perception, alignment
2023	EvalCrafter [142]	2,500	256×256-1280×720	8, 24	2-4	5	700	7	8,647	-	Video quality, Text alignment, Motion quality, Temporal Consistency, Subjective likeness
2023	FETV [143]	2,476	256×256,480×480 512×512, 576×320	8	2-4	4	619	3	28,116	-	Static quality, Temporal quality, Overall alignment, Fine-grained alignment
2023	VBench [144]	6,984	256×256,512×512	8, 10	2-3	4	1,746	-	-	-	Video quality, Video-Condition Consistency
2024	T2VQA-DB [145]	10,000	512×512	4	4	9	1,000	27	270,000	-	Video fidelity, Text alignment
2024	GAIA [146]	9,180	256×256-2048×1536	4-50	2.8	18	510	54	971,244	In-lab	Video action quality
2024	LGVQ [147]	2,808	256×256-1408×768	4-24	8-96	6	468	54	454,896	-	Spatial quality, Temporal quality, Text-to-video alignment

*6) **GAIA**:* GAIA [146] focuses on human action quality in AI-generated videos. It contains 9,180 videos from 18 T2V models generated using 510 GPT-4 [160] prompts, and is rated on three quality aspects by 54 subjects.

*7) **LGVQ**:* LGVQ [147] explores motion of various scenarios by decomposing prompts into foreground, background, and motion categories. 2,808 videos were generated from six T2V models based on 468 prompts, and then rated by 54 subjects for spatial quality, temporal quality, and text alignment.

IV. Objective Video Quality Assessment

In this section, we provide an in-depth review of video quality assessment methods covering both full-reference and no-reference models, with a focus on deep learning-based approaches and commonly used loss functions. We analyze, categorize, and summarize the current mainstream quality assessment methods, highlighting model performance and efficiency in quality prediction. Fig. 7 illustrates an overarching summary of different algorithm categories and their development over time, offering a structured overview of each type of model and their evolution.

A. Full-reference Video Quality Assessment

Full-reference video quality assessment (FR-VQA) models seek to quantify (perceptual) differences between reference videos and their distorted counterparts, and map them to predictions of visual quality. Since FR models have available undistorted content information, they generally demonstrate higher correlations against human judgments than non-reference models. FR models can be classified into knowledge-driven and deep learning-based algorithms: 1) **Knowledge-driven** algorithms are inspired by models of the human visual system, and calculate perceptual distance between pixels, structures, or other relevant properties such as gradient difference. 2) **Deep learning-based** methods accept reference and distorted pairs of videos or video frames as inputs, learning both perceptual quality/distortion representations as well as deep semantic abstractions, and how distorted and semantic data relate when forming quality predictions. In this section, we discuss both kinds of methods by providing a concise overview of IQA models before delving into a comprehensive examination of VQA models.

*1) **Knowledge-driven FR VQA Methods**:* The last two decades have seen the development of a remarkable variety of knowledge-driven FR-IQA models. These models can be extended to FR-VQA tasks by simply using them to estimate frame-wise quality, then temporally pooling the frame-level measurements [161], [162] to obtain overall video quality scores. However, this approach does not capture temporal distortions or their impacts on visual perception. Significant effort has been applied to develop true FR-VQA models by knowledge-based temporal distortion modeling. In the following, existing FR IQA/VQA models are further classified into five sub-categories, and discussed in turn.

Type i: Pixel-error-based VQA. The earliest full-reference image quality assessment (FR-IQA) algorithms measured pixel-level errors, such as the mean squared error (MSE) or peak signal-to-noise ratio (PSNR). However, due to the lack of consideration of how people view and process visual distortions, these simple numerical measures often exhibited poor correlation with human perceptions [163].

Type ii: Structural Similarity-based VQA. Things began to change with the introduction of a perceptually motivated IQA model that is accurate and efficient, the Primetime Emmy Award winning Structural Similarity or **SSIM** [164] Index. Further advancements of SSIM soon followed, including multi-scale structural similarity (**MS-SSIM**) [13], wavelet-domain spectral similarity (**CW-SSIM**) [14], information content weighted (**IW-SSIM**) [15], gradient and phase congruency similarity (**FSIM**) [19], edge strength similarity (**ESSIM**) [16], gradient magnitude similarity [17]–[19], and visual saliency (**VSI**) [20], [21]. Owing to the simplicity and effectiveness of SSIM [164], significant efforts have been dedicated to extending it to the video domain. An early **"video SSIM"** [22] operates by space-time sampling estimates of perceptual video quality at three levels: spatially local, frame, and sequence, employing two weighting pooling methods based on luminance estimation and frame motion. **Wang and Li** [165] proposed an alternative weighting scheme of SSIM based on human perception of motion information. However, weighted pooling of spatial SSIM scores does not necessarily account for temporal distortions. Succeeding works explored modeling temporal information in quality assessment. **V-SSIM** [23] is a motion compensated variant of SSIM [164] that models motion information using optical flow. **MC-SSIM** [166] is another motion-compensated variant of SSIM, which evaluates structural retention between motion-compensated regions in a frame. **Manasa and Channappayya** [24] compute temporal quality estimates using local optical flow statistics and spatial quality estimates using MS-SSIM [13], subsequently pooling both estimates into single video quality scores. Instead of separately analyzing spatial and temporal distortion, **3D-**

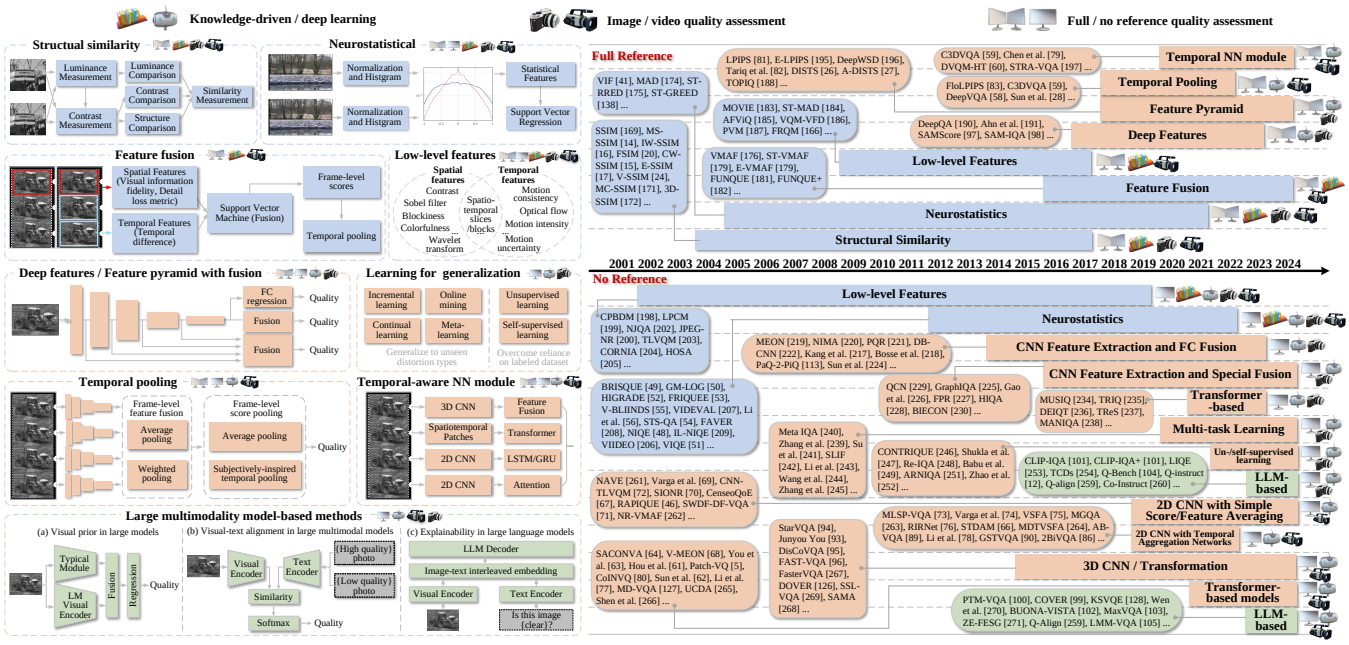


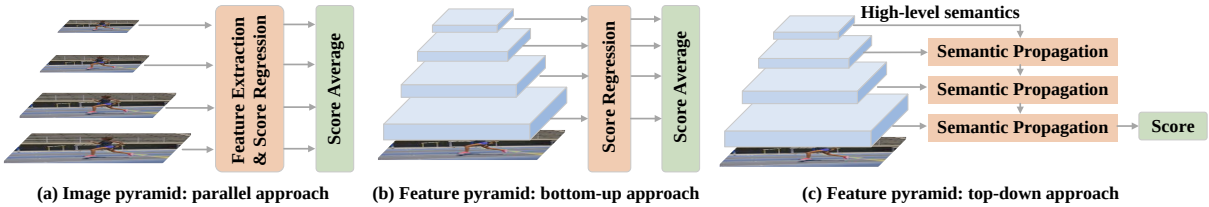
Fig. 7: Classification and evolution of objective quality assessment models. The left figure presents framework diagrams for each category, while the right figure highlights their temporal progression and representative models. Categories are color-coded as follows: knowledge-driven (blue), deep learning-based (orange), and large model-based (green).

SSIM [167] generates a 3D quality map by applying SSIM measurements within local 3D blocks, which are then pooled into an overall video quality score using a weighted scheme based on local information content.

Type iii: Neurostatistics-based VQA. Another major advance was the discovery of neurostatistical models of the responses of visual neurons to visual distortions. This approach was best exemplified by the full reference Visual Information Fidelity (**VIF**) [40], [168] model, which quantifies statistical deviations of visual information arising from distortions, as measured under a modified natural scene statistics model. As discussed later, a variety of popular and powerful no-reference (NR) visual quality prediction models have been devised using similar neurostatistical perceptual distortion models. Another strategy, called most apparent distortion (**MAD**) [169] explicitly separates distortion strength into either apparent or invisible categories, subsequently modeling two distinct measurement strategies on high and low-quality images, respectively. The **ST-RRED** model [170] expands on the single-picture VIF [40] IQA model by modeling temporal distortions using neurostatistical models of spatially-bandpass frame differences. ST-RRED remains highly competitive on most VQA databases and has the added efficiency of allowing for reduced-reference versions via a wavelet-domain subsampling strategy. **ST-GREED** [137] goes further by addressing distortions arising from variable frame rates (VFR) combined with compression, using neurostatistical spatial and temporal band-pass video models. ST-GREED performs remarkably better than any prior existing models on a large subjective database of compressed VFR videos ranging from 30 to 120 frames/sec [107]. However, the same authors show that previous VQA models like SSIM, MS-SSIM, ST-RRED, VMAF [171], and even PSNR can be greatly improved in a simple way by using concepts from ST-GREED [172].

Type iv: Feature fusion-based VQA. VIF and ST-RRED form the foundation of the Netflix Emmy-winning Video Multi-Method Assessment Fusion (**VMAF**) [173] models, which simplifies them by using only four spatial channels of VIF, average absolute frame differences, and a detail-loss measurement feature [174]. These features are used to train VMAF on an internal VQA database using a support vector regression (SVR) model. VMAF has been used to control the quality of all video encodes streamed globally for years, and has been adopted by many other streaming video providers, making it competitive with SSIM/MS-SSIM, which controls the quality of a large percentage of all television content, and much social media, for example picture/video content uploaded and subsequently streamed by Facebook. **ST-VMAF** and **E-VMAF** [175] enhance the performance of VMAF [173] by enriching its temporal features and by employing multiple regression models, respectively. To alleviate the computational burden presented by multiple models in fusion-based VQA methods, **FUNQUE** [176] shares computation by decomposing the features of different models into a common transform domain. The succeeding framework **FUNQUE+** [177] includes low-complexity fused-feature models to boost quality prediction accuracy.

Type v: Low-level motion feature-based VQA. FR VQA models of this type often model temporal distortions within well-designed motion feature extraction modules, based on motion trajectories, wavelet transform, etc. The concept of V-SSIM is expanded upon by the **MOVIE** index [179], which evaluates space-time video quality along motion trajectories, using a deep brain model of extracortical area MT. **ST-**



(a) Image pyramid: parallel approach (b) Feature pyramid: bottom-up approach (c) Feature pyramid: top-down approach

Fig. 8: Three types of IQA framework based on how they extract and employ multi-scale features: the parallel, bottom-up and top-down methods [178].

MAD [180] applies the picture quality model MAD [169] along spatiotemporal slices (STS) or cuts of a space-time video. **AFViQ** [181] conducts perceptual quality assessment of foveated videos by evaluating perceptible contrast-sensitive differences in the wavelet domain. **VQM-VFD** [182] extracts a variety of features from spatiotemporal blocks to identify quality degradations. **PVM** [183] video quality perception by adaptively modeling texture masking and blur detection using a dual-tree complex wavelet transform. Additionally, several studies have focused on predicting video quality at varying frame rates. This is important since many distortions, such as stutter and motion bur, manifest over longer time intervals than existing models like VMAF and SSIM are able to measure. This is of particular importance when there is high motion, as in sports videos. In addition to the aforementioned ST-GREED model, **FRQM** [161] analyzes variable frame rate (VFR) video quality using a temporal wavelet decomposition followed by a spatial subband combination.

*2) **Deep Learning-Based FR IQA Methods**:* While deep learning-based IQA methods can be extended to VQA by applying frame-level scores followed by temporal pooling, deep neural networks and learning strategies offer significantly more scope for tailored design for VQA. The temporal dynamics of videos, including motion and frame dependencies, necessitate sophisticated NN architectures, resulting in a broader range of approaches within the VQA domain. Due to the complexity and diversity of current deep learning-based VQA methods, it is important to review IQA and VQA separately. This separation allows for a more thorough examination of the unique challenges and innovations within each area, providing a clearer understanding of their individual advancements. We summarize deep learning-based FR IQA and VQA models in Table III, which begins with an overview of the core attributes and methodologies of these techniques.

Deep learning-based FR IQA models usually rely on a typical workflow whereby deep features are first extracted from reference images, distorted images, and/or error images, and then fused and regressed into image quality scores. Depending on the level of features used for quality evaluation, these models can be further divided into deep feature-based and feature pyramid-based models, the former of which use only final-layer deep features, while the latter constructs a pyramid of features from multiple layers.

Type i: Deep feature-based IQA. Bosse et al. [184] propose a unified deep network for FR- and NR-IQA, employing a Siamese network for patch-level feature extraction, followed by feature fusion, patch-wise quality regression and pooling. For NR-IQA, the reference branch and feature fusion are omitted. **DeepQA** [185] uses a CNN to model a visual sensitivity function, computes an error map as normalized log differences between reference and distorted images, and applies sensitivity-based weighting to derive a perceptual error map. A nonlinear regressor maps the pooled error map to subjective scores. Building on DeepQA, **Ahn et al.** [186] improve distortion sensitivity prediction by using UNet [187] for spatial preservation, incorporating the reference image into input signals, and adding convolutional layers to expand the receptive field. Quality scores are calculated as the mean perceptual error map values.

Recent advancements in large models have significantly impacted fields like computer vision and natural language processing. The Segment Anything Model (SAM) has extended its utility from semantic segmentation to image quality assessment. **SAMScore** [96] evaluates semantic structural similarity in image translation tasks by computing the spatial-wise cosine similarity of SAM-derived semantic embeddings, averaged into a final score. **SAM-IQA** [97] utilizes SAM's encoder to extract features in both frequency and spatial domains. For full-reference tasks, feature distances in both domains are calculated before quality regression, while for no-reference tasks, the features are directly input into the regression module.

Type ii: Feature pyramid-based IQA. The concept of multi-scale feature extraction in MS-SSIM can be used in deep learning-based models as well, as seen in feature pyramid-based models shown in Fig. 8. These can be further divided into bottom-up approaches [25], [80] and top-down approaches [178], according to the mechanisms used to fuse features from different layers [178]. We first introduce representative bottom-up approaches as follows.

LPIPS [80] measures deep feature distances using networks like SqueezeNet [188], AlexNet [189], and VGG [190]. Reference and distorted patches are processed, channel-wise normalized, and compared using cosine distance, averaged spatially and across layers. Finally, a small network is trained to predict subjective quality using ranking loss, excelling in handling geometric distortions, which makes LPIPS popular for tasks like super-resolution. Kettunen et al. [191] highlight LPIPS's vulnerability to adversarial attacks and introduce **E-LPIPS**, improving robustness with geometric/color transformations, multi-layer distance computation, dropout, and average pooling. **DeepWSD** [192] extends LPIPS by using Wasserstein distance to assess image quality without labeled data. Deep features from VGG-16 are reshaped and compared across five stages, and additional metrics like an "EUL" index,

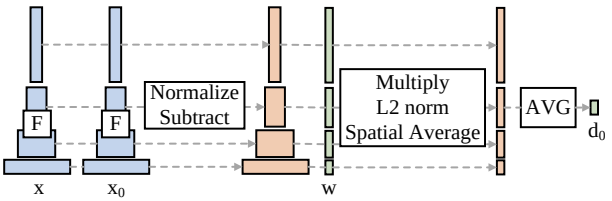


Fig. 9: The framework of LPIPS [80].

combining Euclidean norm with adaptive weights, enhance perceptual fidelity assessment. **Tariq et al.** [81] analyze pre-trained CNN feature maps to identify channels sensitive to visual distortions. Spatial frequency and orientation measures are combined into a single Perceptual Efficacy (PE) score, reflecting channel effectiveness in predicting perceptual quality. **DISTS** [25] integrates sensitivity to structural distortions and tolerance to texture resampling. Using VGG-16, it replaces max pooling with weighted L2 pooling, computes texture and structure distances via global means and correlations, and combines these into a weighted quality score. **A-DISTS** [26] enhances DISTS by incorporating locally adaptive structure and texture similarity using variance-to-mean ratios of features to weight similarities across space and channels.

