Domain-Specific Fusion Of Objective Video Quality Metrics

Aaron Chadha
aaron@isize.co
iSIZE
UK

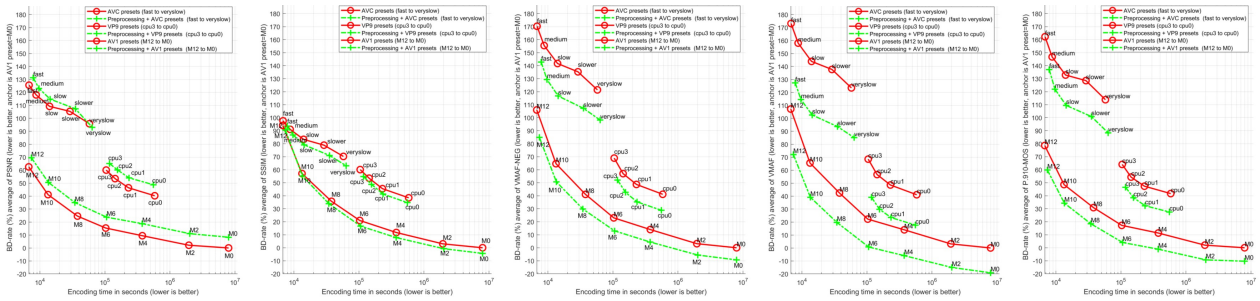
Ioannis Katsavounidis
ikatsavounidis@fb.com
Meta Platforms, Inc.
USA

Ayan Kumar Bhunia
ayan@isize.co
iSIZE
UK

Cosmin Stejerean
cstejerean@fb.com
Meta Platforms, Inc.
USA

Mohammad Umar Khan
umar@isize.co
iSIZE
UK

Yiannis Andreopoulos
yiannis@isize.co
iSIZE
UK



(a) PSNR **(b) SSIM** **(c) VMAF-NEG** **(d) VMAF** **(e) Proposed P.910-MOS**

Figure 1: BD-rate (Bjontegaard Delta-rate) vs. runtime of video encoders when assessed in terms of: PSNR, SSIM, VMAF-NEG, VMAF and the P.910-MOS fused metric derived by our proposal. The utilized encoders (x264 AVC, vpxenc VP9, and svt-av1 AV1, with and without preprocessing) lead to different BD-rate results for each metric. Instead of ad-hoc averaging of BD-rates, we propose to consolidate this difference via domain-specific video quality metric fusion with limited subjective testing.

ABSTRACT

Video processing algorithms like video upscaling, denoising, and compression are now increasingly optimized for perceptual quality metrics instead of signal distortion. This means that they may score well for metrics like video multi-method assessment fusion (VMAF), but this may be because of metric overfitting. This imposes the need for costly subjective quality assessments that cannot scale to large datasets and large parameter explorations. We propose a methodology that fuses multiple quality metrics based on small-scale subjective testing in order to unlock their use at scale for specific application domains of interest. This is achieved by employing pseudo-random sampling of the resolution, quality range and test video content available, which is initially guided by quality metrics in order to cover the quality range useful to each application. The selected samples then undergo a subjective test, such as ITU-T P.910 absolute categorical rating, with the results of the test post-processed and used as the means to derive the best combination of multiple objective metrics using support vector regression. We showcase the benefits of this approach in two applications: video encoding with and without perceptual preprocessing, and deep video denoising & upscaling of compressed content. For both applications, the derived fusion of metrics allows for a more robust alignment to mean opinion scores than a perceptually-uninformed combination of the original metrics themselves. The dataset and code is available at https://github.com/isize-tech/VideoQualityFusion.

