One Transform To Compute Them All: Efficient Fusion-Based Full-Reference Video Quality Assessment

Abhinau K. Venkataramanan, Cosmin Stejerean, Ioannis Katsavounidis, and Alan C. Bovik, *Fellow, IEEE*

Abstract—The Visual Multimethod Assessment Fusion (VMAF) algorithm has recently emerged as a state-of-the-art approach to video quality prediction, that now pervades the streaming and social media industry. However, since VMAF requires the evaluation of a heterogeneous set of quality models, it is computationally expensive. Given other advances in hardware-accelerated encoding, quality assessment is emerging as a significant bottleneck in video compression pipelines. Towards alleviating this burden, we propose a novel Fusion of Unified Quality Evaluators (FUNQUE) framework, by enabling computation sharing and by using a transform that is sensitive to visual perception to boost accuracy. Further, we expand the FUNQUE framework to define a collection of improved low-complexity fused-feature models that advance the state-of-the-art of video quality performance with respect to both accuracy, by 4.2% to 5.3%, and computational efficiency, by factors of 3.8 to 11 times!

Index Terms—Full-Reference Video Quality Assessment, VMAF, FUNQUE, Contrast Sensitivity.

I. INTRODUCTION

The COVID-19 pandemic has accelerated a boom in online video consumption that has engulfed the internet in recent years. Despite returning to in-person work, education, and socializing in 2022, the total volume of internet video grew by 24% over 2021, accounting for more than 65% of the global internet volume. The top five streaming platforms alone accounted for nearly half of all video traffic [1]. Given the stakes of satisfying billions of viewers, modeling the perceptual quality of videos has emerged as a crucial area of research. Moreover, given the rapid rise of Augmented and Virtual Reality (AR/VR) platforms such as the Metaverse and the popularity of video-sharing social media platforms such as Instagram and TikTok, the amount of visual media being uploaded to and downloaded from the internet is expected to grow exponentially.

Objective models of perceptual quality are used to drive key decisions in the process of video delivery that affect the end-user's quality of experience. Perhaps the most important quality-related decision in the video delivery pipeline is the "encoding recipe." The "encoding recipe" refers to the set of parameters used by video codecs like AVC [2], HEVC [3], and AV1 [4] to encode a video clip before transmission.

Typical Adaptive Bitrate (ABR) streaming pipelines control two main parameters that define their encoding recipes -

bitrate, which is related to compression level, and the encoding resolution. The encoding resolution refers to the resolution to which videos are scaled before compression, and it is typically lower than the source resolution. Therefore, viewers often experience a combination of scaling and compression-based distortions such as blocking, blurring, and banding. When low-quality scalers such as bilinear interpolation are used, distortion due to aliasing may also be apparent.

Depending on the encoding complexity of a given source content, a "bitrate ladder" is constructed by analyzing the rate-distortion tradeoff offered by several combinations of bitrates and encoding resolutions. The Dynamic Optimizer (DO) [5] is an example of an algorithm that yields a perceptually-guided "Pareto-optimal" bitrate ladder. In fact, bitrate ladders constructed using DO are convex in the bitrate-distortion domain. The quality of the bitrate ladder depends on how well the rate-distortion tradeoff is characterized, which in turn relies on using an accurate model of perceptual quality.

Perceptual quality models are also used to benchmark video codecs and steer their improvement and development. The improvement in performance offered by one codec over another is measured in terms of the Bjontegaard-Delta Rate (BD-Rate), which is calculated with respect to perceptual quality models such as Structural Similarity (SSIM) [6] and Visual Multimethod Assessment Fusion (VMAF) [7] [8], and PSNR, which is a pixel-fidelity quality model. In short, the BD-Rate quantifies the bitrate savings offered by one codec or bitrate ladder over another when encoding quality is held constant.

Constructing bitrate ladders using algorithms such as DO may require hundreds of resizing/encoding operations and quality evaluations per scene. A state-of-the-art (SOTA) video quality model that is very widely used in this process is VMAF. VMAF is a fusion-based quality model that uses the outputs of "smaller" quality models (called "atom" quality models) as features, based on which the final quality score is computed. Therefore, while VMAF is an accurate predictor of human subjective judgments, it comes at a higher computational cost than models like SSIM and Multiscale SSIM (MS-SSIM) [9]. Coupled with the recent development of custom application-specific hardware for encoding and video processing [10] [11], this has led to the quality assessment step emerging as one of the computational bottlenecks.

We propose novel fusion-based full-reference video quality models for streaming applications based on a new framework that we call Fusion of Unified Quality Evaluators (FUNQUE).

FUNQUE uses a perceptually tuned wavelet-domain transform that is shared by all the atom quality models that are to be fused. In this way, FUNQUE enables extensive computation sharing, leading to an overall low computational complexity while also achieving higher accuracy than other SOTA fusion-based models. A flowchart depicting FUNQUE is shown in Figure 1, and a more detailed description is provided in Section III.

We make four novel contributions that culminate in the development of a suite of SOTA full-reference quality models that we call FUNQUE+:

1) **The FUNQUE framework** - FUNQUE is a novel framework for full-reference quality modeling that we first introduced in [12]. As described above, FUNQUE provides gains in both prediction accuracy and efficiency by using a shared perceptually tuned wavelet transform. The shared wavelet transform is used to efficiently compute a set of atom quality features that are fused to calculate video quality predictions.

2) **Contrast sensitivity functions** - The shared wavelet transform is perceptually sensitized through the use of contrast sensitivity function (CSF) models of the visibility of visual artifacts. We deploy seven CSF models from both the literature and specifically designed for use in FUNQUE.

3) **Atom quality models** - To facilitate the use of the shared transform, we design an extensive set of "atom" quality features that include both novel "quality-aware" features and others drawn from the literature. In particular, we demonstrate that the novel Multi-Scale Enhanced SSIM (MS-ESSIM) model designed here forms the backbone of FUNQUE+ models, and is by itself a leading predictor of video quality. Part of the success of FUNQUE+ also derives from the redesign of several existing quality features as described in Section V, all computed using the shared wavelet transform coefficients.

4) **Scalable Feature Selection** - While an extensive set of candidate features is desirable, we wanted the final fusion quality models to be compact and efficient. The use of traditional feature selection techniques such as exhaustive search or recursive feature elimination (RFE), also known as Greedy Feature Selection (GFS), is precluded by the large number of candidate features. Moreover, these techniques are not guaranteed to produce compact and efficient models. To overcome these limitations, we designed a Constrained Greedy Feature Selection (CGFS) method that uses "feature buckets" to scalably design compact fusion models.

The remainder of this paper is organized as follows. Section II provides background regarding full-reference quality assessment, with a focus on recent fusion-based quality models, and Section III provides a recap of the FUNQUE model. Section IV describes the various methods used to incorporate models of the CSF into the HVS-sensitive wavelet transform, and Section V describes the atom quality features extracted in the transform domain. Following this, Section VI describes the experimental setup and the feature selection method used to develop our models, and Section VII details the results of our experimental validation. Finally, Section VIII concludes with a summary of our findings and possible directions for the future.

II. BACKGROUND

Objective models of image and video quality may be broadly categorized as Full-Reference (FR), No-Reference (NR), or Reduced-Reference (RR), depending on the availability of pristine reference contents. Since streaming services and video compression algorithms have direct access to pristine source content, we only consider the FR quality modeling problem here.

The Structural Similarity (SSIM) index [6] is a widely deployed perceptual objective FR video quality model used to control the quality of broadcast and streaming television content. The pervasive use of SSIM is due to the significant improvements it provides over the legacy PSNR model, which remains in use today, particularly in developing new video codec standards. The success of SSIM helped catalyze the development of more sophisticated FR quality models such as Multi-Scale SSIM (MS-SSIM) [9], the Feature Similarity index (FSIM) [13], Visual Information Fidelity (VIF) [14], and the Detail Loss Metric (DLM) [15], as well as reduced-reference models such as Spatio-Temporal Reduced Reference Entropic Differencing (ST-RRED) [16].

The aforementioned algorithms were generally designed based on models of perception and Natural Scene Statistics (NSS), and do not utilize data-driven machine-learning techniques. In a further advance, the Visual Multimethod Assessment Fusion (VMAF) [7] model deployed fusion-based quality assessment. Fusion-based quality models employ a "mixture-of-experts" approach by using smaller learning-free quality models, called "atom quality models," as features to train a machine-learning model, such as a Support Vector Regressor (SVR). Similar machine learning techniques, e.g., combining "quality-aware" features and models using SVRs was already common practice in the design of NR video quality algorithms, years prior to VMAF [17]–[20].

VMAF uses six features in total - scalar pixel-domain VIF computed at four Gaussian scales, DLM computed using a 4-level Db2 Discrete Wavelet Transform (DWT), and a "Motion" feature that characterizes the amount of temporal information present in the reference video. The computation of two 4-level multi-scale decompositions - the Gaussian pyramid in VIF and the Db2 DWT in DLM - are the primary reasons for VMAF's high computational complexity. Nevertheless, due to its improved accuracy as compared with the aforementioned models, VMAF has emerged as a SOTA FR perceptual video quality model in the streaming and compression space, alongside SSIM and MS-SSIM. Another major driver of VMAF's widespread adoption is the development of an efficient fixed-point implementation that significantly improves its computational complexity. [21] While this technique is not explored here, fixed-point implementations may be, in principle, used to accelerate any of the models we discuss.

Attempts at improving VMAF have typically focused on adapting it as an optimization loss function in perceptual

optimization problems or improving its accuracy. A key hurdle faced when attempting to use VMAF as a training objective for deep learning algorithms is its non-differentiability. To address this, differentiable approximations of VMAF, such as ProxIQA [22] and ProxVQM [23] have been developed, which adapt VMAF into a viable perceptual objective function to train deep image/video processing and compression algorithms. VMAF has also been used to predict Just Noticeable Differences (JND) [24] for better design of ABR bitrate ladders. VMAF's atoms, and in particular VIF and DLM, were designed to capture both quality degradations and quality improvements and as such, VMAF carried over this feature, making it susceptible to reporting higher quality numbers after applying contrast enhancement and sharpening, as was demonstrated in [25]. This prompted the development of the "No Enhancement Gain" VMAF model (VMAF-NEG) [26], which is constrained to only measure quality degradations.

On the other hand, attempts at improving VMAF's accuracy typically involve identifying better atom quality features. The Spatio-Temporal VMAF (ST-VMAF) and Ensemble VMAF (Ens-VMAF) [27] models introduce temporal quality-aware features derived from the Spatial Efficient Entropic Differencing (SpEED) quality models. In addition, Ens-VMAF uses an ensemble of two fusion models that use complementary features to boost accuracy. Enhanced VMAF (Enh-VMAF) [28] further builds on this theme by incorporating temporal information using optical flow-based dynamic texture features (DTFs), ensemble modeling, and adding chroma-channel features to the feature pool.

VMAF has also found applications in domains other than standard HD video streaming, including chroma compression [29], 360 VR [30], high frame rate video [31], and medical videos [32]. Notably, all the aforementioned approaches to improving VMAF involve increasing the model size by adding more computationally complex features. Therefore, while these newer models may advance the SOTA in terms of accuracy, they exacerbate the computational cost problem faced by VMAF. Given the already considerable computational demands of VMAF, we will show that our FUNQUE and FUNQUE+ models are significantly less expensive than other fusion-based models, while also achieving higher accuracy.