Unlike the above bottom-up paradigm, **TOPIQ** [178] adopts a top-down approach for both FR and NR image quality assessment. Multi-scale features from the first five layers of ResNet-50 [193] are unified via a gated local pooling module and enhanced with self-attention blocks. Cross-scale attention is then applied progressively from high-level to low-level features, with the final features pooled and regressed into image-level quality scores.

*3) **Deep Learning Based FR VQA Methods**:* Deep neural networks for video quality assessment aim to capture both spatial and temporal quality variations, which is a challenging task due to larger data inputs and complex frame interactions. A straightforward approach extends FR IQA methods by generating frame-level scores/features followed by temporal fusion, though simple pooling often fails to model temporal distortions effectively. Hence, subjectively-inspired temporal effects have been leveraged by several models to effectively account for temporal quality. Moreover, temporal-aware neural modules, such as 3D CNNs for spatiotemporal features or Transformers for long-range dependencies, have been adopted. Deep learning-based FR VQA methods are thus categorized into temporal pooling-based and temporal NN module-based approaches. Since this review primarily focuses on deep learning-based FR/NR VQA methods, detailed design overviews are provided as below.

Type i: Temporal pooling-based VQA. A common challenge is that the simple average pooling method does not capture complex temporal dynamics. To address this, models like **FloLPIPS** [82] incorporate optical flow to weight the pooling process, enhancing sensitivity to temporal distortions. Additionally, models like **DeepVQA** [57] use spatiotemporal masking effects to adjust pooling weights, thus better identifying frame-level transitions and quality fluctuations. Similarly, **Sun et al.** [27] adopted subjectively-inspired temporal pooling to reflect human perceptual responses.

- **FloLPIPS** [82] adapts LPIPS [80] for video frame interpolation by improving temporal consistency. Optical flow between consecutive reference and distorted frames is computed using PWC-Net to generate weights for spatial pooling in LPIPS. AlexNet extracts LPIPS features, and the weights are normalized based on flow map differences.
- **C3DVQA** [58] models temporal masking effects using both 2D and 3D CNNs(thus it can be categorized to Type ii, too). The 2D CNN processes distorted and residual frames to extract spatial features, which are concatenated to form spatiotemporal features. The 3D CNN learns distortion thresholds from these features, identifying perceptually significant artifacts by masking residuals with thresholds. Fully connected layers map the artifacts to quality scores.
- **DeepVQA** [57] models spatiotemporal masking and temporal memory effects in a two-step process. First, CNNs calculate a spatiotemporal sensitivity map from distorted frames, spatial/temporal error maps, and frame differences, producing frame-level scores via perceptual error mapping. Second, a memory attention mechanism assigns significance weights to frames, enabling weighted aggregation of frame-level scores into global quality scores.
- **Sun et al.** [27] propose a framework for both FR and NR quality assessment of compressed UGC videos. FR mode calculates texture and structure similarities across CNN layers, while NR mode computes global statistics from hierarchical feature maps. Frame-level scores are derived via fully connected layers and aggregated using subjectively-inspired temporal pooling from VSFA [74].

Type ii: Temporal NN module-based VQA. In temporal NN module-based VQA models, the primary approach involves using neural networks to directly model temporal dependencies and quality variations across frames. Unlike simpler pooling methods, these models leverage advanced architectures like 3D CNNs and transformers to more comprehensively learn spatiotemporal interactions. For instance, **C3DVQA** [58] and **Chen et al.** [78] employ 3D CNNs to capture spatiotemporal quality variations, while models like **STRA-VQA** [194] incorporate transformer modules to adapt to changes in spatial and temporal resolutions. A key insight is that these models often use attention mechanisms or adaptive weighting strategies, as seen in **Chen et al.** [78] and **STRA-VQA** [194], to focus on perceptually significant distortions. This helps to enhance a model's ability to account for complex temporal dynamics and quality fluctuations, providing more accurate video quality predictions as compared to simpler pooling methods.

- **Chen et al.** [78] propose a unified framework combining spatiotemporal quality prediction and aggregation. A variant of C3DVQA [58] with a residual attention mechanism highlights noticeable distortions, while an adaptive spatiotemporal aggregation network assigns weights to aggregate quality scores across spatial and temporal dimensions.
- **DVQM-HT** [59] employs hybrid training by using VMAF as a proxy for subjective scores. A 3D CNN learns patch-wise quality measurements, aggregated into frame- and sequence-level scores using a temporal aggregation network. Training

TABLE III: Overview of deep full-reference IQA/VQA methods, indicated by *italic*/orthographic font, respectively.

Type	Method	Architecture (Feature extraction + Quality fusion)	Core block (Pretrained models or crafted modules)	Key idea
Deep feature based IQA	*Sebastian Bosse et al. [184]*	CNN + MLP & Weighted pooling	Siamese Network, Spatial weighted Pooling	Both average pooling and weighted average regression are evaluated in a unified framework for FR and NR image quality assessment.
	DeepQA [185]	CNN + MLP & Weighted pooling	Crafted CNN, Perceptual map generation	Visual sensitivity map weighted pixels on an objective error map produce a perceptual error map.
	Sewoong Ahn et al. [186]	CNN + MLP & Weighted pooling	DeepQA, Skip-Connection	Modifies DeepQA to enrich spatial information, using a Unet structure, adding layers, and directly predicting quality scores.
	SAMScore [96]	SAM + Similarity pooling	Image encoder of SAM, Cosine similarity	Leverage SAM encoder to extract semantic embeddings, based on which spatial-wise cosine similarity is calculated and averaged into the overall quality.
	SAM-IQA [97]	SAM + MLP	Image encoder of SAM, Fourier/classic convolution	Embeddings of SAM encoder are gone through spatial-frequency feature extraction, and the obtained features are regressed into the overall quality.
Feature pyramid based IQA	*LPIPS [80]*	CNN + Distance pooling	AlexNet/VGG, Distance Calculation	Calculates the distance between network activations of pairs of image patches, averaged across spatial dimensions and layers of the network.
	Markus Kettunen et al. [191]	CNN + Distance pooling	Adversarial Attack, LPIPS	Boosts LPIPS by transforming the input images, computing over all layers, applying dropout, and using averaging pooling.
	DeepWSD [192]	CNN + Distance pooling	VGG-16, Wasserstein Distance	The Wasserstein distance and Euclidean distance at each stage of CNN are measured and averaged to predict the final quality scores.
	Taimoor Tariq et al. [81]	CNN + Distance pooling	Pretrained models, Channel selection	Analyzes the capabilities of deep features for estimating quality degradations by measuring frequency and orientation selectivity.
	DISTS [25]	CNN + Similarity pooling	VGG-16, Similarity Calculation	The first FR IQA model insensitive to texture resampling, measures perceptual similarity within a deep representation.
	A-DISTS [26]	CNN + Similarity pooling	Separate Structure and Texture	Adaptively weights local structure and texture separated by a dispersion index.
	TOPIQ [178]	CNN + Feature fusion & MLP	ResNet50, Self-attention, Cross-attention	Exploits multi-scale features and incrementally transfers high-level semantic information to low-level representations in a top-down fashion.
Temporal pooling based VQA	FloLPIPS [82]	CNN + Distance pooling	LPIPS, Optical flow, Weighted spaital average	Re-designs the spatial averaging step of LPIPS by weighting by the differences of estimated optical flow of adjacent frames.
	C3DVQA [58]	2D CNN + 3D CNN & Mask pooling	2D CNN, 3D CNN	A CNN with 2D and 3D kernels learns distortion visibility thresholds which mask residual frames to exaggerate noticeable distortions.
	DeepVQA [57]	CNN + Memory attention pooling	DeepQA, Temporal aggregation	Learns spatial-temporal sensitivity similar DeepQA, and conducts temporal pooling across frames using an aggregation network.
	Wei Sun et al. [27]	CNN + MLP & Memory effect pooling	Similarity Calculation, Quality Regression	Different measurements are calculated within intermediate layers depending on whether FR or NR assessment is required, followed by regression and pooling.
Temporal NN module based VQA	C3DVQA [58]	2D CNN + 3D CNN & Mask pooling	2D CNN, 3D CNN	A CNN with 2D and 3D kernels learns distortion visibility thresholds which mask residual frames to exaggerate noticeable distortions.
	Junming Chen et al. [78]	2D CNN + 3D CNN	2D CNN, Spatiotemporal aggregation	Three distinct approaches are investigated to aggregate patch-level quality scores along both spatial and temporal axes.
	DVQM-HT [59]	3D CNN + MLP	3D CNN, shallow 2D CNN	The first model to employ a 3D CNN to predict video quality on patches, where VMAF is used to enrich the training data.
	STRA-VQA [194]	2D CNN + Transformer	ResNet50, Adaptive Weight Transformer	Spatiotemporal quality is modeled in an adaptive weight Transformer which is sensitive to the downsampling spatial and temporal resolutions.

data for 3D CNN includes a database of 614,400 patch pairs with VMAF values, and the temporal aggregation network is trained on the VMAFplus [173] database.

• **STRA-VQA** [194] addresses spatiotemporal resolution adaptive coding scenarios. Features from source and reconstructed videos are processed through an adaptive weight transformer, which accounts for spatial and temporal downsampling, followed by an FC layer to predict quality scores.

B. No-reference Video Quality Assessment

High-quality reference signals are unavailable in many practical scenarios, and therefore, the development of no-reference video quality assessment models (NR-VQA) is quite important, but remains quite challenging. NR-VQA algorithms need to be highly sensitive to distortions, while also accounting for aspects of visual content that reduce (mask) distortions, or that are more sensitive to distortion. **Knowledge-driven** models often rely on perceptually relevant statistical properties of pictures and videos, similar to those underlying successful FR models like VIF and VMAF. **Deep learning-based** methods learn to predict quality degradations from large databases of distorted visual data. Similar to the approach taken in Section IV-A, we begin by briefly reviewing learning no-reference or blind IQA (BIQA) models and blind VQA (BVQA) methods, then study data-driven deep learning-based blind models.

*1) **Knowledge-driven BVQA Methods**:* Traditional VQA methods typically follow a knowledge-driven approach, whereby handcrafted features are manually extracted to assess video quality. For low-level visual feature-based VQA methods, these features are carefully designed based on prior domain knowledge, aiming to capture key aspects of video quality such as sharpness, contrast, or motion artifacts. The quality evaluation process relies heavily on predefined rules, which, while effective to some extent, may struggle to generalize across diverse video contents and distortion types. These

have been largely replaced by more powerful general-purpose BIQA/BVQA models that are based on measurements of distortion-induced statistical deviations of bandpass processed images/videos from perceptually relevant models of natural scene statistics (NSS) [39], as discussed earlier.

Type i: Low-level visual feature-based VQA. There are a few works performing quality measurement based on distortion-specific or low-level visual features, such as local spatial features, *e.g.*, edge strength, contour shape, and motion intensity. The earliest (BIQA or) BVQA algorithms were designed to analyze and quantify single distortion types, such as blockiness, blur, ringing, sharpness, and space compression [195]–[198], based on the measurement of a small number of image or frame level features, For example, **CPBDM** [195] and **LPCM** [196] were developed for blur evaluation, while **NJQA** [199] and **JPEG-NR** [197] aimed to assess noise and JPEG compression, respectively. **TLVQM** [200] computes diverse handcrafted quality-aware features, including spatial attributes, motion-induced statistics, and aesthetics features, then train a shallow regressor to predict video quality scores. **CORNIA** [201] and **HOSA** [202] employ unsupervised dictionary learning techniques to learn distortion codebooks from local features extracted from image patches, through which global quality-aware image representations are obtained.

Type ii: Neurostatistics-based VQA. NSS-based methods are effective due to the fact that high-quality natural images and videos consistently follow specific statistical patterns, which are systematically disrupted by distortions [48], [203]. For example, **BRISQUE** [48] computes a small set of spatial bandpass and locally normalized features, then uses a support vector regressor (SVR) to learn mapping from them to human opinion scores. Similarly, **GM-LOG** [49] extracts statistical features from smoothed gradient and Laplacian-of-Gaussian bandpass spaces, while **HIGRADE** [51] computes gradient features in the LAB and gradient of LAB color spaces. A variety of color spaces and perception-driven transforms were utilized to extract a larger number of perceptually relevant NSS features in the "bag of features" model called **FRIQUEE** [52].

Building on top of IQA models, a variety of blind VQA methods have been proposed to analyze the perceptual quality based on spatio-temporal scene statistics. **V-BLIINDS** [54] utilizes a spatio-temporal model of the statistics of DCT coefficients of local frame differences to predict perceptual video quality scores. **Li et al.** [55] employ spatiotemporal natural video statistics in the 3D-DCT domain to predict perceptual quality. **VIDEVAL** [204] implements an ensemble model that fuses features selected from several top-performing BVQA models, using a supervised feature selection procedure on a large UGC superset. **STS-QA** [53] extracts spatiotemporal statistical features along different orientations of video space-time slices (STS) that capture directional global motion, then regresses the feature vector into video quality predictions using a shallow learner. **FAVER** [205] is the first NR-VQA model able to account for frame rate variations, using extended models of the natural statistics of space-time wavelet-decomposed video signals.

The majority of existing BIQA or BVQA algorithms belong to the "opinion-aware" category, wherein a learned regression model, either deep or shallow, is trained on databases of distorted videos that have been human-labeled in the form of mean opinion scores (MOS). However, these models can suffer from limited generalization capability on real-world images/videos, where multiple, unknown, commingled distortions may arise that are not present in the training set. Therefore, "opinion-unaware" (OU) or "completely blind" models which do not rely on training on human-labeled videos, have aroused considerable research insterest [47], [203], [206]–[210]. Among these, **NIQE** [47] and **IL-NIQE** [206] were designed using perceptually-inspired neuro-statistical features in the spatial domain that capture deviations from perceptually relevant NSS models. Following the design framework of NIQE [47], **NPQI** [209] explores NSS features from a local binary map and locally normalized coefficients of images, while **SNP-NIQE** [210] measures structural variations as well as naturalness deviations. **VIIDEO** [203] models the temporal regularities of natural videos, using them to assess video quality. A more recent completely blind BVQA model targeting UGC videos, called **STEM** [208], quantifies losses of "perceptual straightness" [211] to measure temporal quality. Similarly, **TPQI** [212] measures the perceptual straightness and compactness of trajectories of video representations. Another type of OU quality predictor is based on pseudo-references, such as **SLEEQ** [207] and **BPRI** [213]. The recent **VIQE** [50] model fuses a variety of patch- and frame-wise video statistics, to capture distortion-predictive space-time statistical irregularities of videos.

*2) **Deep Learning Based BIQA Methods**:* Following Section IV-A2, we review deep learning-based blind IQA and VQA methods separately. Such a focused review will enable a deeper analysis of how these no-reference models handle temporal dynamics, motion artifacts, and frame consistency without relying on reference information, providing a more comprehensive understanding of how deep learning techniques are specifically adapted for video-related challenges. Similarly, we first discuss existing deep learning-based BIQA methods.

Type i: CNN feature extraction and FC fusion. This category of blind IQA models uses CNNs for feature extraction, followed by fully connected (FC) layers for quality score prediction. Some models [112] adopt pre-trained networks, benefiting from the generalization ability gained from large datasets. This allows for faster convergence and robust performance. Others [214]–[216] use end-to-end training with custom-built CNNs, which are tailored to specific IQA tasks, offering greater flexibility and potential for capturing unique distortion patterns. After feature extraction, FC layers fuse the features to produce final quality scores. **Kang et al.** [214] introduce an early BIQA framework using a five-layer CNN to predict patch-level quality, averaged to produce image-level scores. **Bosse et al.** [215] utilize a deep CNN for patch-level quality predictions, aggregated using either average or weighted-average pooling. **MEON** [216] combines distortion identification and quality prediction via two sub-networks sharing learned features, with outputs fused to estimate overall quality. **NIMA** [217] and **PQR** [218] predict probability distributions of quality scores instead of scalar values. **DB-CNN** [219] assesses perceptual quality of synthetic and au-

thentic distortions with two specialized networks, combined using a deep bilinear model. **PaQ-2-PiQ** [112] introduces models leveraging ResNet-18 [193] as the backbone and RoIPool [220] for aggregating patch and whole-image quality. The proposed P2P Feedback model, which concatenates patch and image scores for final prediction, achieves the best performance. **Sun et al.** [221] propose a staircase neural network to capture multi-level features and use an iterative training strategy for mixed database training.

Type ii: CNN feature extraction and special fusion. This class of blind IQA methods builds on CNN-based feature extraction but incorporates specialized fusion techniques to enhance quality prediction. Unlike traditional FC fusion approaches, these models apply innovative strategies such as graph-based learning [222], fuzzy logic [223], or hallucination [224], [225] techniques to better capture the complex relationships between image distortions and perceived quality. **QCN** [226] estimates quality by iteratively updating image embeddings and "score pivots" through geometric order learning, with the final score derived from the nearest pivot. **GraphIQA** [222] uses distortion graph representations for perceptual quality modeling, enhanced by a discrimination network and a fuzzy prediction network. **Gao et al.** [223] introduce a fuzzy-based network that maps VGG16 image embeddings to quality score distributions. **FPR** [224] and **HIQA** [225] assess quality by referencing hallucinated or pristine versions of distorted images. **BIECON** [227] mimics FR-IQA models by predicting local patch scores generated by an FR-IQA algorithm.

Type iii: Transformer-based IQA. Transformers have shown strong potential as a generalist model for broad computer vision tasks [90], [91], [228]–[230]. Transformer-based IQA models leverage the powerful self-attention mechanism of Transformers to capture both global and local dependencies in images, making them particularly well-suited for image quality assessment tasks. **MUSIQ** [231] and **TRIQ** [232] use Transformer encoders to process images of varying resolutions, employing embedding methods and learnable classification (CLS) tokens to predict quality scores via fully connected layers. **DEIQT** [233] extends this approach by using CLS tokens from the Transformer encoder as queries in the decoder to learn quality-aware embeddings, which are regressed into quality scores using an MLP. **TReS** [234] combines a CNN and a Transformer encoder to extract local and global features, with specialized loss functions improving prediction accuracy and robustness. **MANIQA** [235] employs a transpose attention block and Swin Transformer to enhance interactions between global and local features extracted by a Vision Transformer.