CCS CONCEPTS

• **Applied computing**; • **Information systems** → **Multimedia streaming**;

KEYWORDS

datasets, neural networks, quality assessment, video coding, video denoising

ACM Reference Format:
Aaron Chadha, Ioannis Katsavounidis, Ayan Kumar Bhunia, Cosmin Stejerean, Mohammad Umar Khan, and Yiannis Andreopoulos. 2022. Domain-Specific Fusion Of Objective Video Quality Metrics. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Subjective visual quality assessment via well-known methodologies such as ITU-T P.910 or BT.500 [2, 26] is considered to be the gold standard for quality assessment in video signal processing systems like video encoding, restoration, denoising, synthetic video creation, etc. However, all subjective quality assessment methodologies are also well-known to be cumbersome to apply at scale, as they require careful screening and briefing of participants, time-consuming test sessions under controlled lab conditions or in crowdsourced form, and statistical post-processing of the results [15] to ensure raters' bias and inconsistencies are taken into account. In addition, the derived results only characterize the examined test conditions: changes to the dataset or application conditions (e.g., encoding recipe changes), or even the use of different content require new subjective testing rounds.

On the other hand, objective quality metrics have evolved significantly in the last 10 years. Indeed, video quality metrics that encapsulate elements of human perception [2, 15, 23], perceptual modelling of encoding artifacts [4, 14], as well as viewing setup awareness [28] have emerged as strong contenders for the characterization visual quality impairments in video systems. This has also led to the research community now moving away from SNR-optimization in favour of structural similarity (SSIM) [28, 30], video multi-method assessment fusion (VMAF) [14], Apple's advanced video quality tool (AVQT) [25], and others. Even though many of the more accurate metrics are computationally expensive to apply at scale, their deployment can be automated and offers a way to score quality impairments of various video signal processing systems for large datasets and under varying application conditions [14, 25].

However, while such metrics have been evolving, so have algorithms that focus on perceptual quality, such as psychovisually-optimized video coding tools within AV1 and VVC [10, 20, 35, 37], deep perceptual preprocessing for video coding [4, 37], perceptually-optimized video denoising and upscaling [6] and synthetic video generation with generative adversarial networks [17]. Given that many of these systems optimize for variations of cost functions that align to components of widely-used metrics like VMAF and SSIM, this makes their performance diverge when assessed with multiple state-of-the-art metrics. An example is illustrated in Fig. 1, where BD-rates obtained for different encoders (and with and without the use of deep perceptual preprocessing [4]) vary significantly from one metric to another. Such observations have been reported for a variety of applications and test conditions by a number of independent studies [2, 37]. This is further exasperated with the use of generative neural network loss functions [6, 17]. Within the remit of encoding systems, theoretical studies have termed such discrepancies as the perception-distortion trade-off [3].

This is by no means a negative development: after all, aligning to human perception should indeed be the goal of any video signal processing system [1, 10, 14, 20, 25]. However, it opens up the following challenges:

- *What is the best way to combine, or fuse, multiple distortion-oriented and perception-oriented quality metrics in order to match subjective mean opinion scores of a video signal processing application?*
- *How can this be done efficiently, i.e., without the need for extensive subjective tests?*

To address these questions, we propose the use limited subjective testing with guided pseudorandom sequence and parameter selection that spans the operational settings of a narrow-domain problem. This is then used to train a support vector regressor in order to derive a single metric that can be used for system testing at scale. The benefits of this approach are validated in two different domains, i.e., video coding systems with and without perceptual preprocessing, as well as video denoising systems. As shown in the rightmost part of Fig. 1, the results of this approach can mitigate the discrepancies of individual metrics in a manner that is guided by the recovered quality scores from subjective tests.

2 RELATED WORK

Given a reconstructed frame $\hat{x}$ and a ground-truth reference frame x, full-reference distortion metrics quantify the quality of $\hat{x}$ by its pairwise discrepancy from x. Common metrics for distortion include PSNR, SSIM [30], MS-SSIM [31], information fidelity criterion (IFC) [22] and visual information fidelity (VIF) [21]. In the case of video, these metrics are commonly computed per frame and consolidated over the video with an arithmetic or harmonic mean. While distortion metrics are ideal for quantifying fidelity to the source, Blau *et al.* [3] illustrate that they are unable to capture the perceptual quality of the content. Optimizing for PSNR or mean squared error (MSE) renders a blurry output that is the average of all possible solutions in the pixel space weighted by their likelihoods [3, 4, 13]. A blurry output does not necessarily lie on the natural image manifold and may not be perceptually aligned to the source. The perceptual quality of a frame can instead be fully captured by the divergence between the distribution of reconstructed frames $p(\hat{x})$ and reference frames $p(x)$, either in the pixel, wavelet or DCT domains [16, 19]. GAN-based methods [1, 9, 12, 13, 29] employ adversarial training to directly minimize some form of divergence (e.g. Jensen-Shannon) between the generated and reference image distributions and thus generate higher quality images than those achievable by a simple autoencoder trained with MSE. However, the distortion may be high, since divergence is measured between distributions of images and not pairwise instances of x and $\hat{x}$.