Taking inspiration from the pooling strategy used by DLM, which computes quality over only the central 64% area of an image, restricting the spatial region may lead to complexity improvements for any quality model. Furthermore, temporal subsampling [33], [34] has also emerged as a legitimate strategy for further reducing computational complexity, particularly when paired with scene-change-detection algorithms. While these methods are not discussed here, they may be applied to any of the quality models described here.

Deep-learning methods for quality modeling have gained popularity in recent years. Even before application to picture quality assessment, distances between VGG [35] features have been used as a perceptual loss to guide deep image generation tasks such as super-resolution and style-transfer [36], [37]. The use of feature differences as a perceptual similarity metric was formalized by the Learned Perceptual Image Patch Similarity (LPIPS) [38] model, which uses a linear model to map differences between AlexNet [39] features to similarity scores. Adversarially-robust versions of LPIPS such as E-LPIPS [40] and R-LPIPS [41] have also been developed using adversarial retraining. More recently, transformers have also been used to map differences between deep features to quality scores [42], though these methods fall beyond the range of practical computational complexities of interest here.

The Deep Image Structure and Texture Similarity (DISTS) [43] model adopts a different approach to comparing deep feature maps. The mean and standard deviation of VGG feature maps are calculated and compared using SSIM-like metrics, which function as perceptual distance measures. Finally, the Deep Wasserstein Distance (DeepWSD) model [44] compares feature maps using statistical similarity metrics instead of pixel similarity/difference metrics. In particular, DeepWSD estimates the Wasserstein distance between the distributions of deep feature coefficients obtained from a pair of compared images.

III. FUNQUE

The foundation of the FUNQUE framework [12] is a "unified transform" that is designed to be sensitive to properties of the Human Visual System (HVS). The role of this transform is two-fold. First, reusing the output of the unified transform for all of the atom quality models improves efficiency through computation sharing. Secondly, incorporating HVS-sensitivity into all the atom quality models improves accuracy against human judgments of perceptual video quality.

The latter is achieved by combining models of contrast sensitivity with the multi-scale frequency-selectivity of discrete wavelet transforms (DWTs). In particular, FUNQUE uses a 21-tap spatial filter based on the CSF model proposed in [45], and a Haar wavelet transform to obtain the final transform coefficients. A detailed treatment of the CSF model used in FUNQUE, and the new models developed for FUNQUE+, is provided in Section IV.

The FUNQUE model designed in [12] also uses the Self Adaptive Scale Transform (SAST) [46], which is a technique that involves rescaling reference and test input frames prior to quality estimation to account for viewing conditions. Similar to the use of the shared unified transform, SAST also serves the dual purpose of simultaneously improving efficiency and accuracy.

The transformed wavelet coefficients are then used to compute atom quality features for fusion. While FUNQUE is a general framework that may be used with any wavelet-domain features, the model proposed in [12] used Enhanced SSIM (ESSIM) [33], VIF [14], DLM [15] and motion as the atom quality features. Through an ablation study, it was found that ESSIM in particular contributed over 50% of the observed improvement in accuracy. Finally, these features were fused using a non-linear SVR model. A detailed description of the features used in FUNQUE, and the new features developed for FUNQUE+, is provided in Section V.

In summary, FUNQUE provides a framework for efficient, accurate fusion-based FR video quality models. The prototype
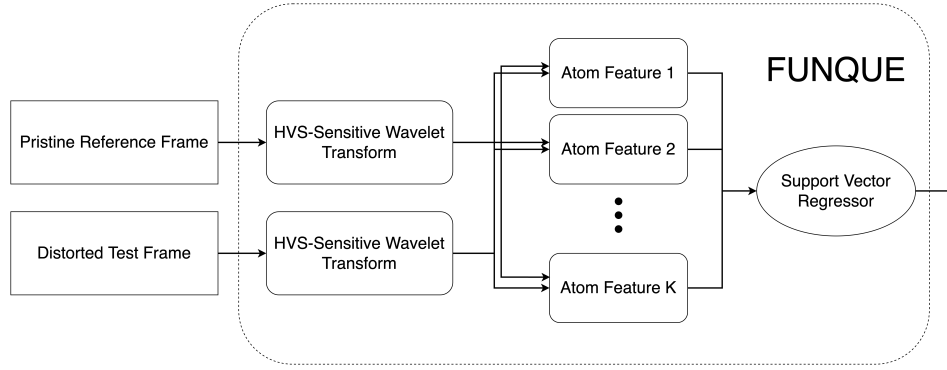
Fig. 1. The FUNQUE framework for efficient fusion-based video quality modeling

model[1] presented in [12] provides a proof-of-concept, but the development and analysis of the model left significant room for improvement, which is addressed here. The feature set that defines FUNQUE consists of pre-existing features that were primarily borrowed from VMAF. By contrast, here we develop a much wider range of novel features with an emphasis on efficiency and wavelet-domain compatibility while incorporating chromatic and spatio-temporal features.

Furthermore, the CSF model used in the original FUNQUE, which was used to demonstrate the impact of the unified transform, was directly borrowed from DLM [15]. Here, we conduct a much more extensive survey and analysis of efficient CSF models, and develop some of our own, towards developing a better Unified Transform model. Finally, we provide a significantly broader and deeper analysis of all the selected models, including extensive cross-database testing, statistical significance tests, computational complexity analysis, and monotonicity analysis. In this manner, we demonstrate significant advances over the prototype FUNQUE model [12] and are able to show that FUNQUE+ significantly exceeds the performances of state-of-the-art FR video quality models.

IV. Contrast Sensitivity and the Unified Transform

A. The FUNQUE Unified Transform

Broadly, the HVS-sensitive unified transform involves filtering the input frames using a model of human contrast sensitivity, followed by processing them with a suitable wavelet transform to conduct perceptually sensitized multi-scale analysis. We deploy a model of the CSF to capture information regarding the visibility of artifacts as a function of spatial frequency. In fact, we tested and compared several CSF models, as described in the following.

FUNQUE is a general modular framework that is based on a unified transform [12] modulated by a simple model of contrast sensitivity, proposed by Ngan et al. [45]:

$$\text{NganCSF}(f) = (0.31 + 0.69f)e^{-0.29f}, \quad (1)$$

where f is expressed in units of cycles/degree.

Since applying Fourier transforms on HD frames is computationally expensive, the Ngan CSF model can be translated

[1]The prototype FUNQUE model was presented at IEEE ICIP 2022.

into a spatial (angle) domain filter by computing an analytical inverse Fourier Transform:

$$\text{NganSpat}(\theta) = \frac{2\left(0.0656 - 23.6910\,\theta^2\right)}{\left(0.0841 + 39.4784\,\theta^2\right)^2}. \quad (2)$$

When viewing video content on a 1080p display of height H, placed at a distance D, the angle subtended by a pixel at the eye is

$$\Delta\theta \approx \tan(\Delta\theta) = \frac{H}{D \times 1080} \times \frac{180}{\pi} \quad (3)$$

in degrees, which is the interval at which the above "continuous-angle" filter must be sampled. The corresponding discrete spatial domain filter is

$$\text{NganSpat}[n] = \Delta\theta \times \text{NganSpat}(n\Delta\theta). \quad (4)$$

Since FUNQUE was trained for TV viewing, a nominal value of $D/H = 3.0$ was used, as in VMAF. Since the discrete spatial filter has infinite support, it was truncated to 21 taps, at which point, the value of the tail is less than 5% of the peak value.

B. Towards Improved HVS Modeling

A key drawback of FUNQUE's unified transform is the use of a 21-tap filter, which runs contrary to the goal of computational efficiency. The spatial filter form of the Ngan CSF, as used in the DLM algorithm [15], assigns a multiplicative weight to each subband. The Li subband weight (LiSW) is computed as a function of the subband $\theta \in \{H, V, D\}$, where H, V, D denote horizontal, vertical, and diagonal wavelet subbands, indexed by wavelet levels $\lambda = 1, 2, \ldots$

$$\text{LiSW}(\lambda, \theta) = \text{NganCSF}\left(f_{nom}(\lambda, \theta)\right), \quad (5)$$

where $f_{nom}(\lambda, \theta)$ is the nominal frequency assigned to subband θ at level λ:

$$f_{nom}(\lambda, \theta) = \frac{1080 \times \pi \times (D/H)}{2^\lambda \times 180 \times (0.15p(\theta) + 0.85)}, \quad (6)$$

where $p(\theta) = 1$ if $\theta \in \{H, V\}$, and $p(\theta) = -1$ if $\theta = D$. This captures the oblique effect [47], whereby the HVS is less sensitive to diagonal subbands than horizontal and vertical subbands.

In addition to the Ngan CSF model, we considered two other frequency-domain models. The second CSF model we considered was developed by Nadenau et al. [48], who used it to augment wavelet-domain color image compression. Since images are typically compressed in opponent color spaces, three models of CSF were developed in [48], corresponding to the "luminance," "red-green," and "blue-yellow" channels. All three CSFs are described in the frequency domain by the functional form

$$\text{NadenauCSF}(f) = \left(1 + 255e^{-bf^c}\right)/256, \qquad (7)$$

where b and c vary across three channels. Note that unlike the NganCSF, which is bandpass, the Nadenau CSF has a lowpass response in all three channels. As with the NganCSF, we would like to apply the NadenauCSF as a spatial filter (NadenauSpat) to avoid computing Fourier Transforms. An analytical inverse proved difficult to compute due to the presence of f^c in the exponential function. Therefore, numerical inverses were calculated using the Fast Fourier Transform (FFT). The filters were truncated to yield 5% error in the tail, leading to 5-tap filters for the Y and Cr channels and a 7-tap filter for the Cb channel. Note that all three NadenauSpat CSF filters are significantly more compact than the 21-tap NganSpat CSF filter.

Similar to the approach followed by Li et al. [15], we also consider a subband-weighted version of the Nadenau CSF as a coarse yet efficient method of incorporating HVS-sensitivity. The Nadenau subband weights are computed as

$$\text{NadenauSW}(\lambda, \theta) = \text{NadenauCSF}\left(f_{nom}(\lambda, \theta)\right), \qquad (8)$$

where f_{nom} is assigned as in (6).

A third model of the CSF that we considered was formulated by Larson et al. [49], which is a modified version of the classic Mannos-Sakrison CSF model [50]. Larson's model is expressed as a function of radial spatial frequency f_r and orientation ϕ:

$$\text{LarsonCSF}(f_r, \phi)$$
$$= \begin{cases} (0.0499 + 0.5928f_\phi)e^{(0.228f_\phi)^{1.1}}, f_r \geq 4 \\ 0.981, \quad f_r < 4 \end{cases}, \quad (9)$$

where $f_\phi = f_r/(0.15\cos(4\phi) + 0.85)$ is a corrected radial frequency that accounts for the oblique effect. Since the LarsonCSF does not yield a small spatial filter when transformed into the spatial domain, we only consider the subband-weighting method, with weights given by

$$\text{LarsonSW}(f, \theta) = \text{LarsonCSF}(f_{nom}(\lambda, \theta), 0). \qquad (10)$$

Once again, the nominal frequencies were computed as in (6), and the orientation ϕ was set to zero since the oblique effect is already accounted for by f_{nom}.

Note that all the aforementioned CSFs encode the visibility of various spatial frequencies. However, since we aim to use wavelet decompositions of the two input frames, we also considered models of wavelet subband visibility. In particular, we consider two models of wavelet subband CSFs, by Watson et al. [51] and Hill et al. [52]. Since both these wavelet CSF

TABLE I
SUMMARY OF CSF METHODS

Method	Base CSF Model	Domain	Color Aware?
NganSpat	Ngan [45]	Pixel	No
LiSW	Ngan [45]	Wavelet	No
NadenauSpat	Nadenau [48]	Pixel	Yes
NadenauSW	Nadenau [48]	Wavelet	Yes
LarsonSW	Larson [49]	Wavelet	No
WatsonSW	Watson [51]	Wavelet	Yes
HillSW	Hill [52]	Wavelet	No

models are applied as subband-weighting mechanisms, we will refer to them as the Watson SW and Hill SW CSF models.