Type iv: Multi-task learning. This category of blind IQA approaches focuses on multi-task learning, where models are designed to address challenges like overfitting, learning efficiency, and the ability to generalize to unseen distortion types, enabling more robust and scalable IQA solutions across multiple domains. They incorporate techniques such as continual learning, meta-learning, and incremental learning to adaptively improve the model's capacity to assess image quality in diverse scenarios. **Zhang et al.** [236] address continual learning in BIQA by enabling models to learn continually from IQA dataset streams. **MetaIQA** [237] uses meta-learning to create a prior NR-IQA model for distortion-specific tasks, fine-tuning it to predict quality for unknown distortions. **Su et al.** [238] propose a two-stream architecture with a shallow network learning distortion manifolds and a deep network refining quality predictions, enhanced by masked distortion labeling and gradual weighting strategies. **SLIF** [239] performs multi-task quality assessment using incremental learning with scalable memory units that prune unimportant neurons during training. **Li et al.** [240] introduce a channel modulation kernel to capture intra- and inter-domain attention, combined with a multi-dataset learning strategy for mixed datasets. **Wang et al.** [241] develop an online mining pipeline where a full and pruned model are trained together, using prediction disagreements to mine hard examples for enhancing the full model. **Zhang et al.** [242] propose task-specific normalization parameters for each dataset to produce weighted quality scores for overall predictions.

Type v: Unsupervised and self-supervised learning. This class of IQA methods focuses on unsupervised and self-supervised learning techniques, aiming to overcome reliance on large labeled datasets for quality prediction. By leveraging unsupervised learning or self-supervised learning, these approaches learn quality-aware representations without requiring explicit quality labels. These methods offer greater flexibility and scalability, particularly in scenarios where labeled data is scarce, while also promoting improved generalization across different datasets and distortion types. **CONTRIQUE** [243] is an unsupervised contrastive learning-based model that leverages a distortion classifier trained on diverse distortions, achieving SOTA performance across leading IQA databases. **Shukla et al.** [244] use a variational autoencoder GAN trained on pristine images to assess quality by comparing latent distributions between pristine and distorted images. **Re-IQA** [245] trains two encoders in an unsupervised setting to capture low-level quality and high-level content features, with a regression model mapping these to quality scores. **Babu et al.** [246] propose a self-supervised contrastive learning framework, pretraining on synthetic data and fine-tuning on authentic data using content separation and positive samples only. **ARNIQA** [247] models the distortion manifold in a self-supervised manner, maximizing similarities between representations of equally distorted images to enhance robustness and generalization across datasets. **Zhao et al.** [248], similar to CONTRIQUE, introduces a quality-aware pretext task and trains an IQA model using diverse augmented distortions constrained by a quality-aware contrastive loss.

Type vi: Large multimodality model-based IQA. **CLIP-IQA** [100] assesses image quality by leveraging the vision-language prior in CLIP [8]. It uses cosine similarity between image embeddings and text embeddings of antonym prompts ('good photo' and 'bad photo') to compute quality scores via softmax. The variant **CLIP-IQA+** fine-tunes prompts for improved accuracy, supporting zero-shot and fine-tuned quality assessments. Inspired by CLIP-IQA, **LIQE** [249] uses a multitask learning scheme to optimize

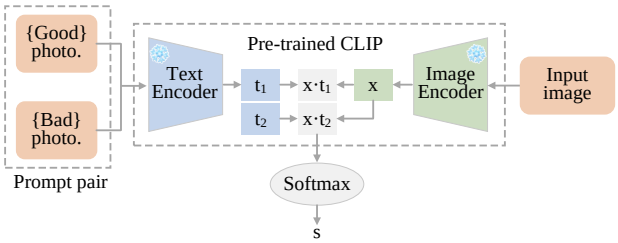


Fig. 10: The framework of CLIP-IQA [100].

CLIP end-to-end, and **TCDs** [250] employs explicit content descriptions for aesthetics evaluation.

Emerging multimodal large language models (MMLLMs) demonstrate strong visual understanding capabilities [251]–[254]. **Q-Bench** [103] explores MMLLMs for image perception, deriving quality scores through softmax pooling of 'good' and 'poor' tokens. However, its performance is limited by the lack of low-level visual datasets. **Q-Instruct** [11] addresses this by fine-tuning MMLLMs on a curated low-level instruction dataset, while **Q-Align** [255] uses a human-emulating syllabus to improve quality score predictions. **Co-Instruct** [256] extends MMLLMs for visual comparisons by fine-tuning on the Co-Instruct-562K dataset with image-text interleaved features, achieving superior performance on multiple tasks.

*3) **Deep Learning Based BVQA Methods**:* Unlike IQA, BVQA must address the temporal dimension, making temporal distortion modeling a critical challenge, especially without reference information. This requires inferring quality solely from distorted content. Temporal aggregation methods range from simple score averaging to advanced techniques like 3D CNNs and attention mechanisms, enabling models to capture temporal dependencies and assess quality effectively. Recent deep learning-based BVQA models, listed in Table IV, are categorized into five classes based on their architectures and temporal aggregation strategies. Following IV-A3, our discussion focuses on their design approaches.

Type i: 2D CNNs with simple score/feature averaging. A straightforward approach for video quality prediction involves averaging frame-level scores or features from spatial-only neural networks. While computationally efficient and leveraging established image-based models, these methods often overlook temporal dynamics. Some approaches, like **SIONR** [69], combine spatial and temporal features or calculate frame-wise variations, but still prioritize spatial analysis. Simple pooling methods, such as **GAP** [66], [69], [70], offer efficiency but may miss critical spatial and temporal details. Advanced techniques, like saliency-weighted pooling in **SWDF-DF-VQA** [70], address this by focusing on regions of interest. These methods balance simplicity and efficiency while limiting temporal complexity.

- **Varga et al.** [68] use Inception-V3 [268] and Inception-ResNet-V2 [269] for frame-level feature extraction, applying various pooling methods (*e.g.*, average, median) to obtain video-level features. SVRs map these features to quality scores.
- **NAVE** [257] employs deep autoencoders to process a global feature matrix of extracted NSS and spatial-temporal features. Frame-level quality scores are obtained via a DeepNet and averaged to estimate video-level quality.
- **CNN-TLVQM** [71] combines TLVQM temporal features with ResNet-50 spatial features which are weighted by a spatial activity map. Temporal quality variations are modeled either using LSTM or pooling with an SVR.
- **SIONR** [69] fuses temporal variations in high-level semantic features (ResNet-50) and low-level statistical features, predicting frame-level scores that are temporally pooled for video-level quality.
- **CenseoQoE** [66] applies a lightweight ConvNet to extract frame-level features, with scores averaged through GAP to produce video-level quality.
- **RAPIQUE** [45] combines handcrafted neurostatistical spatial-temporal features with ResNet-50 deep features to produce frame-level vectors, pooled and mapped via SVR for video quality scores.
- **SWDF-DF-VQA** [70] uses multiple pre-trained CNNs for frame-level feature extraction, applying saliency-weighted GAP to emphasize important regions. Frame-level features are pooled and regressed to predict video quality.
- **NR-VMAF** [258] targets compression and scaling artifacts by extracting patch-level features using CNN/DNN, aggregating frame scores into video-level quality predictions.

Type ii: 2D CNN with temporal aggregation networks. Temporal quality modeling is enhanced in models that capture frame-level interactions and temporal memory effects. Models like **MLSP-VQA** [72], **Varga et al.** [73], and **VSFA** [74] use recurrent networks (LSTMs or GRUs) to aggregate frame-level features, addressing motion and quality fluctuations overlooked by earlier frame-based approaches. **VSFA** [74] and **AB-VQA** [88] emphasize perceptual hysteresis effects, simulating non-linear human responses to quality drops and recoveries for more accurate predictions. Models like **RIRNet** [75] and **STDAM** [65] incorporate spatial pyramid pooling and multi-scale attention mechanisms to capture both local and global temporal features, handling complex video content such as rapid motion or subtle changes. Attention mechanisms in **STDAM** [65] and **GSTVQA** [89] enhance quality predictions by focusing on salient regions and key frames, balancing spatial and temporal distortions for videos with uneven quality.

- **MLSP-VQA** [72] uses Inception-ResNet-v2 [269] to extract spatial features from video frames, which are pooled via GAP to form frame-level feature vectors. Temporal aggregation is achieved using either an LSTM or a temporal CNN, enabling video-level quality predictions.
- **Varga et al.** [73] employs AlexNet [189], Inception-V3 [268], and Inception-ResNet-V2 [269] pre-trained on KoNViD-10K for transfer learning. Frame-level features are extracted and processed through a two-layer LSTM with a fully connected layer to predict video quality.
- **VSFA** [74] combines ResNet-50 [193]-based feature extraction with GRUs to capture temporal dependencies. It models perceptual hysteresis effects by combining memory quality and current quality into frame-level scores, which are pooled into

TABLE IV: Overview of deep no-reference VQA methods.

Type	Method	Architecture (Feature extraction + Quality fusion)	Core block (Pretrained models or crafted modules)	Key idea
2D CNNs with simple score/feature averaging	Domonkos Varga et al. [68]	CNN + SVR	Inception-V3, Inception-ResNet-V2	A pre-trained CNN computes frame-level feature vectors after being fine-tuned on an image database.
	NAVE [257]	Handcrafted features + Autoencoders	Autoencoder, Classification Layer	Deep autoencoder takes NSS and temporal-spatial features as input and generates more descriptive quality features.
	CNNTLVQM [71]	Handcrafted Features & CNN + LSTM/SVR	TLVQM, ResNet-50, Sobel Filters, LSTM	VQA model that combines handcrafted motion features and CNN-based spatial features pooled with spatial activity.
	SIONR [69]	CNN + MLP	ResNet-50, GAP	Temporal variations of semantic features and low-level distortions are computed to predict frame-level quality.
	CenseoQoE [66]	CNN + MLP	Shallow CNN, GAP	A unified model of both FR and NR quality assessment of images and videos based on a backbone network and FC layers.
	RAPIQUE [45]	Handcrafted features & CNN + SVR	ResNet-50	A space-time bandpass statistics model that combines quality-aware NSS features with deep semantic features.
	SWDF-DF-VQA [70]	CNN + SVR/GPR	Pre-trained CNN, SWGA	Parallel pre-trained CNNs extract quality distortions which are then temporally pooled and saliency weighted.
	NR-VMAF [258]	CNN + MLP	Pre-trained CNN, Visual saliency extraction	Deep features are extracted by pre-train CNNs from frame filtered by visual saliency, which are regressed and averaged pooling.
2D CNN with temporal aggregation networks	MLSP-VQA -FF/RN [72]	CNN + MLP or CNN + LSTM	Inception-ResNet-v2, LSTM, Feed Forward Network	Global average pooling is performed on the activation maps of all kernels in the stem of the pre-trained network
	Domonkos Varga et al. [73]	CNN + LSTM	Pre-trained CNN, LSTM	Frame-level feature vectors are computed by a pre-trained CNN and transferred to an LSTM network.
	VSFA [74]	CNN + GRU & Memory effect modeling	ResNet-50, GRU	Content-dependency and temporal-memory effects are modeled by classification CNN and GRU networks, respectively.
	MGQA [259]	CNN + GRU & Memory effect modeling	Grid mini-patch sampling, VSFA	Generate heavily squeezed videos by spatiotemporal sampling, from which a reliable VQA model predict quality scores.
	RIRNet [75]	CNN + RNN & Motion effect modeling	ResNet-50, GRU	Motion effect of multiple temporal frequencies is characterized by a hierarchical recurrent temporal modeling scheme.
	STDAM [65]	CNN + BiLSTM	ResNet-18, Graph Convolution,BiLSTM	A graph convolution and attention module enhances the deep frame-level features which are evolved by a BiLSTM network.
	MDTVSFA [260]	CNN + GRU & Memory effect modeling	VSFA, Nonlinear Mapping	VQA model learned using a mixed datasets training strategy to account for various contents and distortions.
	AB-VQA [88]	CNN + GRU & Memory effect modeling	VGG-16, GRU	Attention module extracts local distortions while GRU and temporal pooling layers model memory effects.
	Junfeng Li et al. [77]	CNN + BiGRU & Memory effect modeling	ResNet-50, Inception-V3, BiGRU	Dual network generates frame features which are processed by BiGRU and two temporal pooling modules.
	GSTVQA [89]	CNN + GRU & Temporal pyramid pooling	VGG-16, GRU	Unified Gaussian distribution constraints are imposed on deep features to obtain more generalized quality representations.
	2BiVQA [85]	CNN + BiLSTM	Pre-trained CNN, Bi-LSTM	Two Bi-LSTM networks capture both short-range dependencies and long-range dependencies to account for memory effects.
3D CNNs / Transformation	SACONVA [63]	Handcrafted features + CNN	3D Shearlet Transform, 1D CNN, Autoencoder	Spatiotemporal features are extracted by a 3D shearlet transform and exaggerated into discriminative features by CNN.
	V-MEON [67]	3D CNN + MLP	3D CNN, GDN	Multi-task DNN framework not only predicts perceptual quality but a probabilistic prediction of codec type.
	Rui Hou et al. [60]	2D CNN + 3D CNN	VGG-16, 3D CNN	Bin-based average pooling is applied on a 3D CNN to calculate spatiotemporal information while reducing over-fitting.
	Patch-VQ [4]	2D & 3D CNN + Time series regression	PaQ-2-PiQ, ResNet3D, InceptionTime	Space-time features are extracted by 2D and 3D network streams, then a time series network processes the pooled features.
	CoINVQ [79]	2D & 3D CNN + FC	EfficientNet-b0, D3D	Video quality is analyzed from multiple aspects: content, distortion, compression level, and temporal aggregation.
	Wei Sun et al. [61]	2D & 3D CNN + FC	ResNet-50, SlowFast, MLP	Frame-level and chunk-level deep features are calculated, regressed, and pooled to obtain video-level quality scores.
	Bowen Li et al. [76]	2D & 3D CNN + GRU & Memory effect modeling	ResNet-50, SlowFast, GRU	Transfers knowledge from IQA datasets of authentic distortions and large-scale action recognition to learn feature extractors.
	MD-VQA [126]	2D & 3D CNN & Handcrafted features + MLP	EfficientNetV2, ResNet3D-18, Handcrafted features	Pretrained 2D CNN, 3D CNN, and handcrafted distortion descriptors separately extract semantic, distortion, and motion quality-aware features.
	UCDA [261]	3D CNN + MLP	C3D, unsupervised domain adaptation	A curriculum-style unsupervised domain adaptation for cross-domain generalization by progressively adapting based on prediction confidence.
	Wenhao Shen et al. [262]	3D CNN + MLP	3D CNN, Spatiotemporal pyramid attention	Hierarchical motion information at different temporal scales is fed into spatiotemporal pyramid attention for cross-scale dependency modeling.
Transformer-based models	StarVQA [93]	Transformer + MLP	Encoder, Time Attention, Space Attention	First work to leverage Transformers on the VQA problem. A vectorized loss function is designed to help train the Encoder.
	Junyong You [92]	CNN + Transformer & MLP	Encoder, Channel Attention, Spatial Attention	A long short-term convolutional Transformer architecture predicts perceptual video quality from frame features extracted by a CNN.
	DisCoVQA [94]	Transformer + Transformer & MLP	Swin-Transformer, Transformer Encoder, Decoder	Temporal variations and quality attention effects are modeled by a distortion extraction module and a content attention module.

(Continued)

Type	Method	Architecture (Feature extraction + Quality fusion)	Core block (Pretrained models or crafted modules)	Key idea
Transformer -based models	FAST-VQA [95]/ FasterVQA [263]	Transformer + MLP	Swin-Transformer, Grid Mini-patch Sampling	A grid mini-patch sampling scheme is used to reduce the computational cost of high-resolution VQA.
	SAMA [264]	Transformer MLP	Multi-granularity sampling, Swin Transformer	A novel visual data sampling strategy is proposed based on multi-granularity pyramid sampling and spatiotemporal masking.
	DOVER [125]	a) Transformer + MLP b) CNN + Cosine Similarity	inflated-ConvNext, Video Swin Transformer	Decompose video quality perception into aesthetic and technical perspectives and extract features from two pre-trained models.
	SSL-VQA [265]	Transformer + 3D CNN & Distribution distance	Video Swin Transformer, Knowledge transfer	Integrate a knowledge transfer mechanism within a semi-supervised learning framework to effectively utilize a limited amount of labeled data.
Large multimodality model-based methods	PTM-VQA [99]	2D & 3D CNN & CLIP visual encoder + MLP	Pretrained model set including CLIP	Features extracted from pretrained models are integrated to learn final quality representations and scores supervised by diverse loss.
	COVER [98]	2D CNN & CLIP visual encoder + MLP	CLIP, ConvNet, Swin Transformer	Model video quality in semantic, aesthetic, and technical branches with features fused by cross-gating blocks and regressed into quality score.
	KSVQE [127]	3D Transformer + Attention & MLP	3D-Swin Transformer, CLIP, Cross-/self-attention	A CLIP-based content understanding module select quality-aware patches to fed into Transformer-based quality regression module for quality prediction.
	Wen Wen [266]	CLIP visual encoder + MLP	CLIP, ResNet-18, SlowFast	A base quality predictor responds to basic visual distortion, with spatial and temporal rectifiers capturing resolution and framerate changes.
	BUONA-VISTA [101]	a) CLIP + Cosine similarity b) Crafted scores + Pooling c) Crafted scores + Normalization	CLIP, NIQE, TPQI	CLIP with multi-pair antonym prompts forms the semantic quality index while spatial and temporal quality is compensated by NIQE and TPQI.
	MaxVQA [102]	CLIP & Transformer + Cosine similarity	CLIP, FAST-VQA	Cosine similarity is computed between CLIP textual features of learnable contrastive prompts and fused features from CLIP and FAST-VQA.
	ZE-FESG [267]	CLIP + Cosine similarity	CLIP	52 descriptions accounting for technical and abstract aspects as well as the video frames are fed into CLIP to obtain features of multiple dimensions.
	Q-Align [255]	LMM + rating level conversion	mPLUG-Owl-2 (LMM)	LMMs are firstly fine-tuned to rate like a human with rating levels and then fed with discrete video frames to rate video quality level.
	LMM-VQA [104]	CLIP & 3D CNN + LLM	Llama-3-8b-Instruct (LLM), CLIP, SlowFast	Spatial features from CLIP and temporal feature from SlowFast are projected into text-guided tokens, and are fed into LLM along with prompt tokens.