As proved by Blau *et al.* [3], there in fact exists a perception-distortion bound where perception must be traded off for distortion or vice-versa. Some full-reference distortion measures such as LPIPS [36], DISTS [8], FUNQUE [27] or VMAF [14], which are more perceptually-oriented due to further training with subjective ratings, may present a weaker tradeoff at the boundary when compared to the likes of PSNR or SSIM. In this paper, we propose to capture the boundary behavior of multiple reference-based metrics from the Netflix libvmaf library [14], and orient towards perceptual quality by learning a fusion with subjective ratings. We achieve this with very limited testing needed, which is shown to be three orders of magnitude lower than the full exploration space of a typical video streaming system. Such subjective testing is carried out in this work by ITU-T P.910 ACR [11] followed by post-processing [15], but other forms of subjective testing like crowdsourced quality score collection [7] can also be employed if necessary.

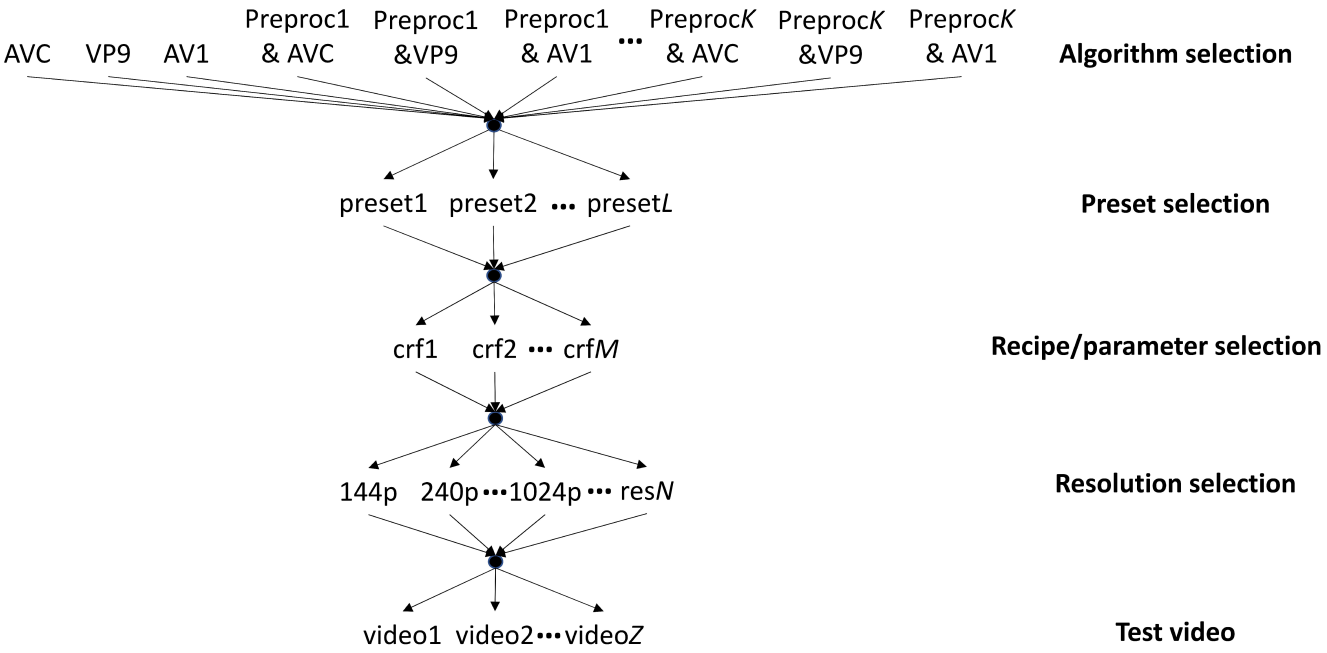


Figure 2: Typical algorithmic, preset, parameter, resolution and content space exploration for video encoding or denoising tests. Parameter crf stands for constant rate factor and Preproc1,...,Preproc$K$ stand for K variations of a preprocessing algorithm prior to encoding (if applicable).