Therefore, in summary, we have experimented with seven methods of applying CSFs as part of defining unified transforms in the FUNQUE framework. Key details regarding the CSF methods used in our experiments are summarized in Table I.

C. Self Adaptive Scale Transform

The Self-Adaptive Scale Transform (SAST) is a preprocessing method proposed in [46] to account for viewing conditions, which are described by the viewing distance and the height of the display. When SAST is applied prior to quality assessment, both the reference and test video frames are downscaled by a factor of

$$\alpha_{\text{SAST}} = \sqrt{4\tan\left(\theta_H/2\right)\tan\left(\theta_W/2\right) \cdot \frac{D}{H} \cdot \frac{D}{W}} \qquad (11)$$

where θ_W and θ_H denote the angular width and height of the field of view, W and H denote the physical width and height of the display, and D denotes the viewing distance.

Assuming a standard 16:9 display, and using typical values of $\theta_H = 40°$ and $\theta_W = 50°$, then

$$\alpha_{\text{SAST}} \approx \frac{D/H}{1.618}. \qquad (12)$$

In the following developments, we assume a D/H ratio of 3.0, which corresponds to typical TV viewing conditions. For simplicity, we round α_{SAST} to the nearest integer, yielding a scale factor of two. As described in Section VII, we have found that applying SAST improves the accuracy of the FUNQUE+ models. This observation mirrors the improvement in accuracy observed by both ESSIM and FUNQUE due to the application of SAST [12], [33].

However, the use of SAST introduces a vulnerability in the FUNQUE+ model that may be exploited when it is used as a performance metric in applications such as perceptual optimization and codec evaluation. For example, since both reference and test frames are processed at half-resolution by FUNQUE+, a codec may achieve high (or even perfect) FUNQUE+ scores by also processing frames at half-resolution. However, this would clearly not lead to lossless coding since operating at half-resolution is inherently lossy. Therefore, we also report "full-scale" (FS) FUNQUE+ models that do not utilize SAST. Such models typically achieve lower accuracies but do not suffer the same vulnerability.

V. FEATURE EXTRACTION FROM THE UNIFIED TRANSFORM DOMAIN

The FUNQUE framework for quality modeling is founded on the use of a shared transform from which all atom quality features are computed. Since the shared transform is based on a perceptually-sensitive wavelet transform, careful development of atom quality models is needed to enable maximum computation sharing. When adopting atom quality models from the literature, wavelet-domain models such as the Detail Loss Metric (DLM) [15] may be used off-the-shelf. However, spatial-domain models are not compatible with FUNQUE.

To adopt spatial domain models, we identify analogous quantities that may be computed in the wavelet domain. For example, horizontal and vertical subbands may be used as substitutes for gradients in gradient-based methods, and local means and variances within square windows may be computed from wavelet subbands using an approach similar to [12], which will also be explained in Section V-A. Such adaptations are crucial to the success of FUNQUE and FUNQUE+, and in this section, we describe the use of these techniques to develop our candidate feature set.

A. Multi-Scale Enhanced Structural Similarity

As shown in [12], a key driver of the accuracy of the prototype model FUNQUE is the use of Enhanced Structural Similarity (ESSIM) [33]. In its original form, ESSIM was computed in the spatial domain similar to SSIM. However, since the FUNQUE framework is predicated on computing all features from wavelet decompositions of input frames from the reference and test videos, we have developed a method in [12] to compute spatial-domain structural similarity (SSIM) directly from Haar wavelet coefficients. Local SSIM scores obtained in this manner are used to compute Enhanced SSIM [33], which uses the Coefficient of Variation (CoV) to conduct spatial aggregation and has been shown to outperform baseline SSIM. In this subsection, we review the wavelet-domain computation of ESSIM and propose a novel wavelet-domain Multi-Scale ESSIM (MS-ESSIM) quality model.

Consider a 2×2 block of an input image x. A 1-level Haar transform of this block may be expressed as

$$
\begin{bmatrix} X_{1,A} \\ X_{1,H} \\ X_{1,V} \\ X_{1,D} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} +1, +1, +1, +1 \\ +1, -1, +1, -1 \\ +1, +1, -1, -1 \\ +1, -1, -1, +1 \end{bmatrix} \begin{bmatrix} x(i,j) \\ x(i,j+1) \\ x(i+1,j) \\ x(i+1,j+1) \end{bmatrix}. \quad (13)
$$

Consider a pair of images x and y of size $M \times N$ such that both M and N are divisible by 2^L. Assuming L-level Haar transforms of each, let $X_{\lambda,\theta}$ and $Y_{\lambda,\theta}$ denote the subband $\theta \in \{A, H, V, D\}$, which corresponds to the approximation, horizontal, vertical, and diagonal subbands, at level $\lambda \in \{1, \ldots, L\}$. Since the transformation matrix is a scaled orthogonal matrix, inner products between 2×2 blocks of the input images may be obtained from inner products between their corresponding wavelet coefficients with appropriate scaling. This argument may be extended further to include $2^L \times 2^L$ blocks when using an L-level Haar transform.

Furthermore, local variances of images and covariances between images may be expressed as scaled inner products between corresponding blocks. Therefore, using the properties described above, local statistics from disjoint $2^L \times 2^L$ blocks may be obtained from Haar wavelet coefficients as follows:

$$
\mu_{x,L}(i,j) = 2^{-L} X_{L,A}(i,j), \quad (14)
$$

$$
\mu_{y,L}(i,j) = 2^{-L} Y_{L,A}(i,j), \quad (15)
$$

$$
\sigma_{x,L}^2(i,j) = 2^{-2L} \sum_{k=1}^{L} \sum_{P_{ij}^k} \sum_{\{H,V,D\}} X_{k,\theta}^2(m,n), \quad (16)
$$

$$
\sigma_{y,L}^2(i,j) = 2^{-2L} \sum_{k=1}^{L} \sum_{P_{ij}^k} \sum_{\{H,V,D\}} Y_{k,\theta}^2(m,n), \quad (17)
$$

$$
\sigma_{xy,L}(i,j) = 2^{-2L} \sum_{k=1}^{L} \sum_{P_{ij}^k} \sum_{\{H,V,D\}} X_{k,\theta}(m,n) Y_{k,\theta}(m,n), \quad (18)
$$

where $P_{ij}^k = \{(m,n) \mid i2^{L-k} \le m < (i+1)2^{L-k}, j2^{L-k} \le n < (j+1)2^{L-k}\}$ denotes disjoint $2^{L-k} \times 2^{L-k}$ blocks. These local statistics may be used to compute both SSIM and ESSIM in the wavelet domain, as in [33].

We extend this framework to compute MS-SSIM and MS-ESSIM by computing the aforementioned statistics at all levels $\lambda \le L$, as opposed to only computing them at level L. Our multi-scale definition of wavelet-domain MS-(E)SSIM corresponds to computing spatial-domain (E)SSIM using windows that double in size between successive scales. This is analogous to the method used by traditional MS-SSIM, where images are downsampled by a factor of two between successive scales. Altering the multi-scale computation in this manner allows the computation of statistics at each scale in an iterative manner as shown below

$$
\sigma_{x,\lambda+1}^2(i,j) = 2^{-2(\lambda+1)} \sum_{k=1}^{\lambda+1} \sum_{P_{ij}^k} \sum_{\{H,V,D\}} X_{k,\theta}^2(m,n) \quad (19)
$$

$$
= 2^{-2} \sum_{m=i}^{i+1} \sum_{n=j}^{j+1} \sigma_{x,\lambda}(m,n)^2 + \tilde{\sigma}_{x,\lambda+1}^2(i,j), \quad (20)
$$

where

$$
\tilde{\sigma}_{x,\lambda+1}^2(i,j) = 2^{-2(\lambda+1)} \sum_{\{H,V,D\}} X_{\lambda+1,\theta}^2(i,j). \quad (21)
$$

A similar iterative method is used to compute $\sigma_{y,\lambda}^2(i,j)$ and $\sigma_{xy,\lambda}(i,j)$ at each level from the previous levels. In this manner, we obtain local statistics at all scales using the same number of operations that computing single-scale (E)SSIM from an L-level wavelet decomposition would require.

Using the multi-scale local statistics computed in this manner, pooled luminance and contrast-structure similarity scores at each level may be obtained as

$$l_{\text{Pool},\lambda} = \text{Pool}\left(\frac{2\mu_{x,\lambda}\mu_{y,\lambda} + K_1}{\mu_{x,\lambda}^2 + \mu_{y,\lambda}^2 + K_1} \right) \tag{22}$$

$$cs_{\text{Pool},\lambda} = \text{Pool}\left(\frac{2\sigma_{xy,\lambda} + K_2}{\sigma_{x,\lambda}^2 + \sigma_{y,\lambda}^2 + K_2} \right), \tag{23}$$

from which the level-L SSIM and ESSIM are computed by pooling the products of the luminance and contrast-similarity scores from level L. Note that the Pool function is set to mean pooling when computing (MS-)SSIM, and to CoV pooling when computing (MS-)ESSIM. The final MS-SSIM and MS-ESSIM scores are obtained by combining scores across levels using the MS-SSIM exponents α_λ [9]:

$$\text{MS-SSIM}_L = (\text{SSIM}_L)^{\alpha_L} \prod_{\lambda=1}^{L-1} (cs_{\text{Mean},\lambda})^{\alpha_\lambda} \tag{24}$$

$$\text{MS-ESSIM}_L = (\text{ESSIM}_L)^{\alpha_L} \prod_{\lambda=1}^{L-1} (cs_{\text{CoV},\lambda})^{\alpha_\lambda}. \tag{25}$$

B. Information-Theoretic Features

As in FUNQUE, we use multi-scale Visual Information Fidelity (VIF) [14] features. The VIF features, which are derived under a scalar Gaussian Scale Mixture (GSM) model, are computed from wavelet subbands at each level. In short, local means $(\mu_{x,\lambda,\theta}, \mu_{y,\lambda,\theta})$, variances $(\sigma_{x,\lambda,\theta}^2, \sigma_{y,\lambda,\theta}^2)$, and covariances $(\sigma_{xy,\lambda,\theta})$ from subbands are computed using highly efficient integral images [33]. Note that this notation extends the denotation of local statistics in Section V-A, where $\lambda = 1 \ldots L$ denotes the wavelet level, and $\theta \in \{A, H, V, D\}$ denotes the wavelet subband. For example, $\mu_{x,1,A}(i,j)$ denotes local means computed from the approximation subband in the first level of the wavelet decomposition of the reference image.

As described in [14], VIF assumes that distortions arise from a channel that is modeled as

$$Y_{\lambda,\theta}(i,j) = g_{\lambda,\theta}(i,j)X_{\lambda,\theta}(i,j) + N_{\lambda,\theta}(i,j), \tag{26}$$

where $g_{\lambda,\theta}(i,j)$ is the gain of the distortion channel and $N_{\lambda,\theta}(i,j) \sim \mathcal{N}(0, \sigma_v^2)$ is additive white Gaussian noise.