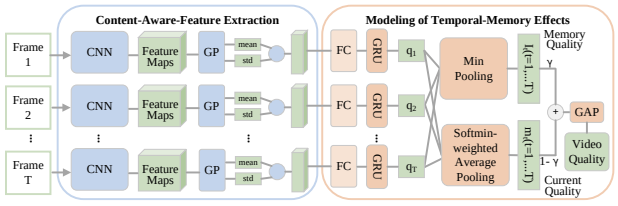


Fig. 11: The framework of VSFA [74].

overall video quality predictions.

- **MGQA** [259] builds on VSFA by incorporating spatial and temporal grid sampling. It evaluates classical temporal sampling methods from video understanding tasks and spatial Grid Mini-patch Sampling (GMS), and uses them to generate squeezed video representations for efficient quality assessment.
- **RIRNet** [75] uses ResNet-50 [193] with spatial pyramid pooling [270] to extract motion-aware features. A recurrent temporal dimension (RTD) module models motion over scales, while a recurrent temporal resolution (RTR) module processes nested temporal features, producing final quality predictions via a linear layer.
- **STDAM** [65] combines graph convolutions, attention mechanisms, and optical flow modules. Graph convolutions extract multi-scale features, while channel and spatial attention enhance salient regions. Motion features are captured using optical flow, and bi-LSTMs aggregate frame-level features for final predictions.
- **MDTVSFA** [260] decomposes VQA into relative, perceptual, and subjective quality tasks. Relative quality is modeled with VSFA as the backbone, while perceptual alignment uses a 4-parameter nonlinear mapping and dataset-specific adjustments via an FC layer account for subjective quality. Mixed data training enhances generalization.
- **AB-VQA** [88] focuses on UGC videos with complex, uneven distortions. VGG16 [190] extracts frame-level features, attention modules capture long-range dependencies, and HVS-inspired temporal pooling in VSFA models perceptual hysteresis effects. Final quality scores are pooled via minimum and softmin-weighted averaging methods.
- **Li et al.** [77] combines ResNet-50 [193] and Inception-V3 [268] for feature extraction, with global pooling and convolution fusion. A bi-GRU models temporal dependencies, while regression blocks, including a temporal memory and Gaussian regression block, refine predictions. Final quality scores are obtained via a weighted sum of components.
- **GSTVQA** [89] extracts frame-level features using VGG16 [190], which are processed with channel attention to create multi-scale features. A Gaussian normalization layer reduces domain gaps, and pyramid pooling aggregates short- and long-term quality features for robust generalization.
- **2BiVQA** [85] extracts patch-level features from video frames using a pre-trained CNN. Bi-LSTM modules handle spatial pooling of patches and temporal pooling of frames, with an FC layer producing overall quality predictions.

Type iii: 3D CNNs / Transformation. 3D CNN-based models process consecutive frames as 3D space-time volumes, enabling simultaneous spatial and temporal feature extraction. This approach, used in models like **V-MEON** [67], **You et al.** [62], and **Hou et al.** [60], naturally integrates motion dynamics and simplifies architectures compared to 2D CNNs with additional temporal pooling mechanisms. Ad-

vanced methods like **Li et al.** [76] and **Sun et al.** [61] further refine predictions by incorporating HVS-inspired hysteresis pooling, aligning better with human perception. Multi-scale approaches, such as **Patch-VQ** [4] and **Shen et al.** [262], leverage pyramidal structures and attention mechanisms to capture fine-grained and global spatiotemporal features. Hybrid models like **CoINVQ** [79] and **MD-VQA** [126] combine spatial, motion, and semantic features to handle diverse distortions and artifacts. **SACONVA** [63] utilizes 3D shearlet transforms to extract spatiotemporal features before feeding them into a CNN.

- **SACONVA** [63] uses 3D shearlet transforms for spatiotemporal feature extraction, and features are then processed by a 1D CNN initialized with autoencoders. Logistic regression predicts quality, while a softmax classifier identifies distortion types.
- **V-MEON** [67] combines 2D and 3D CNN layers to extract features for both quality prediction and codec classification. Final scores are derived as the inner product of probability and quality vectors.
- **You et al.** [62] employs a 3D CNN for clip-level scores, aggregated via an LSTM and fully connected layers to produce video-level quality predictions.
- **Hou et al.** [60] extracts spatial features from eight-frame blocks using VGG-Net [190], with a 3D CNN and bin-based pooling to model temporal and spatial dynamics for quality prediction.
- **Patch-VQ** [4] combines 2D and 3D features with ROIPool and SOIPool layers. An InceptionTime network [271] predicts local and global quality, leveraging labels from the proposed LSVQ database.
- **CoINVQ** [79] uses ContentNet, DistortionNet, and CompressionNet to capture content type, distortion type, and compression levels, aggregated through temporal pooling for final predictions.
- **Sun et al.** [61] combines ResNet-50 for spatial features and SlowFast R50 for motion features. Multi-scale fusion weights quality scores computed at three resolutions for improved accuracy.
- **Li et al.** [76] integrates ResNet-50 [193] spatial features and SlowFast [272] motion features. Temporal pooling via HVS-inspired hysteresis yields final video quality predictions.
- **MD-VQA** [126] extracts hybrid features using EfficientNetV2 for semantics, ResNet3D-18 for motion, and handcrafted features for UGC live video quality predictions.
- **UCDA** [261] uses a pre-trained C3D [273] backbone for feature extraction, applying unsupervised domain adaptation between labeled source videos and unlabeled target videos.
- **Shen et al.** [262] captures multi-scale motion through a spatiotemporal feature pyramid, with pyramid attention modules for channel and spatial selective sensitivity.

Type iv: Transformer-based models. Transformers, built on self-attention mechanisms, excel at modeling long-range and contextual dependencies, making them highly effective for video quality assessment, where spatial and temporal features must be captured across extended sequences. Attention modules in Transformer-based models selectively focus

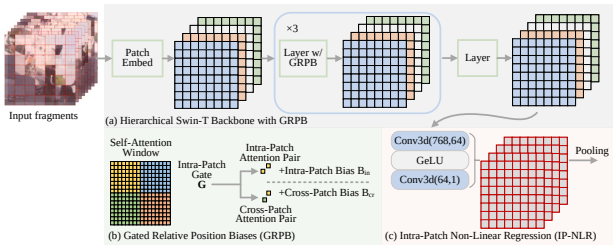


Fig. 12: The framework of FAST-VQA [95].

on important spatial-temporal regions. For example, **DisCoVQA** [94] uses distortion-aware tokens to identify degraded regions, while **StarVQA** [93] applies attention to non-overlapping space-time patches, enhancing robustness across diverse content types and uneven distortions. Efficient handling of large video data is a key focus. Models like **FAST-VQA** [95] and **FasterVQA** [263] reduce computational costs using techniques like Grid Mini-patch Sampling (GMS), preserving sensitivity to quality features while improving efficiency. Multi-level feature extraction and hierarchical attention are also prominent trends. Models like **PHIQNet** [92] and **DisCoVQA** [94] extract features at spatial, temporal, and clip levels, then apply attention mechanisms to fuse these into unified quality predictions. This streamlined overview highlights Transformers' adaptability and efficiency in video quality assessment.

- **StarVQA** [93] divides video frames into non-overlapping space-time patches, which are encoded into spatiotemporal vectors with a learnable label token. Encoding blocks apply temporal and spatial attention, and outputs are processed via LayerNorm and a fully connected (FC) network. A vectorized regression loss converts outputs into quality scores using softmax and linear SNR mapping.
- **PHIQNet** [92] combines ResNet-50 [193] pyramid feature maps with channel and spatial attention blocks to produce global feature vectors. A long short-term convolutional Transformer (LSCT) fuses clip-level features (processed by a 1D CNN) into overall quality predictions, capturing both short- and long-term dependencies.
- **DisCoVQA** [94] utilizes a Spatial-Temporal Distortion Extraction (STDE) module with a Swin-T backbone to extract distortion-aware tokens and scores. A Temporal Content Transformer (TCT) computes attention weights on these scores, yielding final video quality predictions.
- **FAST-VQA** [95] reduces computational cost by 97.6% through Grid Mini-patch Sampling (GMS), extracting spatially aligned mini-patches (fragments). These fragments are processed by a Fragment Attention Network (FANet), which uses Gated Relative Position Biases (GRPB) and intra-patch non-linear regression (IP-NLR) to predict quality. **FasterVQA** [263] introduces a temporal GMS module for 4x greater efficiency than FAST-VQA.
- **SAMA** [264] implements a scalable and masking-based video sampling method to preprocess input data into multi-scale representations, improving compatibility with base IQA/VQA models.
- **DOVER** [125] splits quality evaluation into two branches:

a technical branch using Video Swin Transformer [274] for objective metrics and an aesthetic branch using inflated-ConvNext [275] for subjective preferences. The two branches supervise separate quality components and combine for overall predictions.

• **SSL-VQA** [265] leverages semi-supervised learning with a Video Swin Transformer trained via contrastive loss for quality-aware representations. A regressor and a distance model leverage labeled and unlabeled data to generate final quality scores by averaging predictions.

Type v: Large multimodality model-based VQA methods. The visual encoder of CLIP [8], trained on the WebImageText database, effectively models semantic-level quality of video frames. Models like **PTM-VQA** [99], **COVER** [98], **KSVQE** [127], and **Wen et al.** [266] utilize CLIP's pre-trained encoder for semantic feature extraction to enhance quality prediction. CLIP's natural language supervision further enables it to generate specific quality descriptions by matching visual and textual features. Models like **BUONA-VISTA** [101], **MaxVQA** [102], and **ZE-FESG** [267] use antonym prompts and frame embeddings to calculate quality scores based on feature affinities. Large multimodal models (LMMs) extend this capability by combining robust visual and textual representations. For instance, **Q-Align** [255] predicts human-aligned quality ratings, while **LMM-VQA** [104] integrates spatial and temporal tokens to derive text-based quality scores. These LMM-based approaches showcase the potential to model complex video quality patterns across diverse contexts.

• **PTM-VQA** [99] uses pre-trained models, including CLIP's visual encoder, to extract quality-related features integrated into a unified latent space. Intra-Consistency and Inter-Divisibility (ICID) losses ensure feature consistency and separability for effective clustering, achieving state-of-the-art performance on multiple VQA datasets.

• **COVER** [98] evaluates video quality across technical, aesthetic, and semantic dimensions. A Swin Transformer assesses technical quality, a ConvNet analyzes aesthetics, and CLIP captures semantic features. A cross-gating block fuses these branches to deliver holistic quality scores, excelling on UGC datasets.

• **KSVQE** [127] is designed for short-form video quality, combining a 3D Swin Transformer backbone with modules for quality-aware region selection (QRS), content-adaptive modulation (CaM), and distortion-aware modulation (DaM). CLIP aids in extracting semantic information, enabling superior handling of quality-relevant regions and complex distortions.

• **Wen et al.** [266] combines a base quality predictor (using sparse frames and CLIP for semantic features) with spatial and temporal rectifiers to adjust scores based on resolution and frame rate. This approach ensures adaptability to diverse video attributes.

• **BUONA-VISTA** [101] is an opinion-unaware model integrating high-level semantic metrics from CLIP (via prompts like "high quality" vs. "low quality") with spatial and temporal naturalness indices. Aggregated scores are computed through Gaussian normalization and sigmoid rescaling, effectively capturing technical and aesthetic quality.

• **MaxVQA** [102] combines CLIP's semantic features with FAST-VQA's texture and temporal distortion features. Quality predictions are derived by calculating cosine similarity between learnable prompts and fused features, achieving interpretable, state-of-the-art results.

• **ZE-FESG** [267] leverages CLIP to extract 52-dimensional feature vectors from video frames using 35 technical and 17 abstract descriptions as semantic guidance. Features are aggregated to produce quality scores, offering a zero-shot NR-VQA solution.

• **Q-Align** [255] fine-tunes large multimodal models (LMMs) to align predictions with human ratings by categorizing MOS into discrete levels (*e.g.*, "excellent," "good"). During inference for video quality assessment, the model processes sparsely sampled frames to output weighted scores.

• **LMM-VQA** [104] models temporal information using SlowFast [272] motion features processed into temporal tokens, and spatial features via ViT modules in CLIP. These tokens, combined with text tokens, are decoded into sequences indicating video quality scores or levels.

C. Loss Function

Next, we introduce loss functions that are commonly used as objectives when training deep video quality models.

1) L_1/L_2 ***loss:*** L_p $(p = 1, 2)$ norms have been used to learn most visual tasks due to their simplicity and analytical properties. The L_1 loss is also known as the mean absolute error (MAE) between a batch of predicted quality scores and MOS:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}|q_i - \hat{q}_i|, \tag{1}$$

where N is the batch size, and q_i and $\hat{q}_i$ are the predicted quality scores and the ground truth quality scores of the i-th video in the batch, respectively. Many deep learning based VQA models [4], [61], [66], [74], [77], [79], [94] employ the L_1 loss. Many other deep VQA models [57]–[60], [63]–[65], [69], [72], [73], [75], [78], [85], [92] use the L_2 loss, also known as the mean squared error (MSE) between the predicted and ground-truth quality scores:

$$L_2 = \frac{1}{N}\sum_{i=1}^{N}\|q_i - \hat{q}_i\|_2^2. \tag{2}$$

2) ***Monotonicity Loss:*** This is a type of ranking loss which distinguishes the relative qualities of videos. The rank value between an arbitrary pair of videos in a batch can be formulated as:

$$L_{rank}^{ij} = max(0, |q_i - q_j| - e(q_i, q_j)\cdot(\hat{q}_i - \hat{q}_j)), \tag{3}$$

where

$$e(q_i, q_j) = \begin{cases} 1 & q_i \geq q_j \\ -1 & q_i < q_j. \end{cases} \tag{4}$$

The monotonicity loss is then:

$$L_{monotonicity} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=1}^{N}L_{rank}^{ij}. \tag{5}$$

The authors of [61], [66] calculate the weighted average of monotonicity loss and L_1 loss to jointly optimize the monotonicity and accuracy of their video quality prediction models. They found that using monotonicity loss can accelerate model convergence.

*3) **Cross entropy loss**:* Some authors have cast the VQA problem in a classification setting [276], [277], where different quality levels correspond to different classes. Then the widely used cross entropy loss between the predicted qualities and the discrete set of labels can be used to optimize model efficiency. The cross entropy loss is:

$$L_{cross-entropy} = -\frac{1}{N}[\sum_{i=1}^{N}\sum_{j=1}^{K}\hat{p}(l^{(i)}=j)\log p(l^{(i)}=j)], \tag{6}$$

where K is the number of classes, and $\hat{p}$ and p are K-dimensional probability vectors, where each entry indicates the probability of a quality level of the ground-truth and prediction, respectively. For example, V-MEON [67] predicts the types of codecs used to compress the input videos, ContentNet in [79] outputs the content labels along with quality, and DistortionNet [79] and the Softmax classification module in [63] both predict the distortion types.

*4) **Cosine Similarity Loss**:* Vectorized regression [93] generates two probabilistic quality vectors corresponding to ground-truth $\boldsymbol{q}$ and prediction $\boldsymbol{y}$. In this case, one may deploy the cosine similarity loss between them:

$$L_{cosine} = 1 - \frac{\langle \boldsymbol{q}\cdot\boldsymbol{y}\rangle}{\|\boldsymbol{q}\|\cdot\|\boldsymbol{y}\|}, \tag{7}$$

where $\langle\cdot\rangle$ is the inner product, $\|\cdot\|$ is the L_2-norm.