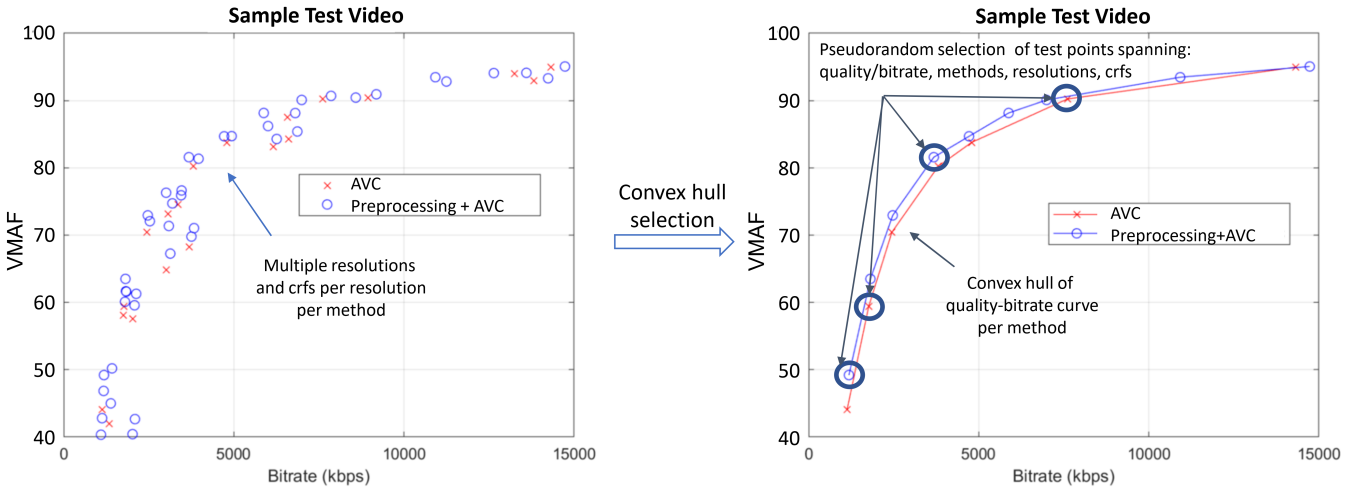


Figure 3: VMAF-vs-bitrate plots for: (left) multiple qualities/bitrates, methods, resolutions and multiple CRFs; (right) convex hull selection across the parameter space and pseudo-random sampling of the convex-hull.

3 FROM LIMITED DOMAIN-SPECIFIC SUBJECTIVE TESTING TO FUSION OF MULTIPLE OBJECTIVE METRICS

We begin by describing the typical parameter space of video streaming systems and the way we pseudo-randomly sample it for limited-scale subjective testing within a specific domain of interest, such as video coding or video denoising (Section 3.1). We then present the subjective testing methodology and the manner via which we fuse multiple subjective metrics (Sections 3.2 and 3.3).

3.1 Sample Space

The typical exploration space of video coding or denoising experiments is shown in Fig. 2. The figure illustrates the use of three

coding standards (AVC, VP9 and AV1) with and without K variants of a video preprocessing algorithm [4]. For each of these, one can select: L encoding presets, M quality settings (in terms of constant rate factors), and N encoding resolutions. Finally, a typical assessment requires the use of a library of Z video test assets that span the video content of interest. Even by selecting modest values of: $K = 2$, $L = 5$, $M = 8$, $N = 6$, and $Z = 100$, this leads to $3 \times (K + 1) \times L \times M \times N \times Z = 216,000$ outputs that need to be assessed. This is clearly infeasible to handle with either lab-based or crowdsourced-based testing. Even if it would be deemed to be feasible, such tests cannot be generalized to more experiments without some form of fitting of objective quality metrics to recovered mean opinion scores.