As described in [14], local statistics are used to estimate channel parameters as:

$$g_{\lambda,\theta}(i,j) = \sigma_{xy,\lambda,\theta}(i,j) / \sigma_{x,\lambda,\theta}^2(i,j) \tag{27}$$

$$\sigma_{v,\lambda,\theta}^2(i,j) = \sigma_{y,\lambda,\theta}^2(i,j) - g_{\lambda,\theta}(i,j)\sigma_{xy,\lambda,\theta}(i,j) \tag{28}$$

The estimated channel parameters are used to compute VIF as a ratio of mutual information measures between the reference and test images. For more details, we refer the reader to [14]. When using approximation subbands at each level λ to compute VIF, we denote it by VIF-A$_\lambda$, and it is computed as:

$$\text{VIF-A}_\lambda = \frac{\sum_{i,j} \log\left(1 + \frac{g_{\lambda,A}^2(i,j)\sigma_{x,\lambda,A}^2(i,j)}{\sigma_{v,\lambda,A}^2(i,j)+\sigma_n^2}\right)}{\sum_{i,j} \log\left(1 + \frac{\sigma_{x,\lambda,A}^2(i,j)}{\sigma_n^2}\right)}, \tag{29}$$

where σ_n^2 is a small constant used to model neural noise. Note that despite the change in notation, these features are identical to the "VIF-Scale" features used in [12].

On the other hand, the definition of VIF in [14] involves applying a vector GSM model on horizontal and vertical subbands of a multi-scale steerable pyramid wavelet decomposition and combining features from all scales. Therefore, analogous to this original formulation of VIF, we compute an L-level "scalar-GSM VIF" on the H and V subbands as

$$\text{VIF-HV}_L = \frac{\sum_{\lambda \leq L} \sum_{\theta \in \{H,V\}} \sum_{i,j} \log\left(1 + \frac{g_{\lambda,\theta}^2(i,j)\sigma_{x,\lambda,\theta}^2(i,j)}{\sigma_{v,\lambda,\theta}^2(i,j)+\sigma_n^2}\right)}{\sum_{\lambda \leq L} \sum_{\theta \in \{H,V\}} \sum_{i,j} \log\left(1 + \frac{\sigma_{x,\lambda,\theta}^2(i,j)}{\sigma_n^2}\right)}. \tag{30}$$

A key drawback of VIF is that it is only a measure of spatial quality, and does not include a temporal component. The Spatio-Temporal Reduced Reference Entropic Differencing (ST-RRED) [16] quality model improves on VIF in this regard by defining measures of both spatial and temporal quality. Inspired by ST-RRED, which is defined in the steerable pyramid [53] domain on both frames and frame differences, we compute multi-scale ST-RRED features on the approximation subbands of the Haar transform.

To explain how this is done, we first discuss the computation of the spatial component, termed spatial RRED (SRRED). Using the same local statistics as used by VIF, local information-weighted entropies in subband θ at level λ are computed as

$$h_{\lambda,\theta}(i,j) = \alpha_{\lambda,\theta}(i,j) \log\left(2\pi e(\sigma_{\lambda,\theta}^2(i,j) + \sigma_n^2)\right), \tag{31}$$

where the weighting factors are given by

$$\alpha_{\lambda,\theta}(i,j) = \log\left(1 + \sigma_{\lambda,\theta}^2(i,j)\right). \tag{32}$$

As with our treatment of VIF above, we consider two versions of SRRED. The first is defined on the approximation subbands at various scales, similar to VIF-A.

$$\text{SRRED-A}_\lambda = \frac{1}{MN} \sum_{i,j} |h_{x,\lambda,A}(i,j) - h_{y,\lambda,A}(i,j)|. \tag{33}$$

The second definition is similar to the definition of VIF-HV, where SRRED is computed on the H and V subbands at all levels.

$$\text{SRRED-HV}_L = \frac{1}{MN} \sum_{\lambda,\theta} \sum_{i,j} |h_{x,\lambda,\theta}(i,j) - h_{y,\lambda,\theta}(i,j)|. \tag{34}$$

To compute Temporal RRED (TRRED), a similar analysis is carried out on the differences between approximation subbands from adjacent frames. Let the local statistics of these "differenced subbands" be denoted by $m_{\lambda,\theta}$, $s_{\lambda,\theta}^2$, and $s_{xy,\lambda,\theta}$. Then, the local weighted entropies are given by

$$g_{\lambda,\theta}(i,j) = \beta_{\lambda,\theta}(i,j) \log\left(2\pi e\left(s_{\lambda,\theta}^2(i,j) + \sigma_n^2\right)\right), \tag{35}$$

where the weighting factors are

$$\beta_{\lambda,\theta}(i,j) = \log\left(1 + \sigma_{\lambda,\theta}^2(i,j)\right) \log\left(1 + s_{\lambda,\theta}^2(i,j)\right). \tag{36}$$

The two variants of TRRED are then given by

$$\text{TRRED-A}_\lambda = \frac{1}{MN} \sum_{i,j} |g_{x,\lambda,A}(i,j) - g_{y,\lambda,A}(i,j)|. \tag{37}$$

and

$$\text{TRRED-HV}_L = \frac{1}{MN} \sum_{\lambda,\theta} \sum_{i,j} |g_{x,\lambda,\theta}(i,j) - g_{y,\lambda,\theta}(i,j)|. \tag{38}$$

Finally, the combined STRRED features are arrived at:

$$\text{STRRED-A}_\lambda = \text{SRRED-A}_\lambda \times \text{TRRED-A}_\lambda \tag{39}$$

and

$$\text{STRRED-HV}_L = \text{SRRED-HV}_L \times \text{TRRED-HV}_L. \tag{40}$$

C. Detail Loss Metric

The Detail Loss Metric [15] is a wavelet-domain full-reference quality model that measures the amount of "detail loss" suffered by a test video frame in comparison to a reference video frame. In the following, we summarize the DLM algorithm. Note that due to the use of the shared HVS-sensitive unified transform, we omit the "contrast sensitivity function" step described in [15].

The first step in computing DLM involves applying a "decoupling step." The decoupling step is based on the following distortion model used to describe wavelet subband coefficients $\theta \in \{H, V, D\}$. Let the images x and y be reference and test images, respectively. The distortion model assumed by DLM is given by

$$Y_{\lambda,\theta}(i,j) = \gamma_{\lambda,\theta}(i,j) X_{\lambda,\theta}(i,j) + A_{\lambda,\theta}(i,j), \tag{41}$$

where the gain factor γ models attenuation of local gradients due to detail loss, $R_{\lambda,\theta}(i,j) = \gamma_{\lambda,\theta}(i,j) X_{\lambda,1}(i,j)$ are the "restored" coefficients, and $A_{\lambda,\theta}(i,j)$ are the "additive impairments."

The "restored" wavelet decomposition is computed from the given frames as

$$\hat{R}_{\lambda,\theta}(i,j) = \hat{\gamma}_{\lambda,\theta}(i,j) X_{\lambda,\theta}(i,j), \tag{42}$$

where

$$\psi_{x,\lambda}(i,j) = \arctan\left(\frac{X_{\lambda,V}(i,j)}{X_{\lambda,H}(i,j)}\right), \tag{43}$$

$$\psi_{y,\lambda}(i,j) = \arctan\left(\frac{Y_{\lambda,V}(i,j)}{Y_{\lambda,H}(i,j)}\right), \tag{44}$$

$$\Delta\psi_\lambda(i,j) = |\psi_{x,\lambda}(i,j) - \psi_{y,\lambda}(i,j)|, \tag{45}$$

and

$$\hat{\gamma}_{\lambda,\theta}(i,j) = \begin{cases} \frac{Y_{\lambda,\theta}(i,j)}{X_{\lambda,\theta}(i,j)}, & \Delta\psi_\lambda(i,j) < 1° \\ \text{clip}\left(\frac{Y_{\lambda,\theta}(i,j)}{X_{\lambda,\theta}(i,j)}, 0, 1\right), & else \end{cases}. \tag{46}$$

The quantity $\Delta\psi$ is used to preserve contrast enhancement, which scales both the horizontal and vertical subband coefficients of the distorted frame. The additive impairment coefficients are then computed as

$$\hat{A}_{\lambda,\theta}(i,j) = Y_{\lambda,\theta}(i,j) - \hat{R}_{\lambda,\theta}(i,j). \tag{47}$$

The additive impairments are used to mask the restored coefficients as

$$\tilde{R}_{\lambda,\theta}(i,j) = \left(\hat{R}_{\lambda,\theta}(i,j) - M_\lambda(i,j)\right)^+, \tag{48}$$

where

$$M_\lambda(i,j) = \sum_\theta \sum_{i-1}^{i+1} \sum_{j-1}^{j+1} w_{ij}(m,n) \left|\hat{A}_{\lambda,\theta}(m,n)\right|, \tag{49}$$

$$w_{ij}(m,n) = \frac{1 + \delta(m-i, n-j)}{30}, \tag{50}$$

$(\cdot)^+$ denotes clipping negative values to zero, and δ is the Kronecker delta function.

Unlike the formulation used in VMAF and FUNQUE, we compute DLM on a "per-scale" basis. The value at scale λ is computed as

$$\text{DLM-S}_\lambda = \frac{\sum\limits_{\theta \in \{H,V,D\}} \left(\sum\limits_{i,j} \tilde{R}_{\lambda,\theta}(i,j)^3\right)^{1/3}}{\sum\limits_{\theta \in \{H,V,D\}} \left(\sum\limits_{i,j} |X_{\lambda,\theta}(i,j)|^3\right)^{1/3}}. \tag{51}$$

During feature selection, the "DLM" feature bucket allows the selection of either all DLM scales (DLM-S$_\lambda$; $\lambda \leq L$), or only the coarsest scale DLM-S$_L$. Due to the multi-scale nature of both MS-ESSIM and the information-theoretic features in VIF and ST-RRED, the single-scale DLM offers a good lightweight feature that does not compromise on accuracy.

D. No-Reference Features

While full-reference (FR) quality models like the ones discussed so far compute local measures of similarity or distortion between reference and test frames, no-reference (NR) quality models compute measures of "naturalness" or quality directly from test frames. Other than recent deep-learning methods [54] [55], the two classical approaches to NR quality assessment involve natural scene statistics (NSS) models and artifact measurement.

NSS-based NR quality models typically involve constructing multi-scale multi-orientation band-pass decompositions and building statistical models of the transform coefficients. While such models have proven to be powerful predictors of quality [18] [34], they often face difficulty in generalizing across databases containing highly diverse user-generated content (UGC).

Artifact-measurement methods such as TLVQM [56] and the method in [57] attempt to characterize the spatial and temporal properties of video distortions, such as motion, blur, sharpness, blockiness, etc. Based on the success of the TI feature in VMAF and its derivatives, as well as the usefulness of the SI feature in Enh-VMAF, we posit that such features may be useful despite not being based on perceptual/statistical principles.

Keeping our desire for low complexity in mind, we construct two features based on TLVQM - the spatial activity index (TL-SAI) and blurriness (TL-Blur). In TLVQM, both features were computed using Sobel filters [58], but we compute these features using the horizontal and vertical high-pass filters and subbands as gradient functions and responses respectively.

The gradient energy is computed as

$$E_\lambda^{(1)} = H_\lambda^2 + V_\lambda^2 \tag{52}$$

and the second-order gradient energy is computed as

$$E_\lambda^{(2)} = (f_{HP} \circledast_H H_\lambda)^2 + (f_{HP} \circledast_V V_\lambda)^2, \qquad (53)$$

where f_{HP} denotes the wavelet high-pass filter and $\circledast_H$ and $\circledast_V$ denote convolution in the horizontal and vertical directions, respectively.