*5) **SROCC/PLCC Loss**:* The Spearman Rank Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (PLCC) are commonly used as metrics to evaluate objective VQA model predictive performance against subjective labels, and hence it is natural to employ them as loss functions [67], [76], [95]. The differentiable SROCC [278] is expressed as:

$$SROCC = \frac{\sum_{i=1}^{N}(q_{pr}^i - \overline{q_{pr}})(q_r^i - \overline{q_r})}{\sqrt{\sum_{i=1}^{N}(q_{pr}^i - \overline{q_{pr}})^2\sum_{i=1}^{N}(q_r^i - \overline{q_r})^2}}, \tag{8}$$

where $\{q_{pr}^i\}_{i=1}^N$ and $\{q_r^i\}_{i=1}^N$ are the ranks of the predicted quality scores $\{q_p^i\}_{i=1}^N$ and ground-truth opinions $\{q^i\}_{i=1}^N$, respectively, and $\overline{q_{pr}}$ and $\overline{q_r}$ denote the mean values of $\{q_{pr}^i\}_{i=1}^N$ and $\{q_r^i\}_{i=1}^N$. The SROCC loss is then defined as:

$$L_{SROCC} = 1 - SROCC. \tag{9}$$

Similarly, the differentiable PLCC is calculated as:

$$PLCC = \frac{\sum_{i=1}^{N}(q_m^i - \overline{q_m})(q^i - \overline{q})}{\sqrt{\sum_{i=1}^{N}(q_m^i - \overline{q_m})^2\sum_{i=1}^{N}(q^i - \overline{q})^2}}, \tag{10}$$

where the $\{q_m^i\}_i^N$ are the fitted predictions from a 4-parameter nonlinear mapping on $\{q_p^i\}_{i=1}^N$, following the recommendation of the Video Quality Experts Group [279], and $\overline{q_m}$ is the mean value of $\{q_m^i\}_i^N$. Then the PLCC loss can be defined as:

$$L_{PLCC} = \frac{1 - PLCC}{2}. \tag{11}$$

*6) **Other Loss Functions**:* A variety of other loss functions have been used to design deep VQA models. A mixed dataset training strategy is adopted in [260] to better generalize VQA models, wherein monotonicity loss, PLCC loss and normalized L_1 loss are averaged to obtain a compound loss on each individual dataset, and where the overall loss used to train the unified VQA model across the multiple datasets is a softmax weighted average of the losses on all the datasets. NIMA [217] applies the squared Earth Mover's Distance (EMD) [280] to measure the distances between the cumulative distributions of the predicted quality ratings and the ground-truth human ratings. CNN-TLVQM [71] leverages Huber loss to train a CNN model to conduct local feature extraction, and to regress segment-level feature vectors into final video quality scores produced by an LSTM network. CoINVQ [79] uses pairwise loss and contrastive loss to evaluate compression level differences in CompressNet, while calculating the pairwise hinge loss in DistortionNet.

V. Performance Benchmark

In this section, we delve deeper into the comparison of both FR and NR IQA/VQA algorithms across several open-source video quality assessment datasets, each curated to challenge algorithms with diverse distortion types that are increasingly relevant in contemporary video streaming and broadcast scenarios.

A. Evaluation Criteria

Generally, an objective model's performance can be evaluated by examining its accuracy, monotonicity, and consistency with human ratings. There are three commonly employed performance metrics: SROCC, PLCC, and the root mean squared error (RMSE). Specifically, SROCC evaluates the monotonicity between the predicted and ground truth quality scores, PLCC evaluates the model's prediction linearity, and RMSE measures the prediction accuracy. Higher SROCC, PLCC, and lower RMSE scores denote better performances. Note that PLCC and RMSE are computed after performing a nonlinear four-parametric logistic regression to linearize the objective predictions to be on the same scale as MOS [130]:

$$f(x) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp\{(-x + \beta_3/|\beta_4|\})}. \tag{12}$$

B. Experimental Setup

For the FR models, we conducted the evaluation using the LIVE-VQA [105] and MCL-V [114] datasets known for their inclusion of conventional distortions such as compression artifacts, transmission, and scaling, and three additional datasets, LIVE-YT-HFR [117], BVI-HFR [116], and BVI-VFI [118], specifically chosen for their representation of temporal distortions, including variable frame rates and frame interpolation. For the NR models, the test beds include three UGC VQA datasets, LIVE-VQC [111], KoNViD-1k [119], and YouTube-UGC [5], and two cutting-edge AIGC VQA datasets, T2VQA-DB [145] and GAIA [146]. Details describing these datasets can be found in Section III.

TABLE V: Performance comparison of full reference IQA/VQA models on FR video quality assessment databases. The image and video quality assessment models are indicated by *italic* and orthographic font, respectively. Metrics are SROCC/PLCC, and the results of the best performing model are boldfaced.

Type	Model	General distortion		Bespoke for temporal distortion		
		LIVE-VQA [105]	MCL-V [114]	LIVE-YT-HFR [117]	BVI-HFR [116]	BVI-VFI [118]
Knowledge-driven (IQA/VQA)	*PSNR*	0.6958/0.7499	0.5410/0.5430	0.7802/0.7481	0.2552/0.3155	0.5200/0.4710
	SSIM [164]	0.7211/0.7883	0.7100/0.7090	0.5566/0.5418	0.1958/0.3532	0.6000/0.5400
	VIF [40]	0.6861/0.7601	0.7430/0.7470	0.6810/0.7020	0.2500/0.2640	0.5350/0.4890
	VMAF [171]	0.8163/0.8115	0.8280/0.8300	0.7782/0.7419	0.1888/0.3703	0.5950/0.5640
	ST-GREED [137]	0.6869/0.7049	0.7817/0.7996	**0.8822/0.8869**	**0.8042/0.8312**	0.1120/0.2140
Deep learning (IQA)	*DeepQA* [185]	0.8678/0.8692	0.6395/0.6384	0.0815/0.3162	0.0222/0.3061	0.4438 /0.4671
	LPIPS [80]	0.5243/0.5658	0.7610/0.7520	0.6920/0.7050	0.2388/0.3753	0.5990/0.5970
	DISTS [25]	0.4582/0.4868	0.7960/0.7820	0.7210/0.7290	0.4046/0.6789	0.5000/0.5500
	TOPIQ [178]	0.7095/0.7345	0.7309/0.7262	0.1216/0.3080	0.0016/0.2823	0.2342/0.3639
Temporal pooling / NN module (VQA)	FloLPIPS [82]	0.5485/0.5653	0.5447/0.5833	0.0716/0.1603	0.1041/0.2456	**0.6830/0.7060**
	DeepVQA [57]	0.9152/0.8952	-/-	0.4331/0.3996	0.1469/0.2013	-/-
	C3DVQA [58]	**0.9261/0.9122**	0.7850/0.7920	0.7300/0.7410	-/-	0.5080/0.3510
	STRA-VQA [194]	-/-	**0.8570/0.8640**	0.7990/0.8060	-/-	-/-

TABLE VI: Performance comparison of no-reference IQA/VQA models on NR video quality assessment databases. The image and video quality assessment models are indicated by *italic* and orthographic font, respectively. Metrics are SROCC/PLCC, and the results of the best performing model are boldfaced.

Type	Model	User-generated Content			AI-generated Content	
		LIVE-VQC [111]	KoNViD-1k [119]	YouTube-UGC [5]	T2VQA-DB [145]	GAIA [146]
Knowledge-driven (IQA/VQA)	*BRISQUE* [48]	0.5948/0.6267	0.6567/0.6576	0.3820/0.3952	0.1880/0.2554	0.0967/0.2120
	NIQE [47]	0.5896/0.6205	0.5417/0.5530	0.2379/0.2776	0.0047/0.2045	0.0615/0.1904
	TLVQM [200]	0.7964/0.7975	0.7729/0.7688	0.6693/0.6590	0.4891/0.4960	0.4655/0.4783
	VIDEVAL [204]	0.7148/0.7258	0.7832/0.7803	0.7787/0.7733	0.5246/0.5435	0.4684/0.4801
	FAVER [205]	0.7888/0.7982	0.7906/0.7912	0.7367/0.7395	0.5105/0.5351	0.2004/0.2691
Deep learning (IQA)	*PaQ-2-PiQ* [112]	0.6436/0.6683	0.6130/0.6014	0.2658/0.2935	0.1518/0.1409	0.2261/0.2348
	DB-CNN [219]	0.6391/0.7125	0.7187/0.7300	0.4793/0.5224	0.0129/0.0552	0.1727/0.1797
	MUSIQ [231]	0.6252/0.7101	0.7266/0.7468	0.5299/0.5622	0.0712/0.0698	0.1572/0.1609
	CLIP-IQA+ [100]	0.7276/0.7789	0.7813/0.7817	0.5374/0.5801	0.0759/0.1351	0.1538/0.1677
2D CNNs with simple score/feature averaging (VQA)	RAPIQUE [45]	0.7287/0.7594	0.8031/0.8175	0.7591/0.7684	0.3130/0.4510	0.2728/0.3246
	SIONR [69]	0.7361/0.7821	0.8109/0.8180	0.3621/0.3949	0.2434/0.2554	0.1263/0.1520
2D CNN with temporal aggregation networks (VQA)	VSFA [74]	0.6978/0.7426	0.7728/0.7754	0.7240/ 0.7430	0.1011/0.1193	0.5085/0.5215
	2BiVQA [85]	0.7610/0.8320	0.8150/0.8350	0.7710/0.7900	-/-	-/-
3D CNNs/ Transformation (VQA)	BVQA [76]	0.8340/0.8420	0.8340/0.8360	0.8180/0.8260	0.7390/0.7486	0.5201/0.5289
	Shen et al. [262]	0.7620/0.7660	0.7920/0.7880	0.7740/0.7660	0.2736/0.3147	0.0811/0.1438
Transformer (VQA)	FAST-VQA [95]	0.8211/0.8359	0.8543/0.8508	0.8617/0.8669	0.7173/0.7295	0.5276/0.5475
	DOVER [125]	0.7989/0.8348	0.8752/0.8816	0.8761/0.8753	0.7609/0.7693	**0.5335/0.5502**
	SAMA [264]	**0.8600/0.8780**	0.8920/0.8920	0.8810/0.8800	0.0136/0.0447	0.2361/0.2432
Large multimodality models (LMMs) (VQA)	COVER [98]	0.8093/0.8478	0.8933/0.8947	**0.9143/0.9165**	0.1276/0.2463	0.2254/0.2318
	MaxVQA [102]	0.8540/0.8730	**0.8940/0.8950**	0.8940/0.8900	0.1941/0.2232	0.2528/0.2509
	Q-align [255]	0.7730/0.8300	0.8650/0.8770	0.8340/0.8480	**0.7601/0.7768**	-/-
	LMM-VQA [104]	0.8310/0.8630	0.8750/0.8760	0.8580/0.8770	-/-	-/-

Representative models from each category of full-reference (FR) and no-reference (NR) video quality assessment algorithms, as previously introduced, are selected for performance comparison on datasets having diverse distortion types. The performance comparison results of FR models and NR models are summarized in Table V and Table VI, respectively. If available, performance data was taken from the original papers, otherwise we conducted the evaluation.

This analysis aims to provide insights into the accuracy and adaptability of these models, presenting a comprehensive overview of current capabilities and emerging trends in video quality assessment. Next, we separately analyzed the performance results of FR and NR models by type and provide insights on the effective employment of neural network modules in VQA models.

C. FR VQA Models Evaluation

As shown in Table V, among general-purpose knowledge-driven models, VMAF achieved the highest SROCC/PLCC scores on the LIVE-VQA and MCL-V databases, indicating

superior effectiveness in handling traditional quality degradations. However, these general-purpose models are less effective in scenarios with frame rate variability or frame interpolation distortions. ST-GREED, bespoke for perceiving framerate variation, yields high correlations with human ratings on two HFR databases, yet shows limitations on measuring quality degradations induced by video frame interpolation.

Deep learning-based IQA models perform reasonably well on general distortion datasets. DeepQA, in particular, shows a notable advantage in terms of SROCC/PLCC scores, suggesting that these models can capture complex spatial artifacts introduced by compression, transmission, and scaling distortions. However, they struggle when applied to HFR and VFI datasets due to limited temporal modeling capabilities.

By incorporating effective temporal pooling strategies or temporally aware NN modules, deep learning-based VQA models show improved performance over knowledge-driven and deep learning-based IQA models. For instance, FloLPIPS, a recently proposed bespoke metric for VFI that combines distortions in optical flow with LPIPS, delivers the highest SROCC/PLCC scores on the BVI-VFI database. C3DVQA performs notably well on the LIVE-VQA dataset due to its robust handling of spatiotemporal interactions, underscoring the importance of temporal awareness in VQA tasks. Moreover, STRA-VQA, which uses a Transformer-based architecture to model spatiotemporal quality, yields remarkable performance on MCL-V and LIVE-YT-HFR databases.

We offer the following design insights for deep learning-based FR VQA models:

- Spatial quality modeling. Reference videos enable precise spatial quality comparisons. Structural similarity metrics, such as SSIM in traditional models and DISTS in deep learning, effectively model spatial quality. Visual sensitivity maps in models like DeepQA also incorporate human perception, enhancing spatial distortion assessment.
- Temporal feature extraction. Incorporating optical flow (*e.g.*, FloLPIPS) or spatiotemporal 3D block features (*e.g.*, C3DVQA) improves performance on datasets with significant temporal distortions.
- Temporal effect modeling. Attention mechanisms, memory architectures, and memory-aware pooling strategies enhance temporal quality prediction. Transformer-based models like STRA-VQA excel at capturing long-range temporal dependencies, particularly in scenarios like variable frame rates and frame interpolation, by dynamically prioritizing critical temporal features.

D. NR VQA Models Evaluation

As shown in Table VI, knowledge-driven IQA models like BRISQUE and NIQE do not perform well in dynamic video environments. By incorporating deep features and end-to-end training, deep learning-based IQA models deliver better performance on UGC databases. However, knowledge-driven VQA models, such as TLVQM, VIDEVAL, and FAVER, can outperform both handcrafted and learning-based IQA models on both UGC and AIGC video datasets, highlighting the effectiveness of handcrafted spatiotemporal features. It can be seen that, in no reference scenarios, traditional IQA models, even deep learning-based, struggle when applied directly to video datasets, particularly those involving complex motions or AI-generated contents. Their reliance on frame-level features without considering temporal dependencies limits applicability in dynamic video environments. Nevertheless, CLIP-IQA+, outperforms handcrafted and learning-based IQA models on UGC databases, benefiting from vision-language prior in CLIP.

Models like RAPIQUE and SIONR incorporate simple temporal pooling strategies, allowing better performance in UGC datasets. VSFA and 2BiVQA use temporal aggregation networks to model temporal dependencies better than the handcrafted methods, improving performance in user-generated content (UGC). However, when tested on AI-generated content, their inability to model complex visual-text alignment or temporal coherence becomes evident. BVQA leverages 3D CNNs for better temporal feature extraction, performing significantly better on UGC datasets and starting to bridge the performance gap on AI-generated content.

Both Transformer-based and LMM-based models demonstrate the most excellent performance as well as a strong generalization to UGC content, and notably, show promise for AI-generated content. These models highlight the potential of Transformers and multimodal perception in VQA, but further optimization is also needed for broader deployment in AI-generated content, which often has lower frame rates and may require more refined temporal modeling.

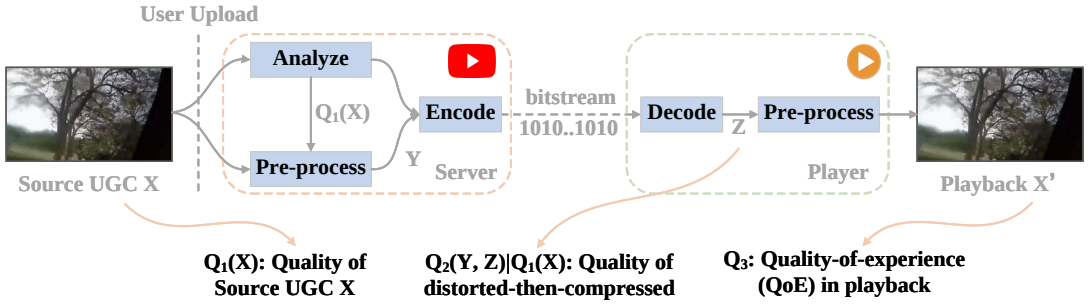
We supply some insights for model design in deep learning-based NR VQA models as follows.

- Temporal quality modeling. Capturing temporal dynamics is a key challenge. While 2D CNNs excel at spatial features, they lack temporal distortion modeling. The use of 3D CNNs and advanced Transformers has significantly improved temporal quality prediction by effectively capturing long-range dependencies.
- Multimodality Integration. Models like COVER and LMM-VQA leverage large multimodal approaches, combining text-video alignment and video fidelity assessment, and are particularly essential for UGC and AIGC evaluation. Such models are increasingly important given the rise of generative AI, highlighting how multimodal-based approaches can bridge the gap between human perception and objective quality models.