However, not all experimental conditions are expected to be met in real-world conditions. For example, testing very-low crf values for very low resolutions is unnecessary, as they would not be used in practice: a higher resolution and lower crf value would offer better quality at lower bitrate [33]. Similarly, a very complex encoding preset at low resolution for a very static input video would also not occur in reality, as a faster preset at higher resolution would suffice for high-quality video representation at a modest bitrate range. Therefore, we propose the use of convex-hull selection for the pruning of the subset of test conditions are that most likely to occur in practice. An example of such convex-hull pruning is shown in Fig. 3, where it is shown that multiple quality-resolution-bitrate points can be reduced to the convex hull of points that are optimal using convex-hull selection [33]. In our work, we utilize VMAF as the metric to select quality-bitrate points for each encoder and each preset of interest. This is because VMAF has been shown to offer a good balance between perception and distortion. Therefore, convex-hull selection based on VMAF is expected to provide for points with a better trade-off in the perception-distortion bound [3]. From the surviving points, we then pseudo-randomly sample across encoders (and also using preprocessing & encoding, if applicable), resolutions, encoding presets, and test video files, in order to have a balanced representation of the viable experimental space in our tests. For the exploration space of Fig. 2, this typically results in as few as 450 sample videos, which can be easily handled via a small-scale subjective test. When comparing to the 216,000 possible test results needed to explore the sample space of Fig. 2, this corresponds to a reduction of effort by three orders of magnitude.

3.2 Subjective Testing

Once content has been selected, we follow a standard subjective testing methodology, such as ITU-T P.910 or BT.500 absolute category rating (ACR) or degradation category rating (DCR). This involves setting a standard viewing angle and distance from the screen, controlled lighting and screen conditions, rater prescreening for color blindness and eyesight, and rater briefing for the scoring task. In our work, we used P.910 ACR with a five-scale rating and hidden reference, as this has been shown to allow for the best use of rating effort, i.e., the maximum number of samples within the allocated time. The SUREAL package from Netflix is used for post-processing the ACR ratings [15]. Essentially, SUREAL recovers the subjective quality scores from the raw ACR ratings by assuming a linear model and using maximum likelihood estimation (MLE) to jointly estimate the subjective quality of videos and rater bias and inconsistency [15].

The obtained scatter plots for the video coding application are shown in Fig. 4 along with their corresponding correlation coefficients (CC), Spearman's ranked order correlation coefficients (SRCC) and root mean squared error (RMSE). VMAF and VMAF-NEG are shown to be significantly more correlated to quality scores than SSIM (FB-MOS) and PSNR. Similarly, the obtained scatter plots for the video denoising application are shown in Fig. 5. Similarly as before, VMAF and VMAF-NEG are shown to be significantly closer to the recovered quality scores. SSIM is mildly correlated to the recovered quality from the subjective test and PSNR is completely uncorrelated. This is not an unexpected result, given that the utilized neural network denoising and upscaling architecture (described in the experimental section) is trained for perceptual and generative adversarial loss functions rather than signal distortion. However, this emphasizes the need for consolidation between subjective scores and objective metrics for state-of-the-art video signal processing systems that go significantly beyond signal-to-noise ratio minimization.