Using these quantities, the two features are computed as

$$\text{TL-SAI}_\lambda = \text{std}\left(\sqrt{E_\lambda^{(1)}}\right)^{1/4} \qquad (54)$$

and

$$\text{TL-Blur}_\lambda = \frac{\text{mean}\left(\text{perc}\left(E_\lambda^{(2)}, 1\right)\right)}{\text{mean}\left(\text{perc}\left(E_\lambda^{(1)}, 1\right)\right)}, \qquad (55)$$

where $\text{perc}(\cdot, 1)$ denotes the set of the largest 1% of values.

NR quality models tend to achieve lower accuracy in comparison to FR quality models since they lack useful information regarding the pristine source content. Since FUNQUE+ operates in the FR regime, we incorporate source information into the artifact features by computing them on both the reference and the test frames and using their difference as features for quality prediction.

$$\Delta\text{TL-SAI}_\lambda = \text{TL-SAI}_{x,\lambda} - \text{TL-SAI}_{y,\lambda} \qquad (56)$$

$$\Delta\text{TL-Blur}_\lambda = \text{TL-Blur}_{x,\lambda} - \text{TL-Blur}_{y,\lambda} \qquad (57)$$

E. Other Features

In addition to the aforementioned features, we include low-complexity features based on the differences of subband coefficients. As in FUNQUE, we consider an analog of the "Motion" feature of VMAF that we compute as the mean absolute difference (MAD) between approximation subbands of successive frames from the reference video. We denote this feature "MAD-Ref." Similarly, we define "MAD-Dis," which is computed similarly using frames from the distorted video. In addition, we include a simple feature "MAD," which is the difference between the approximation subbands of frames from the reference and distorted videos.

In addition to the MAD features, we include the "blur" and "edge" features from [28], which are computed as

$$\text{Blur}_\lambda = \frac{1}{MN} \sum_{i,j} \left(\sum_{\theta \in \{H,V,D\}} |X_{\lambda,\theta}(i,j)| - |Y_{\lambda,\theta}(i,j)| \right)^+ \qquad (58)$$

and

$$\text{Edge}_\lambda = \frac{1}{MN} \sum_{i,j} \left(\sum_{\theta \in \{H,V,D\}} |Y_{\lambda,\theta}(i,j)| - |X_{\lambda,\theta}(i,j)| \right)^+. \qquad (59)$$

TABLE II
DATABASES USED FOR MODEL SELECTION AND EVALUATION

Database	Size	Codec(s)	Scaling?	Bitdepth
BVI-HD	192	HM	No	8
CC-HD	108	HM, AV1, VTM	No	10
CC-HDDO	90	HM, AV1	Yes	10
IVP	100	Dirac, JM, MPEG-2	No	8
MCL-V	96	x264	Yes	8
NFLX-P	70	x264	Yes	8
SHVC	64	HM	No	8, 10
VQEG	72	MPEG-2, JM	No	8

VI. EXPERIMENTS

A. Databases Used for Subjective Validation

To perform model selection and to conduct a thorough evaluation of the FUNQUE+ models, we use the following set of eight databases consisting of full HD videos subjected to scaling and compression distortions.

1) BVI High Definition Database (BVI-HD) [59] - The BVI-HD database consists of 32 1080p reference videos that have been compressed using the HEVC [60] codec at various QP values to yield 192 test videos. The database also consists of a "texture synthesis" sub-database, which is not used here due to our focus on streaming artifacts.

2) BVI Codec Comparison Databases (BVI-CC) [61] - BVI-CC is a codec-comparison database that evaluated the HM [62], AV1 [4], and VTM [63] codecs under three test conditions. Of the three, we use two sub-databases in this work - CC-HD [61] and CC-HDDO [64], both of which contain nine 10-bit reference videos.
The CC-HD database consists of 108 test videos generated using the three codecs at the native 1080p resolution. The CC-HDDO database consists of 90 videos that were generated using the HM and AV1 codecs, and it includes scaling artifacts in addition to compression. The various scale factors used for each video were obtained using the Dynamic Optimizer [5].

3) IVP Subjective Quality Video Database (IVP) [65] - IVP consists of ten 1920×1088 source contents that have been compressed using MPEG2 [66], Dirac wavelet [67], and JVET JM H.264 [68] encoders, yielding 100 test videos.

4) MCL Video Quality Database (MCL-V) [69] - MCL-V consists of twelve 1080p source contents distorted using the x264 [70] encoder at four compression levels each. Scaling is introduced by encoding the videos at 720p in addition to the native 1080p resolution, yielding a total of 96 test videos.

5) Netflix Public Database (NFLX-P) [7] - NFLX-P consists of nine 1080p source contents distorted using the x264 encoder at ten combinations of bitrate and encoding resolution. However, not all combinations are applied to each video, and the database consists of 70 test sequences in total.

6) SHVC High Definition Database (SHVC) [71] - SHVC contains 1080p videos developed by the Motion Picture

Experts Group (MPEG) to evaluate the performance of Scalable HEVC encoding (SHVC). The database consists of 64 test videos after removing videos that overlap with other databases and includes both 8-bit and 10-bit content.

7) VQEG High Definition Database 3 (VQEG) [72] - VQEG consists of nine 1080p source contents distorted using the MPEG-2 [66] and JM [68] encoders to yield 72 test videos.

The test databases used to develop and validate the FUNQUE+ models cover a wide range of scenarios, using a variety of encoders, including both 8-bit and 10-bit content, and also modeling the effect of spatial scaling on quality. A summary of the salient features of the eight databases is given in Table II.

B. Evaluation Methodology

Cross-database generalization is a key property that is required of practical learning-based quality models. While models that have been tuned on individual databases are useful in scenarios where a representative database is available, general-purpose quality models must demonstrate robust cross-database generalization. Indeed, part of the success of VMAF may be attributed to its extensive usage off-the-shelf.

The FUNQUE model was evaluated in [12] by training on the CC-HDDO database and testing on the other seven databases. CC-HDDO is the largest database that was created using more than one codec, and it includes both compression and scaling distortions. However, that evaluation framework could be interpreted as biased towards the CC-HDDO database, since that is what the feature selection and training were performed using. Therefore, to conduct a more thorough evaluation, we applied a comprehensive cross-database evaluation, where every model was trained on each database and then tested on the other seven, leading to a total of 56 train-test pairs.

We use the Spearman Rank Order Correlation Coefficient (SROCC) as the performance metric and Fisher averaging [27] to compute the average over the set of train-test pairs. The Pearson Correlation Coefficient (PCC) is another metric that may be used as the performance objective. Since fusion-based models are trained to predict subjective scores, we observe that these models achieve similar SROCC and PCC values. Moreover, using SROCC allows easy comparison with learning-free ("atomic") models such as SSIM, without resorting to training and testing logistic mapping functions [33] across databases.

Therefore, given a set of databases D and a feature set/model F to be evaluated, the cross-database SROCC is computed as

$$\text{CrossDB-SROCC}(F, D) = \text{FMean}\left(\{\text{SROCC}(F, D_i, D_j)\}\right), \tag{60}$$

where $\text{SROCC}(F, D_i, D_j)$ denotes the SROCC achieved by training the model on database D_i and testing on database D_j, and the Fisher average of a set of SROCC values is given by

$$\text{FMean}\left(\{SROCC_i\}\right) = \frac{e^{2E[Z_i]} - 1}{e^{2E[Z_i]} + 1}, \tag{61}$$

TABLE III
RUNTIME COMPLEXITIES OF FEATURE SELECTION METHODS

Method	Runtime Complexity
Exhaustive Feature Search	$\mathcal{O}\left(2^{NK}\right)$
Constrained Exhaustive Feature Search	$\mathcal{O}\left(N^K\right)$
Greedy Feature Search	$\mathcal{O}\left(K^2 N^2\right)$
Constrained Greedy Feature Search	$\mathcal{O}\left(K^2 N\right)$

where

$$Z_i = \frac{1}{2} \log\left(\frac{1 + SROCC_i}{1 - SROCC_i}\right) \tag{62}$$

is Fisher's z-transformation of correlation coefficients.

C. Constrained Greedy Feature Selection

Algorithm 1 Constrained Greedy Feature Selection

Input:
    $S = \{B_1, \ldots, B_K\}$ - Feature "buckets."
    $D = \{D_1, \ldots, D_T\}$ - Databases for cross-database testing.
Output:
    F_{greedy} - The greedy-selected feature set.

$F_{greedy} \leftarrow \phi$
$S_{avail} \leftarrow S$
$srocc_{best} \leftarrow -1$
while $S_{avail} \neq \phi$ **do**
    $F_{best} \leftarrow F_{greedy}$
    for $B \in S_{\text{avail}}$ **do**
        for $f \in B$ **do**
            $F_{cand} \leftarrow F_{greedy} \cup \{f\}$
            if $\text{CrossDB-SROCC}\left(F_{cand}, D\right) > srocc_{best}$ **then**
                $srocc_{best} \leftarrow \text{CrossDB-SROCC}\left(F_{cand}, D\right)$
                $F_{best} \leftarrow F_{cand}$
                $B_{chosen} \leftarrow B$
            end if
        end for
    end for
    if $F_{greedy} \neq F_{best}$ **then**
        $F_{greedy} \leftarrow F_{best}$
        $S_{avail} \leftarrow S_{avail} \setminus \{B_{chosen}\}$
    else
        break
    end if
end while

As described in the preceding sections, we considered a wide range of CSF models, numbers of wavelet levels $L = 1, \ldots, 4$, and the application of SAST. Since CSF sharing provided a significant boost in performance in [12], we included the CSF in the unified transform to share it across all the atom features.

Due to the wide range of design choices available to us, it is required to perform both model selection, i.e., choosing a CSF, number of levels, and the application of SAST, and feature selection for each model. As identified during the development of FUNQUE [12], the use of exhaustive feature search (EFS) is not feasible due to the large number of features

TABLE IV
FEATURE BUCKETS USED FOR FEATURE SELECTION

SSIM	Info	DLM	Sharpness	MAD
$SSIM_L$, $ESSIM_L$, $MS\text{-}SSIM_L$, $MS\text{-}ESSIM_L$	$VIF\text{-}HV_L$, $(VIF\text{-}A_\lambda; \lambda \le L)$, $STRRED\text{-}HV_L$, $(SRRED\text{-}HV_L, TRRED\text{-}HV_L)$, $(STRRED\text{-}A_\lambda; \lambda \le L)$, $(SRRED\text{-}A_\lambda, TRRED\text{-}A_\lambda; \lambda \le L)$	$DLM\text{-}S_L$, $(DLM\text{-}S_\lambda; \lambda \le L)$	$Blur_L$, $Edge_L$, $(Blur_L, Edge_L)$, $\Delta TL\text{-}SAI_L$, $\Delta TL\text{-}Blur_L$	$MAD\text{-}Ref_L$, $MAD\text{-}Dis_L$, MAD_L

being considered. This matter was addressed by adopting a Constrained Exhaustive Feature Selection (CEFS) approach.

Under this approach, the set of features under consideration was partitioned into four "buckets," and EFS was used under the constraint that at most one feature may be selected from each bucket. This constraint introduces an inductive bias that identifies robust features of each "type" and automatically constrains model size, which improves generalization. For the set of features considered by FUNQUE, this was reported to have reduced the search space by a factor of seven.

However, due to the significantly larger pool of features being considered for FUNQUE+, even CEFS proves to be computationally intractable. So, we instead used a Constrained Greedy Feature Selection (CGFS) algorithm that is significantly more scalable than EFS. A similar greedy feature selection (GFS) method was used in [28], but it does not include the feature constraints that were found to be crucial in FUNQUE.