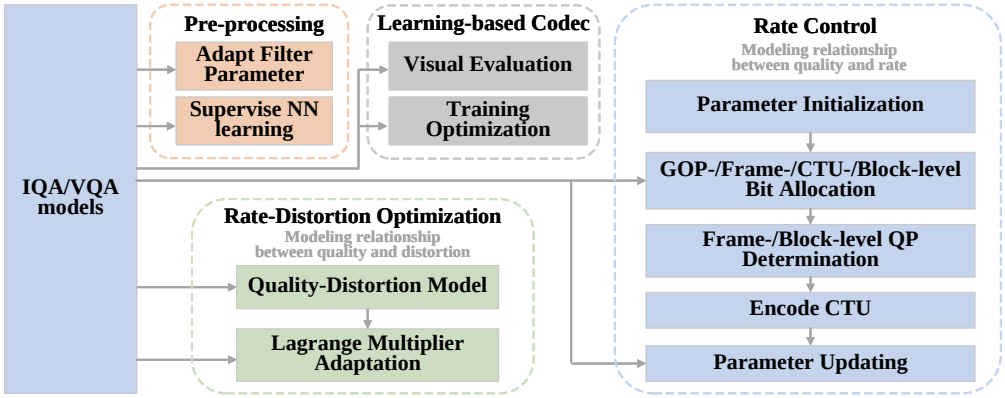
VI. Applications and Challenges

A. Server-Side UGC Video Coding and Transcoding

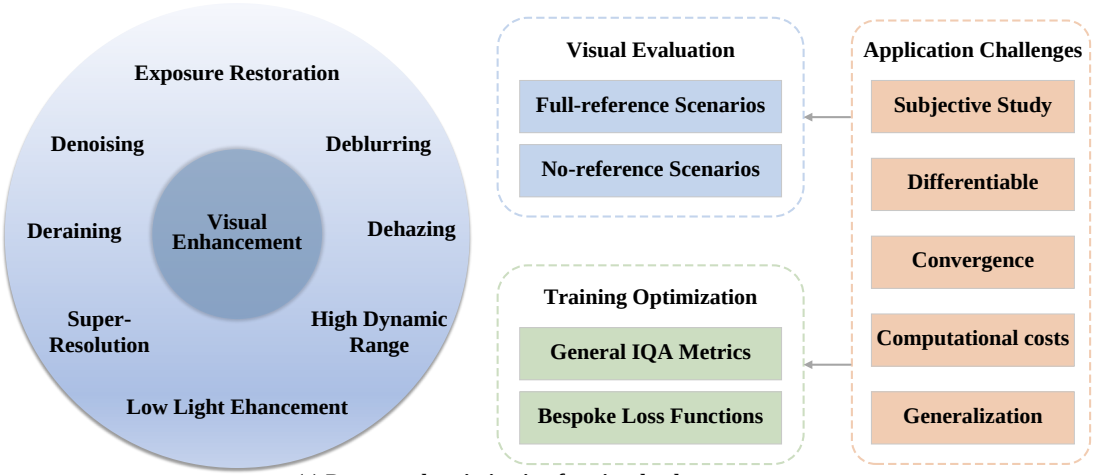
Streaming platforms like YouTube, TikTok, Facebook, and Bilibili use video transcoding to deliver high-quality, low-bitrate videos. Modern codecs like H.26x, VPx, AV1, and AVSx employ a hybrid framework (prediction, transformation, quantization, entropy coding) guided by rate-distortion optimization (RDO) to minimize distortion under bitrate constraints. However, defining distortion in this context can be challenging. Traditional pixel-based metrics like SAD, MSE, and PSNR poorly correlate with human perception. Advanced full-reference VQA models like SSIM and VMAF provide more accurate predictions of perceptual quality and are widely



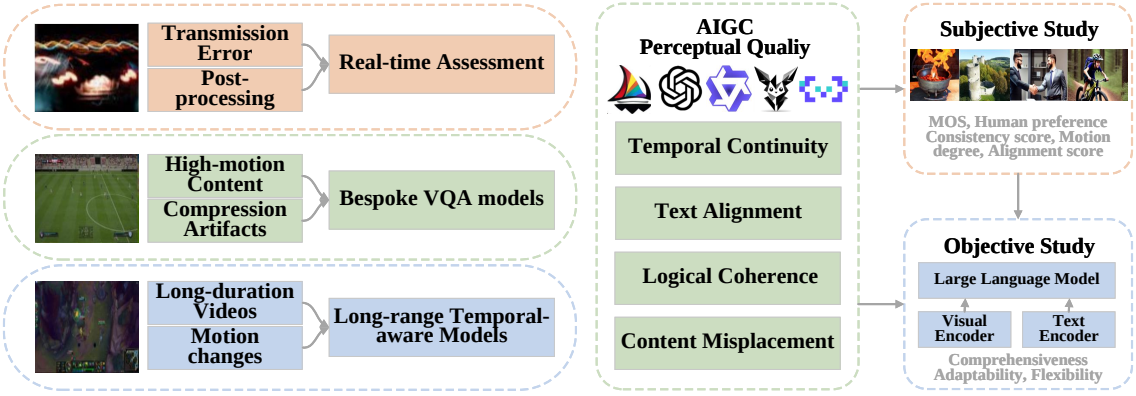
(a) Diagram of an exemplar server-side UGC video coding/transcoding framework.



(b) Perceptually optimized video coding in pre-processing, rate-control, and rate-distortion optimization.



(c) Perceptual optimization for visual enhancement.



(d) Live-streaming video quality monitoring. (e) AI-generated image/video quality assessment.

Fig. 13: Application overview of video quality assessment.

used for compressing professionally-generated content (PGC). However, user-generated content (UGC) often lacks high-quality reference videos and suffers from diverse distortion types. Compressing UGC using distorted originals as references can lead to inaccurate quality predictions.

Towards resolving this issue, consider the general task of predicting the quality of distorted-then-compressed videos [281]–[283]. If the existing distortion is also compression, then this may be viewed as a type of video transcoding problem (albeit, likely without decoding the original compression). Fig. 13 (a) illustrates the typical server-side processing flow of a UGC coding and communication service. In this pipeline, if we disregard other pre-processing and post-processing stages, a source UGC video X, previously distorted and/or compressed by unknown processes, can be quality-evaluated using a blind VQA model to generate a no-reference quality estimate, $Q_1(X)$. The overall quality of the subsequent distorted-then-compressed UGC video can then be modeled using a Bayesian approach [284], allowing for prediction of the final quality of the compressed UGC video, based on the estimated compressed video quality $Q_2(Y, Z)$ conditioned on the input quality $Q_1(X)$. Being able to more accurately predict and monitor the compressed quality of UGC encodes or transcodes can enable more efficient compression rate, extending the limits of traditional UGC coding units.

B. Perceptually Optimized Video Coding

Significant progress in perceptual image/video quality assessment models has advanced their integration into video codec design, optimization, and testing, enhancing coding efficiency. We refer the readers to [285] for systematic discussions on the rationale and methodology of utilizing perceptual VQA algorithms, like SSIM, in video coding. Below, we review key areas: preprocessing, rate control, rate-distortion optimization, and end-to-end codec evaluation and optimization, as shown in Fig. 13 (b).

Preprocessing input video frames before encoding can enhance the perceptual quality of the reconstructed/decoded videos. For instance, sharpening filters [286] counteract encoding-induced blur, with parameters tuned using VMAF. Deep preprocessors [287] employ perceptual loss functions like MS-SSIM and NIMA, while JND thresholds [288] guide adaptive Gaussian filtering for perceptual improvements.

Rate control in video encoding allocates bit budgets by constant bit-rate (CBR), average bit-rate (ABR), or variable bit-rate (VBR) control. A deep learning-based framework proposed in [289], [290] predicts rate factors (RF) for video segments, ensuring constant high quality measured by VMAF. Similarly, Netflix's Dynamic Optimizer [291] generates R-D optimal convex hulls on a per-shot basis, optimizing bit allocation while maintaining VMAF-based distortion metrics. Deng et al. [286] demonstrated that higher motion content allows higher distortion thresholds for the same VMAF score, enabling QP adjustments to balance quality and bitrate. SSIM-based perceptual rate control [292] models QP-perceptual quality relationships, optimizing QP settings per frame, while [293] employs SSIM in a rate-distortion model to meet bit budget targets.

The main objective of video compression is to achieve optimal trade-offs between rate and distortion. A commonly known strategy to achieve this balance is via rate-distortion optimization (RDO), which can be transformed into an unconstrained optimization problem using a Lagrange multiplier approach [294]. Thus find:

$$\min\{J\} \quad \text{where} \quad J = D + \lambda \cdot R, \tag{13}$$

where J, D, λ, and R denote rate-distortion cost, distortion, Lagrange multiplier, and rate, respectively. In previous work focusing on RDO based on perceptual distortion, the distortion component D has generally been replaced by VMAF [286] or SSIM [293], [295], [296].

Deep learning-based end-to-end image and video compression frameworks have garnered attention for efficiently encoding visual data with high perceptual quality. IQA/VQA metrics are essential for evaluating and optimizing these frameworks, employing full-reference metrics like PSNR, SSIM, MS-SSIM, and LPIPS, and no-reference metrics like NIQE and CLIP-IQA [34], [35], [297], [298]. These metrics also function as loss functions during training, guiding perceptually optimized compression [34], [35], [299]. For example, ProxIQA [299] approximates IQA metrics like VMAF and SSIM, serving as a perceptual loss layer. Integrating these metrics advances end-to-end compression, improving both coding efficiency and perceptual quality.

C. Perceptual Optimization for Visual Enhancement

Visual enhancement technologies aim to restore and improve degraded content, significantly enhancing perceptual quality in industries from entertainment to real-time applications. In mobile photography, for instance, advancements in hardware and computational algorithms address inherent limitations, such as small sensors and constrained optics. Key techniques, including denoising, HDR, super-resolution, color correction, and white balance, play a pivotal role in improving visual quality. Integrating human visual perception into these methods ensures that the results are not only computationally efficient but also visually appealing. IQA/VQA metrics serve as critical tools, functioning as both evaluation benchmarks and loss functions to refine visual enhancement algorithms effectively, as illustrated in Fig. 13 (c).

IQA/VQA metrics are critical for assessing how well enhancement algorithms improve perceptual quality. Full-reference metrics like PSNR and SSIM are widely used for tasks such as denoising [300], [301], deblurring [302], [303], dehazing [304], [305], and super-resolution [306], [307]. Advanced metrics like MS-SSIM and LPIPS are applied to HDR and multiple restoration tasks [308], [309]. In scenarios lacking reference data, such as autonomous driving under day-night transitions, no-reference metrics like NIQE, MUSIQ, and NIMA evaluate low-light enhancement and HDR [310], [311], highlighting the necessity of no-reference metrics in challenging scenarios.

IQA/VQA-based perceptual loss functions are pivotal in training deep learning models for visually natural results. For example, SSIM preserves structural details [312], NLPD

optimizes HDR tone mapping [313], [314], and PTQE finetune video frame interpolation models [315]. Custom loss functions like 1D-Wasserstein distances between CNN activations aid multi-task learning for tasks such as denoising and JPEG artifact removal [316]. Specialized loss functions, including those mimicking human eye response, improve exposure restoration [317], showcasing the versatility of IQA metrics in perceptual optimization.

Despite their utility, IQA/VQA metrics face several challenges: 1) Subjective gaps. Existing metrics may not align with perceptual quality when measuring restoration results [318], [319], requiring large-scale subjective studies to benchmark and calibrate metrics. 2) Optimization issues. Non-differentiability and non-smooth gradients complicate model training. 3) Computational Costs. Deep learning-based metrics are resource-intensive, limiting real-time applications. 4) Generalization. Metrics tailored for specific tasks, like super-resolution, often underperform for others, such as video enhancement, where temporal consistency is crucial. Developing adaptable metrics remains a significant challenge.

D. Live-Streaming Video Quality Monitoring

Live-streaming platforms like Twitch, TikTok, Facebook Live, and YouTube Live have surged in popularity, offering diverse content from gaming and sports to concerts and personal vlogs. Ensuring consistent perceptual video quality is critical but challenging [134] due to several inherent factors, as depicted in Fig. 13 (d).

Firstly, real-time transmission over User Datagram Protocol (UDP) can result in quality degradation, such as frame drops and flickering. Addressing these issues requires robust error recovery mechanisms and real-time VQA metrics to monitor user experiences. Secondly, unlike video-on-demand (VoD), live-streaming has limited time for pre- and post-processing. Videos must be encoded, transmitted, and decoded almost instantly, leaving little room for quality measurement and optimization. Additionally, high-motion live-streams, such as sports or eSports, are prone to distortions like stutter, motion blur, and deinterlacing motion mismatches under constrained bandwidth. Complex temporal variations further challenge traditional VQA models, which may struggle with such dynamic content. Given these challenges, VoD-optimized VQA techniques often fall short for live-streaming due to real-time and long-duration requirements. Effective VQA models for live-streaming must efficiently handle long-range temporal distortions and sudden motion changes.

E. AI-Generated Content Picture/Video Quality Assessment

The rapid advancements in machine learning (ML) have revolutionized AI-Generated Content (AIGC) [320], [321], encompassing images, videos, and other media created by models like DALLE-2, Imagen, Imagen-Video, and Stable Diffusion. These platforms enable efficient, customized, and diverse content production but also introduce unique challenges, as seen in Fig. 13 (e).

Unlike traditional content, AIGC faces novel distortions such as Uncanny Valley [322], [323], unrealistic object placement, and a lack of temporal coherence in videos, where frame sequences may lack logical continuity. These issues, absent in human-generated content, significantly impact perceptual quality and fall outside the scope of existing IQA/VQA algorithms, which requires new benchmarks and algorithms tailored to these unique challenges. Recent datasets, such as FETV [143], VBench [144], GAIA [146], and LGVQ [147], provide a foundation by focusing on spatiotemporal consistency, text alignment, and motion quality. As AIGC evolves, new distortions may emerge, necessitating flexible and adaptable IQA/VQA models. Addressing these challenges will enhance user acceptance and ensure high perceptual quality, fostering growth in this transformative domain, whereby robust IQA/VQA algorithms and benchmarks will be key to unlocking the full potential of AIGC.

VII. Conclusion and Future Directions

We have offered a comprehensive survey of deep learning-based video quality assessment studies, focusing on both subjective and objective quality assessment methods. The general workflows of subjective quality assessment data gathering and the existing panoply of popular databases were summarized. We then examined objective full-/no-reference algorithms from the past two decades, with a heavy focus on more recent deep learning-based VQA models. We also conducted a comprehensive comparison of the effectiveness and adaptability of existing FR and NR algorithms on databases of emerging content, providing insights on the employment of effective neural network modules in VQA models. We discussed current limitations and challenges in practical applications of deep learning VQA research, and elaborated on significant directions for future research.

Deep learning-based IQA/VQA models face several challenges in datasets, model architectures, and training strategies, with ample opportunities for future improvement. The limitations of existing video quality assessment datasets hinder the potential of data-driven learning. Constructing large-scale, unbiased datasets for VQA is labor-intensive, requiring accurate data collection methodologies and substantial human involvement in psychometric studies. Future datasets are required to match the ever-expanding demands of streaming videos, encompassing advanced formats like high framerate (HFR), high dynamic range (HDR), and virtual/extended reality (VR/XR). Future datasets should also represent varied content types, including screen content, UGC, PGC, AIGC, telepresence data, and point cloud data, ensuring relevance to emerging video technologies and consumer expectations.

In terms of model architectures, video quality assessment presents unique challenges distinct from high-level computer vision tasks. While pre-trained neural networks excel at extracting spatial features, bespoke architectures are needed to address low-level degradation in video content. Future architectures should effectively integrate psychovisual principles such as memory effects and perceptual straightening to align predictions with human visual experiences. Additionally, leveraging large models, including prompt-driven and feature-based methods, offers promising avenues. However, challenges such as aligning textual outputs with human-like quality judgments

and optimizing feature extraction for dynamic content must be addressed. Furthermore, designing models that balance prediction accuracy with computational efficiency remains a critical goal, as demonstrated by efficient frameworks like RAPIQUE, FAST-VQA, and Faster-VQA.

Training strategies also warrant further exploration. The scarcity of adequately annotated datasets remains a bottleneck, emphasizing the importance of patch-/clip- level training, leveraging proxy scores, and unsupervised learning. The selection of suitable loss functions is equally critical, with approaches like Huber loss or compound loss functions potentially mitigating the challenges posed by unbounded loss functions in challenging scenarios. These innovations can drive the development of robust, accurate VQA models capable of handling diverse content types and distortion phenomena.

By addressing existing challenges and exploring new avenues, future research can unlock the full potential of VQA in applications ranging from media production to telecommunications, ultimately enhancing the quality of visual experiences for end-users worldwide. We hope this survey stimulates further research within the visual analysis community, encouraging interdisciplinary collaborations to advance the state of video quality assessment.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[3] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[4] 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.

[5] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC dataset for video compression research," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–5.

[6] X. Min, G. Zhai, J. Zhou, M. C. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Transactions on Image Processing*, vol. 29, pp. 6054–6068, 2020.

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 34 892–34 916. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

[10] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13 040–13 051.

[11] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, K. Xu, C. Li, J. Hou, G. Zhai, G. Xue, W. Sun, Q. Yan, and W. Lin, "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 25 490–25 500.

[12] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun, Q. Yan, X. Min, G. Zhai, and W. Lin, "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," 2023. [Online]. Available: https://arxiv.org/abs/2312.17090

[13] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402.

[14] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385–2401, 2009.

[15] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2010.

[16] X. Zhang, X. Feng, W. Wang, and W. Xue, "Edge strength similarity for image quality assessment," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 319–322, 2013.

[17] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2013.

[18] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1500–1512, 2011.

[19] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[20] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[21] Z. Shi, K. Chen, K. Pang, J. Zhang, and Q. Cao, "A perceptual image quality index based on global and double-random window similarity," *Digital Signal Processing*, vol. 60, pp. 277–286, 2017.

[22] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal processing: Image communication*, vol. 19, no. 2, pp. 121–132, 2004.

[23] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2007, pp. I–869.

[24] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2480–2492, 2016.

[25] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567 – 2581, 2020.

[26] K. Ding, Y. Liu, X. Zou, S. Wang, and K. Ma, "Locally adaptive structure and texture similarity for image quality assessment," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2483–2491.

[27] W. Sun, T. Wang, X. Min, F. Yi, and G. Zhai, "Deep learning based full-reference and no-reference quality assessment models for compressed UGC videos," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2021, pp. 1–6.

[28] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end Optimized Image Compression," 2017. [Online]. Available: https://arxiv.org/abs/1611.01704

[29] M. Chen, T. Goodall, A. Patney, and A. C. Bovik, "Learning to compress videos without computing motion," *Signal Processing: Image Communication*, vol. 103, p. 116633, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596522000029

[30] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature neuroscience*, vol. 14, no. 9, pp. 1195–1201, 2011.

[31] 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.

[32] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath, "Design of linear equalizers optimized for the structural similarity index," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 857–872, 2008.

[33] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "Mcvd - masked conditional video diffusion for prediction, generation, and interpolation," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 23 371–23 385. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/944618542d80a63bbec16dfbd2bd689a-Paper-Conference.pdf

[34] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 11 913–11 924. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/8a50bae297807da9e97722a0b3fd8f27-Paper.pdf

[35] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 64 971–64 995. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/ccf6d8b4a1fe9d9c8192f00c713872ea-Paper-Conference.pdf

[36] Y. Lu, "The level weighted structural similarity loss: A step away from mse," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 9989–9990, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5131

[37] L. Galteri, L. Seidenari, M. Bertini, and A. Del Bimbo, "Deep generative adversarial compression artifact removal," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[38] H. Shi, L. Wang, N. Zheng, G. Hua, and W. Tang, "Loss functions for pose guided person image generation," *Pattern Recognition*, vol. 122, p. 108351, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321005318

[39] D. L. Ruderman, "The statistics of natural images," *Netw.: Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.

[40] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[41] R. Reininger and J. Gibson, "Distributions of the Two-Dimensional DCT Coefficients for Images," *IEEE Transactions on Communications*, vol. 31, no. 6, pp. 835–839, 1983.

[42] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[43] M. J. Wainwright and E. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., vol. 12. MIT Press, 1999. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1999/file/6a5dfac4be1502501489fc0f5a24b667-Paper.pdf

[44] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. Cock, "VMAF: The journey continues," *Netflix Technology Blog*, vol. 25, no. 1, 2018.

[45] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and accurate video quality prediction of user generated content," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 425–440, 2021.

[46] Q. Zheng, Z. Tu, Y. Fan, X. Zeng, and A. C. Bovik, "No-reference quality assessment of variable frame-rate videos using temporal bandpass statistics," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1795–1799.

[47] 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, 2013.

[48] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[49] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, 2014.

[50] Q. Zheng, Z. Tu, X. Zeng, A. C. Bovik, and Y. Fan, "A completely blind video quality evaluator," *IEEE Signal Processing Letters*, vol. 29, pp. 2228–2232, 2022.

[51] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2957–2971, 2017.

[52] D. Ghadiyaram, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17(1), no. 32, pp. 1–25, 2017.

[53] Q. Zheng, Z. Tu, Z. Hao, X. Zeng, A. C. Bovik, and Y. Fan, "Blind video quality assessment via space-time slice statistics," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 451–455.

[54] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, 2014.

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

[56] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009, vol. 2.

[57] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 219–234.

[58] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3dvqa: Full-reference video quality assessment with 3D convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4447–4451.

[59] C. Feng, F. Zhang, and D. R. Bull, "Deep VQA based on a novel hybrid training methodology," *arXiv preprint arXiv:2202.08595*, 2022.

[60] R. Hou, Y. Zhao, Y. Hu, and H. Liu, "No-reference video quality evaluation by a deep transfer CNN architecture," *Signal Processing: Image Communication*, vol. 83, p. 115782, 2020.

[61] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for UGC videos," *arXiv preprint arXiv:2204.14047*, 2022.

[62] J. You and J. Korhonen, "Deep neural networks for no-reference video quality assessment," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2349–2353.

[63] Y. Li, L.-M. Po, C.-H. Cheung, X. Xu, L. Feng, F. Yuan, and K.-W. Cheung, "No-reference video quality assessment with 3d shearlet transform and convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1044–1057, 2015.

[64] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2244–2255, 2018.

[65] J. Xu, J. Li, X. Zhou, W. Zhou, B. Wang, and Z. Chen, "Perceptual quality assessment of internet videos," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1248–1257.

[66] S. Wen and J. Wang, "A strong baseline for image and video quality assessment," *arXiv preprint arXiv:2111.07104*, 2021.

[67] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," *ACM Multimedia*, pp. 546–554, 2018.

[68] D. Varga, "No-reference video quality assessment based on the temporal pooling of deep features," *Neural Processing Letters*, vol. 50, no. 3, pp. 2595–2608, 2019.

[69] W. Wu, Q. Li, Z. Chen, and S. Liu, "Semantic information oriented no-reference video quality assessment," *IEEE Signal Processing Letters*, vol. 28, pp. 204–208, 2021.

[70] D. Varga, "No-reference video quality assessment using multi-pooled, saliency weighted deep features and decision fusion," *Sensors*, vol. 22, no. 6, p. 2209, 2022.

[71] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 3311–3319.

[72] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "No-reference video quality assessment using multi-level spatially pooled features," *arXiv preprint arXiv:1912.07966*, 2019.

[73] D. Varga and T. Szirányi, "No-reference video quality assessment via pretrained cnn and lstm networks," *Signal, Image and Video Processing*, vol. 13, no. 8, pp. 1569–1576, 2019.

[74] 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.

[75] 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.

[76] 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.

[77] J. Li and X. Li, "Study on no-reference video quality assessment method incorporating dual deep learning networks," *Multimedia Tools and Applications*, pp. 1–20, 2022.

[78] J. Chen, H. Wang, M. Xu, G. Li, and S. Liu, "Deep neural networks for end-to-end spatiotemporal video quality prediction and aggregation," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[79] Y. Wang, J. Ke, H. Talebi, J. G. Yim, N. Birkbeck, B. Adsumilli, P. Milanfar, and F. Yang, "Rich features for perceptual quality assessment of UGC videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 435–13 444.

[80] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[81] T. Tariq, O. T. Tursun, M. Kim, and P. Didyk, "Why are deep representations good perceptual quality features?" in *European Conference on Computer Vision*. Springer, 2020, pp. 445–461.

[82] D. Danier, F. Zhang, and D. Bull, "FlOLPIPS: A bespoke video quality metric for frame interpoation," *arXiv preprint arXiv:2207.08119*, 2022.

[83] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[84] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 1597–1600.

[85] A. Telili, S. A. Fezza, W. Hamidouche, and H. F. Meftah, "2BiVQA: Double BI-LSTM based Video Quality Assessment of UGC Videos," *arXiv preprint arXiv:2208.14774*, 2022.

[86] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[87] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[88] F. Yi, M. Chen, W. Sun, X. Min, Y. Tian, and G. Zhai, "Attention based network for no-reference ugc video quality assessment," in *IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1414–1418.

[89] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1903–1916, 2022.

[90] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[91] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[92] J. You, "Long short-term convolutional transformer for no-reference video quality assessment," in *ACM International Conference on Multimedia*, 2021, pp. 2112–2120.

[93] F. Xing, Y.-G. Wang, H. Wang, L. Li, and G. Zhu, "StarVQA: Space-time attention for video quality assessment," *arXiv preprint arXiv:2108.09635*, 2021.

[94] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, and W. Lin, "DisCoVQA: Temporal Distortion-Content Transformers for Video Quality Assessment," *arXiv preprint arXiv:2206.09853*, 2022.

[95] 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," *arXiv preprint arXiv:2207.02595*, 2022.

[96] Y. Li, M. Chen, W. Yang, K. Wang, J. Ma, A. C. Bovik, and Y. Zhang, "SAMscore: A semantic structural similarity metric for image translation evaluation," *arXiv preprint arXiv:2305.15367*, 2023.

[97] X. Li, T. Jiang, H. Fan, and S. Liu, "Sam-iqa: Can segment anything boost image quality assessment?" *arXiv preprint arXiv:2307.04455*, 2023.

[98] C. He, Q. Zheng, R. Zhu, X. Zeng, Y. Fan, and Z. Tu, "Cover: A comprehensive video quality evaluator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 5799–5809.

[99] K. Yuan, H. Liu, M. Li, M. Sun, M. Sun, J. Gong, J. Hao, C. Zhou, and Y. Tang, "Ptm-vqa: Efficient video quality assessment leveraging diverse pretrained models from the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 2835–2845.

[100] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 2555–2563, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25353

[101] H. Wu, L. Liao, J. Hou, C. Chen, E. Zhang, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring opinion-unaware video quality assessment with semantic affinity criterion," 2023. [Online]. Available: https://arxiv.org/abs/2302.13269

[102] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Towards explainable in-the-wild video quality assessment: a database and a language-prompted approach," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1045–1054.

[103] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai *et al.*, "Q-bench: A benchmark for general-purpose foundation models on low-level vision," *arXiv preprint arXiv:2309.14181*, 2023.

[104] Q. Ge, W. Sun, Y. Zhang, Y. Li, Z. Ji, F. Sun, S. Jui, X. Min, and G. Zhai, "Lmm-vqa: Advancing video quality assessment with large multimodal models," *arXiv preprint arXiv:2408.14008*, 2024.

[105] 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.

[106] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics on Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[107] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108 069–108 082, 2021.

[108] L. K. Choi and A. C. Bovik, "Video quality assessment accounting for temporal visual masking of local flicker," *Signal Processing: image communication*, vol. 67, pp. 182–198, 2018.

[109] D. Y. Lee, S. Paul, C. G. Bampis, H. Ko, J. Kim, S. Y. Jeong, B. Homan, and A. C. Bovik, "A subjective and objective study of space-time subsampled video quality," *IEEE Transactions on Image Processing*, vol. 31, pp. 934–948, 2021.

[110] J. P. Ebenezer, Z. Shang, Y. Chen, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "HDR or SDR? a subjective and objective study of scaled and compressed videos," *IEEE Transactions on Image Processing*, 2024.

[111] 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, feb 2019. [Online]. Available: https://doi.org/10.1109%2Ftip.2018.2869673

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

[113] Z. Ying, D. Ghadiyaram, and A. Bovik, "Telepresence video quality assessment," *European Conference on Computer Vision, Tel Aviv*, pp. 327–347, Oct 2022.

[114] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *Journal of Visual Communication and Image Representation*, vol. 30, pp. 1–9, 2015.

[115] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014—a database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, 2016.

[116] A. Mackin, F. Zhang, and D. R. Bull, "A study of high frame rate video formats," *IEEE Trans. Multimedia.*, vol. 21, no. 6, pp. 1499–1512, 2018.

[117] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108 069–108 082, 2021.

[118] D. Danier, F. Zhang, and D. R. Bull, "Bvi-vfi: A video quality database for video frame interpolation," *IEEE Transactions on Image Processing*, vol. 32, pp. 6004–6019, 2023.

[119] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k),"

in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[120] J. Xu, J. Li, X. Zhou, W. Zhou, B. Wang, and Z. Chen, "Perceptual quality assessment of internet videos," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1248–1257.

[121] G. Li, B. Chen, L. Zhu, Q. He, H. Fan, and S. Wang, "PUGCQ: A Large Scale Dataset for Quality Assessment of Professional User-Generated Content," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3728–3736.

[122] Y. Wang, J. Ke, H. Talebi, J. G. Yim, N. Birkbeck, B. Adsumilli, P. Milanfar, and F. Yang, "Rich features for perceptual quality assessment of UGC videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 435–13 444.

[123] X. Yu, Z. Ying, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective analysis of streamed gaming videos," *IEEE Transactions on Games*, vol. 16, no. 2, pp. 445 – 458, 2022.

[124] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1045–1054. [Online]. Available: https://doi.org/10.1145/3581783.3611737

[125] ——, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 20 144–20 154.

[126] Z. Zhang, W. Wu, W. Sun, D. Tu, W. Lu, X. Min, Y. Chen, and G. Zhai, "Md-vqa: Multi-dimensional quality assessment for ugc live videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 1746–1755.

[127] Y. Lu, X. Li, Y. Pei, K. Yuan, Q. Xie, Y. Qu, M. Sun, C. Zhou, and Z. Chen, "Kvq: Kwai video quality assessment for short-form videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 25 963–25 973.

[128] B. Series, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, vol. 500, no. 13, 2012.

[129] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Data compression conference (DCC)*, 2017, pp. 52–61.

[130] 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.

[131] D. Y. Lee, S. Paul, C. G. Bampis, H. Ko, J. Kim, S. Y. Jeong, B. Homan, and A. C. Bovik, "A subjective and objective study of space-time subsampled video quality," *IEEE Transactions on Image Processing*, vol. 31, pp. 934–948, 2022.

[132] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Visual Communications and Image Processing 2003*, T. Ebrahimi and T. Sikora, Eds., vol. 5150, International Society for Optics and Photonics. SPIE, 2003, pp. 573 – 582. [Online]. Available: https://doi.org/10.1117/12.509908

[133] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," in *Advanced Image and Video Communications and Storage Technologies*, N. Ohta, H. U. Lemke, and J. C. Lehureau, Eds., vol. 2451, International Society for Optics and Photonics. SPIE, 1995, pp. 90 – 101. [Online]. Available: https://doi.org/10.1117/12.201231

[134] Z. Shang, J. P. Ebenezer, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "Study of the subjective and objective quality of high motion live streaming videos," *IEEE Transactions on Image Processing*, vol. 31, pp. 1027–1041, 2021.

[135] Z. Shang, J. P. Ebenezer, A. K. Venkataramanan, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "A study of subjective and objective quality assessment of hdr videos," *IEEE Transactions on Image Processing*, vol. 33, pp. 42–57, 2024.

[136] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "Chipqa: No-reference video quality prediction via space-time chips," *IEEE Transactions on Image Processing*, vol. 30, pp. 8059–8074, 2021.

[137] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction," *IEEE Trans. on Image Processing*, vol. 30, pp. 7446–7457, 2021.

[138] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "Making video quality assessment models robust to bit depth," *IEEE Signal Processing Letters*, vol. 30, pp. 488–492, 2023.

[139] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015.

[140] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[141] I. Chivileva, P. Lynch, T. E. Ward, and A. F. Smeaton, "Measuring the quality of text-to-video model outputs: Metrics and dataset," 2023. [Online]. Available: https://arxiv.org/abs/2309.08009

[142] Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan, "Evalcrafter: Benchmarking and evaluating large video generation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 22 139–22 149.

[143] Y. Liu, L. Li, S. Ren, R. Gao, S. Li, S. Chen, X. Sun, and L. Hou, "Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 62 352–62 387. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/c481049f7410f38e788f67c171c64ad5-Paper-Datasets_and_Benchmarks.pdf

[144] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu, "Vbench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 21 807–21 818.

[145] T. Kou, X. Liu, Z. Zhang, C. Li, H. Wu, X. Min, G. Zhai, and N. Liu, "Subjective-aligned dataset and metric for text-to-video quality assessment," 2024. [Online]. Available: https://arxiv.org/abs/2403.11956

[146] Z. Chen, W. Sun, Y. Tian, J. Jia, Z. Zhang, J. Wang, R. Huang, X. Min, G. Zhai, and W. Zhang, "Gaia: Rethinking action quality assessment for ai-generated videos," 2024. [Online]. Available: https://arxiv.org/abs/2406.06087

[147] Z. Zhang, X. Li, W. Sun, J. Jia, X. Min, Z. Zhang, C. Li, Z. Chen, P. Wang, Z. Ji, F. Sun, S. Jui, and G. Zhai, "Benchmarking aigc video quality assessment: A dataset and unified model," 2024. [Online]. Available: https://arxiv.org/abs/2407.21408

[148] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: the new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, p. 64–73, jan 2016. [Online]. Available: https://doi.org/10.1145/2812802

[149] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017. [Online]. Available: http://arxiv.org/abs/1705.06950

[150] Taobao Alibaba, Inc., "TaoLive." [Online]. Available: https://taolive.taobao.com

[151] Kuaishou Technology, Inc., "Kwai." [Online]. Available: https://www.kwai.com/foryou

[152] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 15 903–15 935. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf

[153] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 36 652–36 663. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/73aacd8b3b05b4b503d58310b523553c-Paper-Conference.pdf

[154] Z. Zhang, C. Li, W. Sun, X. Liu, X. Min, and G. Zhai, "A perceptual quality assessment exploration for aigc images," in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2023, pp. 440–445.

[155] J. Wang, H. Duan, J. Liu, S. Chen, X. Min, and G. Zhai, "Aigciqa2023: A large-scale image quality assessment database for ai generated images: From the perspectives of quality, authenticity and correspondence," in *Artificial Intelligence*, L. Fang, J. Pei, G. Zhai, and R. Wang, Eds. Singapore: Springer Nature Singapore, 2024, pp. 46–57.

[156] Z. Chen, W. Sun, H. Wu, Z. Zhang, J. Jia, Z. Ji, F. Sun, S. Jui, X. Min, G. Zhai, and W. Zhang, "Exploring the naturalness of ai-generated images," 2024. [Online]. Available: https://arxiv.org/abs/2312.05476

[157] C. Li, Z. Zhang, H. Wu, W. Sun, X. Min, X. Liu, G. Zhai, and W. Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6833–6846, 2024.

[158] C. Li, T. Kou, Y. Gao, Y. Cao, W. Sun, Z. Zhang, Y. Zhou, Z. Zhang, H. Wu, W. Zhang, X. Liu, X. Min, and Z. Guangtao, "AIGIQA-20K: A Large Database for AI-Generated Image Quality Assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[159] C. Li, X. Wu, H. Wu, D. Feng, Z. Zhang, G. Lu, X. Min, X. Liu, G. Zhai, and W. Lin, "Cmc-bench: Towards a new paradigm of visual signal compression," 2024. [Online]. Available: https://arxiv.org/abs/2406.09356

[160] OpenAI, "GPT-4 technical report," https://openai.com/research/gpt-4, 2023.

[161] F. Zhang, A. Mackin, and D. R. Bull, "A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 300–304.

[162] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," in *IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 141–145.

[163] "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[164] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[165] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *JOSA A*, vol. 24, no. 12, pp. B61–B69, 2007.

[166] A. K. Moorthy and A. C. Bovik, "Efficient video quality assessment along temporal trajectories," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1653–1658, 2010.

[167] K. Zeng and Z. Wang, "3d-ssim for video quality assessment," in *2012 19th IEEE International Conference on Image Processing*, 2012, pp. 621–624.

[168] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[169] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.

[170] R. Soundararajan and A. C. Bovik, "Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, 2013.

[171] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[172] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Making Video Quality Assessment Models Sensitive to Frame Rate Distortions," *IEEE Signal Processing Letters*, vol. 29, pp. 897–901, 2022.

[173] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, no. 2, 2016.

[174] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[175] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2256–2270, 2018.

[176] A. K. Venkataramanan, C. Stejerean, and A. C. Bovik, "Funque: Fusion of unified quality evaluators," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 2147–2151.

[177] A. K. Venkataramanan, C. Stejerean, I. Katsavounidis, and A. C. Bovik, "One transform to compute them all: Efficient fusion-based full-reference video quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 509–524, 2024.

[178] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, and W. Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 2404–2418, 2024.

[179] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2009.

[180] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *IEEE International Conference on Image Processing*, 2011, pp. 2505–2508.

[181] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 200–213, 2013.