A detailed description of the CGFS algorithm is provided in Algorithm 1. In short, at each step of CGFS, the feature that achieves the greatest increase in cross-database SROCC is added to the selected feature set, while enforcing the bucket constraints. Considering the case of K buckets having N features each, the runtime complexities of the four feature selection algorithms are shown in Table III. Note that CGFS retains the inductive bias introduced by CEFS due to the continued use of feature buckets, while scaling more gracefully on larger candidate feature sets.

We considered five "feature buckets" in the feature selection process, grouping features by "type." The five types of features are "SSIM," "Info," which denotes information-theoretic quality models, "DLM," "Sharpness," which includes all sharpness and blur-related features, and "MAD." The feature buckets are described in detail in Table IV. Note that entries in Table IV that are in the form of tuples are groups of features that must be included or excluded collectively.

D. Chroma-aware Modeling

A key weakness of SOTA quality models such as MS-SSIM and VMAF is that they operate only on the luma channel. Improvements to VMAF like ST-VMAF and Ens-VMAF also rely on extracting better features from the luma channel. However, the superior performance of chromatic SSIM [33] and Enh-VMAF motivate the use of features extracted from the Cb and Cr chroma channels. The tradeoff, however, is that building such "three-channel" (3C) quality models requires the processing of all three color channels, which increases the computational complexity.

Therefore, in addition to a luma-only Y-FUNQUE+ model, we have also developed a 3C-FUNQUE+ model that utilizes features extracted from all three color channels Y, Cb, and Cr. To perform feature selection for the 3C-FUNQUE+ model, we created three copies of each feature bucket described in Section VI-C, corresponding to the three channels. This yielded a total of fifteen feature buckets to be used for feature selection.

VII. RESULTS

A. The FUNQUE+ Models

In this section, we report and analyze the performance of the FUNQUE+ models, and compare them against SOTA baseline models. The luma-only (Y) and three-channel (3C) FUNQUE+ models identified using CGFS are described in Table V. In addition, we also describe "Full-Scale" models that do not utilize SAST. As a result, these models are less perceptually accurate but are better suited to measure pixel fidelity.

Furthermore, both Y-FUNQUE+ and 3C-FUNQUE+ utilize subband-weighting CSFs, which are much more computationally efficient than the 21-tap spatial filter used in FUNQUE [12]. Therefore, overall, both the models are computationally very efficient. This aspect of the FUNQUE+ models is analyzed in further detail in Section VII.

The Full-Scale models are less efficient, both because they are applied at the native 1080p resolution, and since FS-Y-FUNQUE+ uses a spatial filter CSF (albeit only 5-tap). In addition, the two FS models also utilize a larger set of features than the models constructed using SAST.

TABLE V
FUNQUE+ MODELS

Model	Features	CSF	SAST Used?
Y-FUNQUE+	$MS\text{-}ESSIM_2 + MAD\text{-}Ref_2 + DLM\text{-}S_2$	NadenauSW	Yes
3C-FUNQUE+	$Y\text{-}MS\text{-}ESSIM_2 + Y\text{-}MAD\text{-}Dis_2 + Y\text{-}DLM\text{-}S_2 + Y\text{-}SRRED\text{-}HV_2 + Y\text{-}TRRED\text{-}HV_2 + Cb\text{-}Edge_2 + Cr\text{-}MAD_2$	LiSW	Yes
FS-Y-FUNQUE+	$MS\text{-}ESSIM_2 + \Delta TL\text{-}SAI_2 + MAD\text{-}Dis_2 + DLM\text{-}S_2 + STRRED\text{-}HV_2$	NadenauSpat	No
FS-3C-FUNQUE+	$Y\text{-}MS\text{-}ESSIM_3 + \Delta Y\text{-}TL\text{-}SAI_3 + Y\text{-}DLM\text{-}S_3 + Cb\text{-}MAD\text{-}Dis_3 + Cb\text{-}SRRED\text{-}HV_3 + Cb\text{-}TRRED\text{-}HV_3 + Cb\text{-}Edge_3 + Cr\text{-}MAD_3 + Cr\text{-}Blur_3$	WatsonSW	No

TABLE VI
CROSS-DATABASE SROCC ACHIEVED BY Y-FUNQUE+

Database	BVI-HD	CC-HD	CC-HDDO	IVP	MCL-V	NFLX-P	SHVC	VQEG	Average
BVI-HD	-	0.8356	0.8774	0.9182	0.7486	0.9408	0.9011	0.8912	0.8844
CC-HD	0.7990	-	0.8776	0.9073	0.7942	0.9233	0.9172	0.8473	0.8752
CC-HDDO	0.8038	0.8606	-	0.9212	0.7617	0.9205	0.9150	0.8806	0.8777
IVP	0.7982	0.8258	0.8745	-	0.7837	0.9412	0.8916	0.8861	0.8674
MCL-V	0.7420	0.8664	0.8559	0.8917	-	0.8984	0.8941	0.8038	0.8579
NFLX-P	0.7598	0.7363	0.7927	0.9113	0.7614	-	0.8412	0.8855	0.8237
SHVC	0.7938	0.8828	0.8879	0.9047	0.7617	0.9334	-	0.8884	0.8750
VQEG	0.7909	0.7648	0.8375	0.9126	0.7415	0.9541	0.8742	-	0.8586
Average	0.7849	0.8307	0.8603	0.9100	0.7652	0.9327	0.8930	0.8717	**0.8660**

TABLE VII
CROSS-DATABASE SROCC ACHIEVED BY 3C-FUNQUE+

Database	BVI-HD	CC-HD	CC-HDDO	IVP	MCL-V	NFLX-P	SHVC	VQEG	Average
BVI-HD	-	0.8834	0.9064	0.9129	0.7754	0.9398	0.8833	0.8896	0.8920
CC-HD	0.7977	-	0.9167	0.8838	0.8348	0.9249	0.8834	0.8762	0.8799
CC-HDDO	0.7961	0.8846	-	0.9078	0.7676	0.9269	0.8861	0.8819	0.8736
IVP	0.7983	0.7901	0.8648	-	0.7902	0.9472	0.8458	0.8812	0.8571
MCL-V	0.7721	0.9216	0.9112	0.8748	-	0.9183	0.8672	0.8757	0.8844
NFLX-P	0.7759	0.8127	0.8786	0.9064	0.7874	-	0.8209	0.8878	0.8455
SHVC	0.7793	0.9331	0.9345	0.8877	0.7928	0.9204	-	0.8820	0.8879
VQEG	0.7893	0.8717	0.9371	0.8721	0.8056	0.9317	0.8556	-	0.8766
Average	0.7872	0.8795	0.9101	0.8933	0.7943	0.9306	0.8648	0.8822	**0.8754**

TABLE VIII
COMPARISON OF AVERAGE TEST ACCURACY (SROCC) OF FUNQUE+ MODELS AGAINST THE SOTA BASELINE MODELS.
* DENOTES DEEP LEARNING MODELS.

Model	BVI-HD	CC-HD	CC-HDDO	IVP	MCL-V	NFLX	SHVC	VQEG	Average
SSIM	0.5986	0.7179	0.8028	0.7145	0.3971	0.7014	0.5624	0.7357	0.6685
LPIPS*	0.6683	0.6663	0.7335	0.6240	0.6319	0.8277	0.4216	0.8212	0.6921
PSNR	0.6125	0.6145	0.7497	0.8169	0.4630	0.7476	0.7542	0.7516	0.7024
FSIM	0.6860	0.6671	0.7183	0.6945	0.6426	0.8566	0.5842	0.7397	0.7082
ST-VMAF	0.7213	0.7748	0.8011	0.5730	0.7048	0.8536	0.8088	0.7657	0.7603
MS-SSIM	0.7849	0.6909	0.7989	0.8900	0.6781	0.8254	0.8567	0.6403	0.7849
Ens-VMAF	0.7185	0.7601	0.8003	0.7458	0.7881	0.9008	0.7773	0.8166	0.7956
DISTS*	0.7686	0.6695	0.7202	0.7930	0.7703	0.8750	0.7979	0.8943	0.7971
DeepWSD*	0.7979	0.7242	0.7413	0.8756	0.7528	0.8397	0.8304	0.8451	0.8070
VMAF	0.7793	0.8283	0.8633	0.8352	0.7726	0.8924	0.8192	0.8291	0.8312
Enh-VMAF	0.7644	0.7889	0.8634	0.8703	0.7750	0.9100	0.8383	0.8386	0.8377
FUNQUE	0.7624	0.7932	0.8041	0.8895	0.7210	0.9206	0.8608	0.8818	0.8409
MS-ESSIM	0.7693	0.7565	0.8482	0.9026	0.6485	0.9110	0.8874	0.8904	0.8441
FS-Y-FUNQUE+	0.7675	0.8220	0.8619	0.8892	0.7159	0.9110	0.8891	0.8499	0.8484
Y-FUNQUE+	0.7849	0.8307	0.8603	0.9100	0.7652	**0.9327**	**0.8930**	0.8717	0.8660
FS-3C-FUNQUE+	**0.7996**	0.8501	**0.9155**	**0.9125**	0.7332	0.9229	0.8602	**0.8836**	0.8707
3C-FUNQUE+	0.7872	**0.8795**	0.9101	0.8933	**0.7943**	0.9306	0.8648	0.8822	**0.8754**

A closer look at the selected feature sets reveals that all four models utilize MS-ESSIM, which improves upon the ESSIM feature used by FUNQUE. As reported in Table VIII, MS-ESSIM achieves high prediction accuracy even as a standalone quality model. Therefore, the development of MS-ESSIM is one of the main contributions of this work. The high accuracy of MS-ESSIM is attributed to the use of SAST and CoV pooling, similar to ESSIM, as discussed in [33]. It may also be observed that three of the four models utilize other features drawn from SRRED, TRRED, and ΔTL-SAI. In addition, the two 3C models utilize the blur and edge features proposed in [28]. Together, these observations demonstrate the impact of the expanded feature set that was developed to create FUNQUE+.

B. Cross-Database Results

We tabulate the detailed cross-database testing results of the Y- and 3C-FUNQUE+ models are shown in Tables VI and VII. Each row refers to one choice of the training database, and each column refers to a choice of test database. Table VIII provides a comparison of the average test SROCC achieved by the FUNQUE+ models and the baseline models against which FUNQUE+ is compared. For brevity, we condensed the cross-database SROCC tables into one row for each model, where each column presents the average test accuracy on that database over all choices of training databases, and the "Average" column presents the overall average.

From this table, it may be observed that all of the FUNQUE+ models outperformed all of the baseline models, including VMAF and FUNQUE, as well as more computationally complex models including ST-VMAF and Enh-VMAF.

Furthermore, by comparing the performance of the Y-only and 3C models, it may be observed that incorporating the chroma features further improved accuracy. Finally, the FS versions of both models achieved lower SROCC than the other FUNQUE+ models, which demonstrates the tradeoff between modeling viewing conditions and making the model robust to hacking. However, the FS models still significantly outperform the prior SOTA models.

In addition, one may make two observations regarding the accuracies of baseline models. Firstly, SSIM achieves a lower overall accuracy than all other metrics, even PSNR. While this may appear anomalous, similar observations were made in [28]. Moreover, the higher accuracy of MS-SSIM indicates that the single-scale nature of SSIM is responsible for its lower accuracy. Secondly, all deep methods achieve lower accuracy than VMAF, with LPIPS even dipping below PSNR. We attribute this to the fact that these methods were not designed for the assessment of compression quality. Rather, they often focus on aspects such as geometric and texture distortions.