[182] M. H. Pinson, L. K. Choi, and A. C. Bovik, "Temporal video quality model accounting for variable frame delay distortions," *IEEE Transactions on Broadcasting*, vol. 60, no. 4, pp. 637–649, 2014.

[183] F. Zhang and D. R. Bull, "A perception-based hybrid model for video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017–1028, 2015.

[184] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[185] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1676–1684.

[186] S. Ahn, Y. Choi, and K. Yoon, "Deep learning-based distortion sensitivity prediction for full-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 344–353.

[187] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[188] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and less than 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[189] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997*, 2014.

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

[191] M. Kettunen, E. Härkönen, and J. Lehtinen, "E-LPIPS: robust perceptual image similarity via random transformation ensembles," *arXiv preprint arXiv:1906.03973*, 2019.

[192] X. Liao, B. Chen, H. Zhu, S. Wang, M. Zhou, and S. Kwong, "Deepwsd: Projecting degradations in perceptual space to wasserstein distance in deep feature space," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 970–978.

[193] 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.

[194] H. Zhu, B. Chen, L. Zhu, P. Chen, L. Song, and S. Wang, "Video quality assessment for spatio-temporal resolution adaptive coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6403–6415, 2024.

[195] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *2009 International Workshop on Quality of Multimedia Experience*, 2009, pp. 87–91.

[196] R. Hassen, Z. Wang, and M. M. A. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2798–2810, 2013.

[197] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1, 2002, pp. I–I.

[198] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3, 2000, pp. 981–984.

[199] S. A. Golestaneh and D. M. Chandler, "No-reference quality assessment of jpeg images via a quality relevance map," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 155–158, 2014.

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

[201] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

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

[203] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, 2015.

[204] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.

[205] Q. Zheng, Z. Tu, P. C. Madhusudana, X. Zeng, A. C. Bovik, and Y. Fan, "FAVER: Blind Quality Prediction of Variable Frame Rate Videos," *Signal Processing: Image Communication*, vol. 122, 2024.

[206] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, 2015.

[207] D. Ghadiyaram, C. Chen, S. Inguva, and A. Kokaram, "A no-reference video quality predictor for compression and scaling artifacts," in *IEEE International Conference on Image Processing*, 2017, pp. 3445–3449.

[208] P. Kancharla and S. S. Channappayya, "Completely blind quality assessment of user generated video content," *IEEE Trans. Image Process.*, vol. 31, pp. 263–274, 2021.

[209] Y. Liu, K. Gu, X. Li, and Y. Zhang, "Blind image quality assessment by natural scene statistics and perceptual characteristics," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 3, pp. 1–91, 2020.

[210] Y. Liu, K. Gu, Y. Zhang, X. Li, G. Zhai, D. Zhao, and W. Gao, "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 929–943, 2019.

[211] O. J. Hénaff, R. L. Goris, and E. P. Simoncelli, "Perceptual straightening of natural videos," *Nature Neuroscience*, vol. 22, no. 6, pp. 984–991, 2019.

[212] L. Liao, K. Xu, H. Wu, C. Chen, W. Sun, Q. Yan, and W. Lin, "Exploring the effectiveness of video perceptual representation in blind video quality assessment," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 837–846. [Online]. Available: https://doi.org/10.1145/3503161.3547849

[213] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2018.

[214] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[215] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3773–3777.

[216] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[217] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[218] H. Zeng, L. Zhang, and A. C. Bovik, "A probabilistic quality representation approach to deep blind image quality prediction," *arXiv preprint arXiv:1708.08190*, 2017.

[219] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2018.

[220] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[221] W. Sun, X. Min, D. Tu, S. Ma, and G. Zhai, "Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 6, pp. 1178–1192, 2023.

[222] 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.

[223] Y. Gao, X. Min, Y. Zhu, J. Li, X.-P. Zhang, and G. Zhai, "Image quality assessment: From mean opinion score to opinion score distribution," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 997–1005. [Online]. Available: https://doi.org/10.1145/3503161.3547872

[224] K.-Y. Lin and G. Wang, "Hallucinated-iqa: No-reference image quality assessment via adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[225] B. Chen, L. Zhu, C. Kong, H. Zhu, S. Wang, and Z. Li, "No-reference image quality assessment by hallucinating pristine features," *IEEE Transactions on Image Processing*, vol. 31, pp. 6139–6151, 2022.

[226] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Blind image quality assessment based on geometric order learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 12 799–12 808.

[227] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics on Signal Processing*, vol. 11, no. 1, pp. 206–220, 2016.

[228] 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.

[229] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 459–479.

[230] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 124–12 134.

[231] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157.

[232] J. You and J. Korhonen, "Transformer for image quality assessment," in *IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1389–1393.

[233] G. Qin, R. Hu, Y. Liu, X. Zheng, H. Liu, X. Li, and Y. Zhang, "Data-efficient image quality assessment with attention-panel decoder," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 2091–2100, Jun. 2023.

[234] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 1220–1230.

[235] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 1191–1200.

[236] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, "Continual learning for blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[237] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 143–14 152.

[238] S. Su, Q. Yan, Y. Zhu, J. Sun, and Y. Zhang, "From distortion manifold to perceptual quality: a data efficient blind image quality assessment approach," *Pattern Recognition*, vol. 133, p. 109047, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320322005271

[239] R. Ma, Q. Wu, K. N. Ngan, H. Li, F. Meng, and L. Xu, "Forgetting to remember: A scalable incremental learning framework for cross-task blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 25, pp. 8817–8827, 2023.

[240] H. Li, L. Liao, C. Chen, X. Fan, W. Zuo, and W. Lin, "Continual learning of blind image quality assessment with channel modulation kernel," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[241] Z. Wang, Q. Jiang, S. Zhao, W. Feng, and W. Lin, "Deep blind image quality assessment powered by online hard example mining," *IEEE Transactions on Multimedia*, vol. 25, pp. 4774–4784, 2023.

[242] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Task-specific normalization for continual learning of blind image quality models," *IEEE Transactions on Image Processing*, vol. 33, pp. 1898–1910, 2024.

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

[244] A. Shukla, A. Upadhyay, S. Bhugra, and M. Sharma, "Opinion unaware image quality assessment via adversarial convolutional variational autoencoder," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 2153–2163.

[245] A. Saha, S. Mishra, and A. C. Bovik, "Re-iqa: Unsupervised learning for image quality assessment in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 5846–5855.

[246] N. C. Babu, V. Kannan, and R. Soundararajan, "No reference opinion unaware quality assessment of authentically distorted images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 2459–2468.

[247] L. Agnolucci, L. Galteri, M. Bertini, and A. Del Bimbo, "Arniqa: Learning distortion manifold for image quality assessment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 189–198.

[248] K. Zhao, K. Yuan, M. Sun, M. Li, and X. Wen, "Quality-aware pre-trained models for blind image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22 302–22 313.

[249] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "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 (CVPR)*, June 2023, pp. 14 071–14 081.

[250] J. Hou, W. Lin, Y. Fang, H. Wu, C. Chen, L. Liao, and W. Liu, "Towards transparent deep image aesthetics assessment with tag-based content descriptors," *IEEE Transactions on Image Processing*, pp. 1–1, 2023.

[251] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[252] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023. [Online]. Available: https://arxiv.org/abs/2304.10592

[253] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2305.06500

[254] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2305.03726

[255] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun *et al.*, "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," *arXiv preprint arXiv:2312.17090*, 2023.

[256] H. Wu, H. Zhu, Z. Zhang, E. Zhang, C. Chen, L. Liao, C. Li, A. Wang, W. Sun, Q. Yan *et al.*, "Towards open-ended visual quality comparison," *arXiv preprint arXiv:2402.16641*, 2024.

[257] H. B. Martinez, M. C. Farias, and A. Hines, "A no-reference autoencoder video quality metric," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1755–1759.

[258] A. De Decker, J. De Cock, P. Lambert, and G. Van Wallendael, "No-reference vmaf: A deep neural network-based approach to blind video quality assessment," *IEEE Transactions on Broadcasting*, pp. 1–0, 2024.

[259] J. Yan, L. Wu, Y. Fang, X. Liu, X. Xia, and W. Liu, "Video quality assessment for online processing: From spatial to temporal sampling," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

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

[261] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, "Unsupervised curriculum domain adaptation for no-reference video quality assessment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 5178–5187.

[262] W. Shen, M. Zhou, X. Wei, H. Wang, B. Fang, C. Ji, X. Zhuang, J. Wang, J. Luo, H. Pu, X. Huang, S. Wang, H. Cao, Y. Feng, T. Xiang, and Z. Shang, "A blind video quality assessment method via spatiotemporal pyramid attention," *IEEE Transactions on Broadcasting*, vol. 70, no. 1, pp. 251–264, 2024.

[263] 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.

[264] Y. Liu, Y. Quan, G. Xiao, A. Li, and J. Wu, "Scaling and masking: A new paradigm of data sampling for image and video quality assessment," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, pp. 3792–3801, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28170

[265] S. Mitra and R. Soundararajan, "Knowledge guided semi-supervised learning for quality assessment of user generated videos," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, pp. 4251–4260, Mar. 2024.

[266] W. Wen, M. Li, Y. Zhang, Y. Liao, J. Li, L. Zhang, and K. Ma, "Modular blind video quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 2763–2772.

[267] Y. Mi, Y. Li, Y. Shu, and S. Liu, "Ze-fesg: A zero-shot feature extraction method based on semantic guidance for no-reference video quality assessment," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 3640–3644.

[268] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[269] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-Resnet and the impact of residual connections on learning," in *Thirty-first AAAI Conference on Artificial Intelligence*, 2017.

[270] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[271] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[272] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[273] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[274] 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 (CVPR)*, June 2022, pp. 3202–3211.

[275] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 966–11 976.

[276] H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 609–613.

[277] Z. Tu, C.-J. Chen, L.-H. Chen, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Regression or classification? new methods to evaluate no-reference picture and video quality models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 2085–2089.

[278] M. Blondel, O. Teboul, Q. Berthet, and J. Djolonga, "Fast differentiable sorting and ranking," in *International Conference on Machine Learning*, 2020, pp. 950–959.

[279] V. Q. E. Group *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment," in *VQEG meeting, Ottawa, Canada, March, 2000*, 2000.

[280] L. Hou, C.-P. Yu, and D. Samaras, "Squared earth mover's distance-based loss for training deep neural networks," *arXiv preprint arXiv:1611.05916*, 2016.

[281] X. Yu, N. Birkbeck, Y. Wang, C. G. Bampis, B. Adsumilli, and A. C. Bovik, "Predicting the quality of compressed videos with pre-existing distortions," *IEEE Transactions on Image Processing*, vol. 30, pp. 7511–7526, 2021.

[282] M. Smirnov, A. Gushchin, A. Antsiferova, D. Vatolin, R. Timofte, Z. Jia, Z. Zhang, W. Sun, J. Qian, Y. Cao *et al.*, "Aim 2024 challenge on compressed video quality assessment: Methods and results," *arXiv preprint arXiv:2408.11982*, 2024.

[283] M. V. Conde, S. Zadtootaghaj, N. Barman, R. Timofte, C. He, Q. Zheng, R. Zhu, Z. Tu, H. Wang, X. Chen *et al.*, "Ais 2024 challenge on video quality assessment of user-generated content: Methods and

results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5826–5837.
[284] A. C.Bovik, "Assessing quality of images or videos using a two-stage quality assessment," Jan. 7 2020, US Patent 10,529,066.
[285] T. Zhao, K. Zeng, A. Rehman, and Z. Wang, "On the use of SSIM in HEVC," in *Asilomar Conference on Signals, Systems and Computers*. IEEE, 2013, pp. 1107–1111.
[286] S. Deng, J. Han, and Y. Xu, "VMAF based rate-distortion optimization for video coding," in *IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.
[287] A. Chadha and Y. Andreopoulos, "Deep perceptual preprocessing for video coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 852–14 861.
[288] "Video pre-processing with JND-based Gaussian filtering of superpixels, author=Ding, Lei and Li, Ge and Wang, Ronggang and Wang, Wenmin," in *SPIE Visual Information Processing and Communication VI*, vol. 9410, 2015, pp. 20–25.
[289] C. Cai, Y. Wang, X. Li, and T. Ye, "Quality-constant per-shot encoding by two-pass learning-based rate factor prediction," *arXiv preprint arXiv:2208.10739*, 2022.
[290] H. Xing, Z. Zhou, J. Wang, H. Shen, D. He, and F. Li, "Predicting rate control target through a learning based content adaptive model," in *Picture Coding Symposium (PCS)*. IEEE, 2019, pp. 1–5.
[291] I. Katsavounidis, "Dynamic optimizer-A perceptual video encoding optimization framework," *The NETFLIX tech blog*, 2018.
[292] M. Wang, S. Wang, J. Li, L. Zhang, Y. Wang, and S. Ma, "SSIM Motivated Quality Control for Versatile Video Coding," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 1122–1127.
[293] T.-S. Ou, Y.-H. Huang, and H. H. Chen, "SSIM-based perceptual rate control for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 5, pp. 682–691, 2011.
[294] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.
[295] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Rate-SSIM optimization for video coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 833–836.
[296] ——, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 516–529, 2011.
[297] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5718–5727.
[298] X. Jiang, W. Tan, T. Tan, B. Yan, and L. Shen, "Multi-modality deep network for extreme learned image compression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, pp. 1033–1041, Jun. 2023.
[299] L.-H. Chen, C. G. Bampis, Z. Li, A. Norkin, and A. C. Bovik, "Proxiqa: A proxy approach to perceptual optimization of learned image compression," *IEEE Transactions on Image Processing*, vol. 30, pp. 360–373, 2021.
[300] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
[301] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Cycleisp: Real image restoration via improved data synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[302] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 4641–4650.
[303] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, and Y. Wang, "Intriguing findings of frequency selection for image deblurring," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 1905–1913, Jun. 2023.
[304] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5769–5780.
[305] Y. Zhong, J. Liu, X. Huang, J. Liu, Y. Fan, and M. Wu, "Cdcnet: A fast and lightweight dehazing network with color distortion correction," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 3020–3024.
[306] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 1833–1844.
[307] Z. Zhang, H. Wang, M. Liu, R. Wang, J. Zhang, and W. Zuo, "Learning raw-to-srgb mappings with inaccurately aligned supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 4348–4358.
[308] S.-K. Chen, H.-L. Yen, Y.-L. Liu, M.-H. Chen, H.-N. Hu, W.-H. Peng, and Y.-Y. Lin, "Learning continuous exposure value representations for single-image hdr reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 12 990–13 000.
[309] Z. Zhang, S. Zhang, R. Wu, W. Zuo, R. Timofte, X. Xing, H. Park, S. Song, C. Kim, X. Kong *et al.*, "Ntire 2024 challenge on bracketing image restoration and enhancement: Datasets methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 6153–6166.
[310] J. Li, B. Li, Z. Tu, X. Liu, Q. Guo, F. Juefei-Xu, R. Xu, and H. Yu, "Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 205–15 215.
[311] R. Zhu, S. Xu, P. Liu, S. Li, Y. Lu, D. Niu, Z. Liu, Z. Meng, Z. Li, X. Chen, and Y. Fan, "Zero-shot structure-preserving diffusion model for high dynamic range tone mapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 26 130–26 139.
[312] P. Shyam and H. Yoo, "Pair: Perception aided image restoration for natural driving conditions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 7459–7470.
[313] V. Laparra, A. Berardino, J. Ballé, and E. P. Simoncelli, "Perceptually optimized image rendering," *J. Opt. Soc. Am. A*, vol. 34, no. 9, pp. 1511–1525, Sep 2017. [Online]. Available: https://opg.optica.org/josaa/abstract.cfm?URI=josaa-34-9-1511
[314] P. Cao, C. Le, Y. Fang, and K. Ma, "A perceptually optimized and self-calibrated tone mapping operator," *arXiv preprint arXiv:2206.09146*, 2022.
[315] L. Liao, K. Xu, H. Wu, C. Chen, W. Sun, Q. Yan, C.-C. Jay Kuo, and W. Lin, "Blind video quality prediction by uncovering human video perceptual representation," *IEEE Transactions on Image Processing*, vol. 33, pp. 4998–5013, 2024.
[316] M. Delbracio, H. Talebei, and P. Milanfar, "Projected distribution loss for image enhancement," in *2021 IEEE International Conference on Computational Photography (ICCP)*, 2021, pp. 1–12.
[317] K. Panetta, S. K. K. M., S. P. Rao, and S. S. Agaian, "Deep perceptual image enhancement network for exposure restoration," *IEEE Transactions on Cybernetics*, vol. 53, no. 7, pp. 4718–4731, 2023.
[318] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao, "Pipal: A large-scale image quality assessment dataset for perceptual image restoration," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 633–651.
[319] G. Zhai, W. Sun, X. Min, and J. Zhou, "Perceptual quality assessment of low-light image enhancement," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 17, no. 4, Nov. 2021. [Online]. Available: https://doi.org/10.1145/3457905
[320] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," *arXiv preprint arXiv:2303.04226*, 2023.
[321] R. Li, P. Pan, B. Yang, D. Xu, S. Zhou, X. Zhang, Z. Li, A. Kadambi, Z. Wang, and Z. Fan, "4k4dgen: Panoramic 4d generation at 4k resolution," *arXiv preprint arXiv:2406.13527*, 2024.
[322] A. F. Di Natale, M. E. Simonetti, S. La Rocca, and E. Bricolo, "Uncanny valley effect: A qualitative synthesis of empirical research to assess the suitability of using virtual faces in psychological research," *Computers in Human Behavior Reports*, vol. 10, p. 100288, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2451958823000210
[323] J. Rao, C. Qiu, and M. Xiong, "Research on image processing and generative teaching in the context of aigc," in *2022 3rd International Conference on Information Science and Education (ICISE-IE)*, 2022, pp. 20–25.