C. Ablation Analysis

To understand the performance of FUNQUE+, we conducted ablation experiments by identifying three key design elements that characterize the suite of FUNQUE+ models

1) The incorporation of perceptual sensitivity using a CSF model.
2) The use of SAST to account for viewing conditions.
3) The addition of chroma features.

It is important to note that the inclusion of chroma features triples the size of the candidate feature set. Therefore, the development of "three-channel" models is contingent upon the use of the scalable CGFS feature selection method described in Section VI-C. As a result, the improvement in accuracy by including chroma features also reflects the usefulness of CGFS. To quantify the impact of each of the three factors, we evaluate 8 models that correspond to whether a CSF was used, SAST was used, and whether chroma features were included.

The cross-database accuracy achieved by the eight models is presented in Table IX. From the Table, it may be observed that the addition of each of the three factors improves the accuracy of the quality model in all cases. While including the CSF and chroma features comes at additional computational cost, using SAST improves both accuracy and efficiency, due to downscaling by a factor of 2. In this way, FUNQUE+ attains superior accuracy through the use of CSF, SAST, and chroma features.

TABLE IX
AVERAGE TEST SROCC ACHIEVED BY FUNQUE+ VARIANTS

CSF Used?	SAST Used?	Chroma Included?	Test SROCC
No	No	No	0.8441
No	No	Yes	0.8527
No	Yes	No	0.8570
No	Yes	Yes	0.8640
Yes	No	No	0.8484
Yes	No	Yes	0.8707
Yes	Yes	No	0.8660
Yes	Yes	Yes	0.8754

D. Computational Optimization and Analysis

The FUNQUE+ models are made quite efficient by sharing computation by calculating all features from a common wavelet decomposition, by restricting model size during the feature selection process, by using lightweight models of the CSF, and by optimizing the implementations of the atom features for efficiency. Note that the optimizations discussed in this section were applied only to the FUNQUE and FUNQUE+ models, not any baseline models that were used for evaluation.

In particular, we used integral images [33], [73] to remove computationally expensive filtering operations from all features. In this way, all local statistics in the VIF and STRRED algorithms were computed using uniform rectangular windows instead of the usual Gaussian windows.

Filtering using rectangular kernels was eliminated by computing local sums using integral images as follows. Let $X(i,j)$ denote an image whose local sums over $k \times k$ windows are to be calculated. First, construct the integral image

$$I(i,j) = \begin{cases} \sum_{m \leq i} \sum_{n \leq j} X(m,n) & i,j > 0 \\ 0 & \text{otherwise} \end{cases}. \quad (63)$$

The sum over any $k \times k$ window with top left corner at (i,j) may be computed as

$$S_k(i,j) = I(i+k-1, j+k-1) + I(i-1, j-1) \\ - I(i+k-1, j-1) - I(i-1, j+k-1). \quad (64)$$

The sum of these optimizations, coupled with the use of SAST by Y-FUNQUE+ and 3C-FUNQUE+, provides significant boosts in the efficiency of the FUNQUE+ models as compared to other SOTA quality models. To demonstrate this, we have analyzed all the models in Table VIII to obtain estimates of their computational complexity, which we describe in three ways.

First, we describe the asymptotic computational complexity of the algorithms, expressed in $\mathcal{O}(\cdot)$ ("Big O") notation, expressed in terms of relevant design parameters. In addition, we conducted a thorough analysis of their implementations to estimate the number of Giga Floating Point Operations (GLOPs) required to run feature extraction for each model on 150 frames of a Full HD video, corresponding to a typical six-second clip played at 25 fps.

A detailed analysis of the "Big O" computational complexity and the number of Giga Floating Point Operations (GFLOPs) is provided in Table X. In addition, we measured the run time of all the compared models on an Intel Core i7-8700 CPU having a clock frequency of 3.2GHz. All of the models were implemented in Python for a fair comparison, and the tradeoff between run time and model accuracy has been visualized in Fig. 2. From these results, it may be seen that the proposed FUNQUE+ (and FUNQUE) models offer the highest average test SROCC, while also being the most efficient fusion-based models. Note that the run time on the horizontal axis is shown in a log scale to accommodate the wide range of run times observed during the experiment.

TABLE X
ANALYZING THE COMPUTATIONAL COMPLEXITY OF FUNQUE+ AGAINST THE BASELINE MODELS.
GFLOPs CORRESPONDING TO COMPUTING FEATURES ON 150 FRAMES OF A 1920x1080 VIDEO

Model	Asymptotic Computational Complexity	GFLOPs	Time (s)
PSNR	$\mathcal{O}(NT)$	0.933	6.46
Y-FUNQUE+	$\mathcal{O}\left(k_W\left(1-2^{-P_W}\right)NT/D^2\right)$	2.245	7.68
FUNQUE [12]	$\mathcal{O}\left(\left(k_W\left(1-2^{-P_W}\right)+k_{CSF}^2\right)NT/D^2\right);\ k_{CSF}$: CSF filter size	16.908	13.79
3C-FUNQUE+	$\mathcal{O}\left(k_W\left(1-2^{-P_W}\right)NT/D^2\right)$	16.913	21.96
SSIM [6]	$\mathcal{O}(k_G NT)$	75.894	29.38
MS-SSIM [9]	$\mathcal{O}\left(k_G(1-2^{-P_G})NT\right)$	104.398	43.38
FS-Y-FUNQUE+	$\mathcal{O}\left(\left(k_W\left(1-2^{-P_W}\right)+k_{CSF}^2\right)NT\right);\ k_{CSF}$: CSF filter size	79.393	61.37
FS-3C-FUNQUE+	$\mathcal{O}\left(k_W\left(1-2^{-P_W}\right)NT\right)$	86.795	79.75
VMAF [7]	$\mathcal{O}\left(\left(k_G\left(1-2^{-P_G}\right)+\left(k_W+k_C^2\right)\left(1-2^{-P_W}\right)\right)NT\right)$	201.804	84.40
Ens-VMAF [27]	$\mathcal{O}\left(\left(\left(k_G+B^4\right)\left(1-2^{-P_G}\right)+\left(k_W+k_C^2\right)\left(1-2^{-P_W}\right)\right)NT\right);\ B$: SpEED block size	514.018	561.52
ST-VMAF [27]	$\mathcal{O}\left(\left(\left(k_G+B^4\right)\left(1-2^{-P_G}\right)+\left(k_W+k_C^2\right)\left(1-2^{-P_W}\right)\right)NT\right);\ B$: SpEED block size	483.590	612.59
Enh-VMAF [28]	$\mathcal{O}\left(\left(k_G\left(1-2^{-P_G}\right)+\left(k_W+k_C^2\right)\left(1-2^{-P_W}\right)+k_O^2 W\left(1-2^{-P_O}\right)\right)NT\right);$ k_O: Optical flow search size, P_O: Optical flow pyramid height, W: Warps per level	324.896	737.77
FSIM [13]	$\mathcal{O}\left(\left(\log(N)+P_{LG}O+k_{Grad}^2\right)NT\right);\ P_{LG}$: Height of log-Gabor pyramid, O: Number of orientations, k_{Grad}: Gradient filter size	429.484	1047.35
LPIPS [38]	$\mathcal{O}\left(\left(\sum_{i=1}^{5}k_i^2 C_i C_{i-1}/s_i^2\right)NT\right);\ k_i$: Convolution kernel sizes, C_i: Convolution channels, s_i: Sub-sampling factors due to striding and max-pooling.	18205.206	389.56
DISTS [43]	$\mathcal{O}\left(\left(\sum_{i=1}^{5}L_i k_i^2 C_i^2/s_i^2+\sum_{i=1}^{4}k_H^2 C_i/s_i^2\right)NT\right);\ L_i$: Layers in block i, k_i: Convolution kernel sizes, C_i: Convolution channels, s_i: Sub-sampling factors due to striding and max-pooling, k_H: Hanning window size.	382213.417	484.48
DeepWSD [44]	$\mathcal{O}\left(\left(k_R^2+\sum_{i=1}^{5}L_i k_i^2 C_i^2/s_i^2+\sum_{i=1}^{4}\log\left(W^2\right)C_i/s_i^2\right)NT/s_R^2\right);\ k_R$: Resize kernel size, s_R: Resize factor (8), L_i: Layers in block i, k_i: Convolution kernel sizes, C_i: Convolution channels, s_i: Sub-sampling factors due to striding and max-pooling, k_H: Hanning window size, W: Wasserstein distance block size (16).	12014.977	1589.46

* Note: Wherever applicable, N denotes the number of pixels per frame, T denotes the number of frames in the video,
D denotes the SAST downscaling factor, k_G, k_W and k_C denote the size of Gaussian, wavelet, and contrast-masking filters,
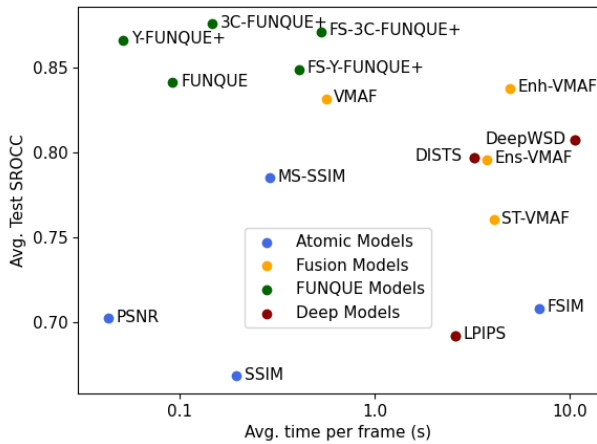and P_G and P_W denote the heights of the Gaussian and wavelet pyramids.

Fig. 2. SROCC vs. Run Time

distortion optimization, a quality model must satisfy two basic monotonicity properties related to compression. Let $Q(r,c)$ denote the prediction of a video quality model for a video compressed at resolution r and a given compression factor c (e.g., CRF in libx264). Then, the model is monotonic if whenever $r_1 \leq r_2$, then $Q(r_1,c) \leq Q(r_2,c)$, and whenever $c_1 \leq c_2$, where lower c denotes less compression, then $Q(r,c_1) \geq Q(r,c_2)$.

To analyze the monotonicity of the FUNQUE+ models, we used the nine 8-bit 1080p source contents from the NFLX-P database. Each source video was encoded at six encoding resolutions - 1080p, 720p, 480p, 360p, 240p, and 144p and eleven CRF values equally spaced in the range 20-40 using the x264 encoder, yielding a total of 792 test videos. Quality predictions were made using VMAF v0.6.1 and the two FUNQUE+ models Y-FUNQUE+ and 3C-FUNQUE+, which were trained on the CC-HDDO database. The CC-HDDO database was chosen for this analysis for the reasons described in Section VI-B, and also because it does not include the x264 encoder.

The monotonicity of the predictions made by all three models was analyzed both in terms of encoding resolution and CRF. It was observed that for three of the nine source
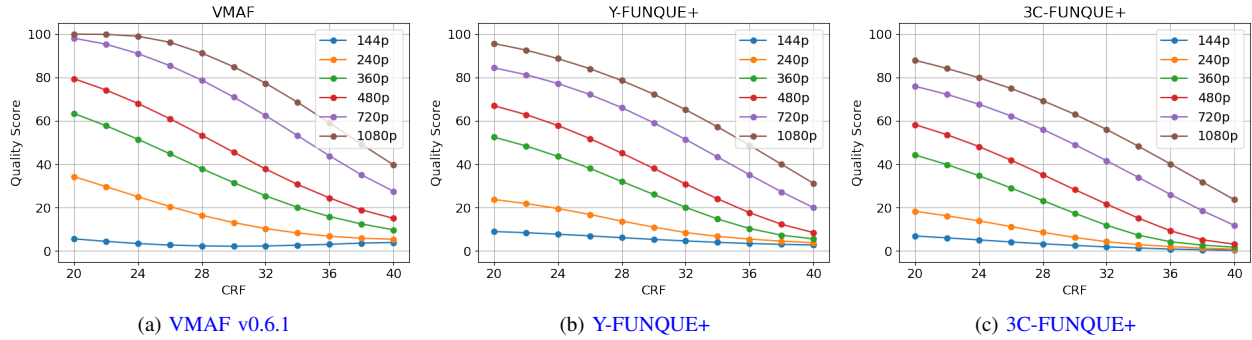
E. Monotonicity Analysis

Finally, we evaluate the monotonic behavior of the proposed FUNQUE+ models. To be suitable for use in perceptual rate-

(a) VMAF v0.6.1   (b) Y-FUNQUE+   (c) 3C-FUNQUE+

Fig. 3. Quality Predictions Made by VMAF, Y-FUNQUE+, and 3C-FUNQUE+ on the "Seeking" Source Content.

contents, VMAF v0.6.1 provided non-monotonic estimates of visual quality at low encoding resolutions. In other words, the quality predicted by VMAF did not decrease with an increase in CRF. An example of this behavior is illustrated in Figure 3a, which shows the variation of predicted visual quality against encoding resolution and CRF for the source video named "Seeking." Interestingly, retraining the VMAF regressor on the CC-HDDO database restored monotonicity to VMAF's predictions. Therefore, we attribute the source of non-monotonicity to the VMAF v0.6.1 regressor and training dataset.

On the other hand, as shown in Figures 3b and 3c, it was found that both FUNQUE+ models predicted perfectly monotonic variations of quality with CRF at all resolutions. This demonstrates that the FUNQUE+ models are robust even in the low quality regime, despite their efficiency.

VIII. CONCLUSION AND FUTURE WORK

We have described a class of efficient fusion-based full-reference video quality models that have been designed to have exceptional practical performance attributes when applied to large-scale video streaming. These models were based on the prototype FUNQUE framework and were developed using novel wavelet-domain features and a novel feature selection algorithm that yields small, diverse fusion-based models. Through extensive analysis and experimental validation, we have been able to show that the resulting FUNQUE+ models achieve higher accuracy than SOTA FR models with significantly reduced computational cost.

We hope that in addition to finding use in the video streaming and sharing industries, the success of the FUNQUE+ models will lead to the development of other low-complexity, high-accuracy models using similar principles. As described in Section II, VMAF has grown to find use beyond streaming, particularly in emerging video modalities. In the future, we envision FUNQUE-based models being used across domains, to perceptually optimize the delivery of 360 VR videos, high-resolution videos including 8K, and High Dynamic Range (HDR) videos.

IX. ACKNOWLEDGMENT

REFERENCES

[1] Global internet phenomena report 2023. [Online]. Available: https://www.sandvine.com/global-internet-phenomena-report-2023 [Accessed: Feb. 2023]

[2] T. Wiegand, "Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H. 264— ISO/IEC 14496-10 AVC)," *JVT-G050*, 2003.

[3] D. Mukherjee, J. Bankoski, A. Grange, J. Han, J. Koleszar, P. Wilkins, Y. Xu, and R. Bultje, "The latest open-source video codec VP9 - an overview and preliminary results," in *Picture Coding Symposium (PCS)*, 2013, pp. 390–393.

[4] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi, C. Chiang, Y. Wang, P. Wilkins, J. Bankoski, L. Trudeau, N. Egge, J. Valin, T. Davies, S. Midtskogen, A. Norkin, and P. de Rivaz, "An overview of core coding tools in the AV1 video codec," in *Picture Coding Symposium (PCS)*, 2018, pp. 41–45.

[5] I. Katsavounidis, "Dynamic optimizer - a perceptual video encoding optimization framework," *The Netflix Tech Blog*, 2018.

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

[7] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, p. 2, 2016.

[8] D. Vatolin, D. Kulikov, M. Erofeev, A. Antsiferova, E. Sklyarov, A. Gushchin, N. Safonov, and N. Alutis, "MSU video codecs comparison 2021 part 1: FullHD, objective," 09 2021.

[9] Z. Wang, E. Simoncelli, and A. 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 Vol.2.

[10] K. Lee, V. Rao, and W. Arnold, "Accelerating facebook's infrastructure with application-specific hardware." [Online]. Available: https://engineering.fb.com/2019/03/14/data-center-engineering/accelerating-infrastructure/ [Accessed: January 2023]

[11] The YouTube Team, "Reimagining video infrastructure to empower youtube." [Online]. Available: https://blog.youtube/inside-youtube/new-era-video-infrastructure/ [Accessed: January 2023]

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 2010.

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

[19] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

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

[21] A. ., J. N. Shingala, N. K. Thangudu, P. Konda, and V. G. R., "Energy efficient perceptual video quality measurement (VMAF) at scale," in *Applications of Digital Image Processing XLIII*, vol. 11510, International Society for Optics and Photonics. SPIE, 2020, p. 115100G.

[22] 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.

[23] D. Ramsook, A. Kokaram, N. O'Connor, N. Birkbeck, Y. Su, and B. Adsumilli, "A differentiable estimator of VMAF for video," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

[24] H. Amirpour, R. Schatz, and C. Timmerer, "Between two and six? towards correct estimation of JND step sizes for VMAF-based bitrate laddering," in *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*, 2022, pp. 1–4.

[25] A. Zvezdakova, S. Zvezdakov, D. Kulikov, and D. Vatolin, "Hacking VMAF with video color and contrast distortion," *ArXiv*, vol. abs/1907.04807, 2019.

[26] Z. Li, K. Swanson, C. Bampis, L. Krasula, and A. Aaron, "Toward a better quality metric for the video community," *The Netflix Tech Blog*, p. 2, 2020.

[27] 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, 2019.

[28] F. Zhang, A. Katsenou, C. Bampis, L. Krasula, Z. Li, and D. Bull, "Enhancing VMAF through new feature integration and model combination," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

[29] L.-H. Chen, C. G. Bampis, Z. Li, J. Sole, and A. C. Bovik, "Perceptual video quality prediction emphasizing chroma distortions," *IEEE Transactions on Image Processing*, vol. 30, pp. 1408–1422, 2021.

[30] M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito, and N. García, "Video multimethod assessment fusion (VMAF) on 360VR contents," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 1, pp. 22–31, 2020.

[31] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "High frame rate video quality assessment using VMAF and entropic differences," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.

[32] M. A. Usman and M. G. Martini, "On the suitability of VMAF for quality assessment of medical videos: Medical ultrasound & wireless capsule endoscopy," *Computers in Biology and Medicine*, vol. 113, p. 103383, 2019.

[33] A. K. Venkataramanan, C. Wu, A. C. Bovik, I. Katsavounidis, and Z. Shahid, "A hitchhiker's guide to structural similarity," *IEEE Access*, vol. 9, pp. 28 872–28 896, 2021.

[34] 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.

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

[36] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 694–711.

[37] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[38] 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 (CVPR)*, June 2018.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[40] M. Kettunen, E. Härkönen, and J. Lehtinen, "E-LPIPS: robust perceptual image similarity via random transformation ensembles," *CoRR*, vol. abs/1906.03973, 2019.

[41] S. Ghazanfari, S. Garg, P. Krishnamurthy, F. Khorrami, and A. Araujo, "R-LPIPS: An adversarially robust perceptual similarity metric," *arXiv preprint arXiv:2307.15157*, 2023.

[42] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 433–442.

[43] 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, 2022.

[44] 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*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 970–978.

[45] K. N. Ngan, K. S. Leong, and H. Singh, "Adaptive cosine transform coding of images in perceptual domain," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1743–1750, 1989.

[46] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and image resolution," *IEEE Transactions on Broadcasting*, vol. 61, no. 3, pp. 520–531, 2015.

[47] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *Human Vision, Visual Processing, and Digital Display III*, vol. 1666, International Society for Optics and Photonics. SPIE, 1992, pp. 2 – 15.

[48] M. J. Nadenau and J. Reichel, "Compression of color images with wavelets considering the HVS," in *Human Vision and Electronic Imaging IV*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 3644, International Society for Optics and Photonics. SPIE, 1999, pp. 129 – 140.

[49] 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.

[50] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 525–536, 1974.

[51] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1164–1175, 1997.

[52] P. Hill, A. Achim, M. E. Al-Mualla, and D. Bull, "Contrast sensitivity of the wavelet, dual tree complex wavelet, curvelet, and steerable pyramid transforms," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2739–2751, 2016.

[53] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proceedings., International Conference on Image Processing*, vol. 3. IEEE, 1995, pp. 444–447.

[54] J. C. Mier, E. Huang, H. Talebi, F. Yang, and P. Milanfar, "Deep perceptual image quality assessment for compression," *CoRR*, vol. abs/2103.01114, 2021.

[55] Z. Ying, M. Mandal, D. Ghadiyaram, and A. C. Bovik, "Patch-VQ: 'patching up' the video quality problem," *CoRR*, vol. abs/2011.13544, 2020.

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

[57] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 4, pp. 533–546, 2015.

[58] N. Kanopoulos, N. Vasanthavada, and R. Baker, "Design of an image edge detection filter using the sobel operator," *IEEE Journal of Solid-State Circuits*, vol. 23, no. 2, pp. 358–367, 1988.

[59] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2620–2630, 2018.

[60] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[61] F. Zhang, A. V. Katsenou, M. Afonso, G. Dimitrov, and D. R. Bull, "Comparing VVC, HEVC and AV1 using objective and subjective assessments," *ArXiv*, vol. abs/2003.10282, 2020.

[62] Joint Video Experts Team, "HEVC test model (HM)." [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/HM [Accessed: Dec. 2022]

[63] J. Chen, Y. Ye, and S. Kim, "Algorithm description for versatile video coding and test model 11 (VTM 11)," 2021.

[64] A. V. Katsenou, F. Zhang, M. Afonso, and D. R. Bull, "A subjective comparison of AV1 and HEVC for adaptive video streaming," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 4145–4149.

[65] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan. (2009) IVP subjective quality video database. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/ [Accessed: Dec. 2022]

[66] International Telecommunication Union, "Generic coding of moving pictures and associated audio information-part 2: Video," *Int. Standards Org./Int. Electrotech. Comm.(ISO/IEC) JTC 1, Rec. H. 262 and ISO/IEC 13 818-2 (MPEG-2 Video)*, 1994.

[67] T. Davies, "The dirac algorithm," 2008. [Online]. Available: https://dirac.sourceforge.net/documentation/algorithm/algorithm/index.htm [Accessed: Dec. 2022]

[68] K. Suehring. [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/JM [Accessed: Dec. 2022]

[69] J. Y. Lin, R. Song, C.-H. Wu, T. J. 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.

[70] VideoLAN, "x264." [Online]. Available: https://code.videolan.org/videolan/x264.git [Accessed: Dec. 2022]

[71] Y. He, Y. Ye, F. Hendry, Y. K. Wang, and V. Baroncini, "SHVC verification test results," *JCT-VC Meeting*, no. JCTVC-W0095, 2016.

[72] (2010) Report on the validation of video quality models for high definition video content. Video Quality Experts Group.

[73] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.