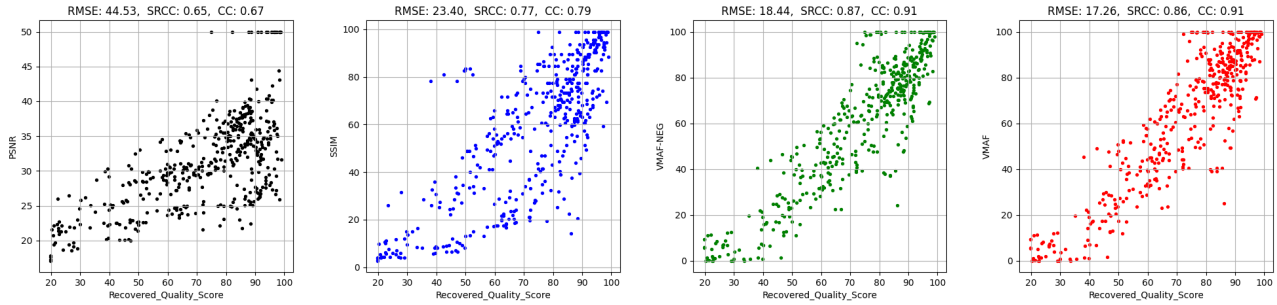
3.3 Fusion of Multiple Objective Quality Metrics

We propose to improve the quality assessment of any individual quality metric, by fusing metrics into a single P.910-MOS metric that conforms to the sample space and subjective testing methodology of its input features. A ν-support vector regressor (ν-SVR) model [5, 24] is trained to estimate the recovered quality scores via the use of multiple quality metrics. We use PSNR, SSIM, VMAF-NEG, and VMAF as input features, as they are readily derived at scale via the libvmaf library of FFmpeg. Contrary to a ϵ-SVR, where we control the error term ϵ, with a ν-SVR we are directly able to control the number of support vectors required for optimization. Given a set of training feature vectors $\boldsymbol{x}_i \in \mathbb{R}^p$, where p represents the number of input features and corresponding target outputs $y_i \in \mathbb{R}^1$, the primal problem can be written as:

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi,\xi^*} \quad & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\left(\nu\epsilon + \frac{1}{l}\Sigma_{i=1}^{l}(\xi_i + \xi_i^*)\right) \\ \text{s.t.} \quad & (\boldsymbol{w}^\top \phi(\boldsymbol{x})_i + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\boldsymbol{w}^\top \phi(\boldsymbol{x})_i + b) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, i = 1, \ldots, l, \epsilon \geq 0 \end{aligned} \tag{1}$$

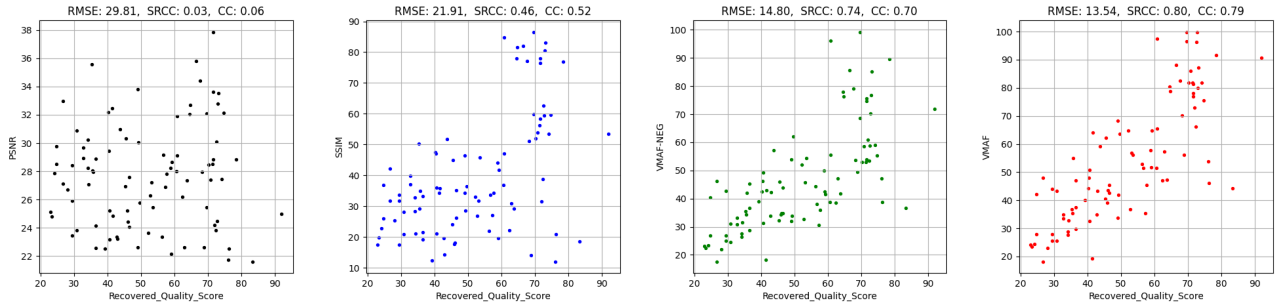
where $\boldsymbol{w}$ and b are learned weights and biases respectively, ξ and ξ^* are slack variables and l is the number of videos. There are three hyperparameters: ν = the proportion of the number of support vectors versus the total number of samples, C = regularization parameter on the loss function, γ = the radius parameter of the radial basis function (RBF) kernel, denoted as ϕ in (1). In short, the loss function represents a tradeoff, where the first term enforces flatness by penalizing large $\boldsymbol{w}$, while the second term penalizes deviations larger than the desired maximum error. The constraints ensure that errors in prediction are less than ϵ (plus some slack). We optimize the loss function on a training set by constructing a Lagrange function with the constraints and solving the dual problem [24].

We follow a standard 80-20 training/test split over AV2CTC sequences. The datapoints that survive the pseudo-random sampling



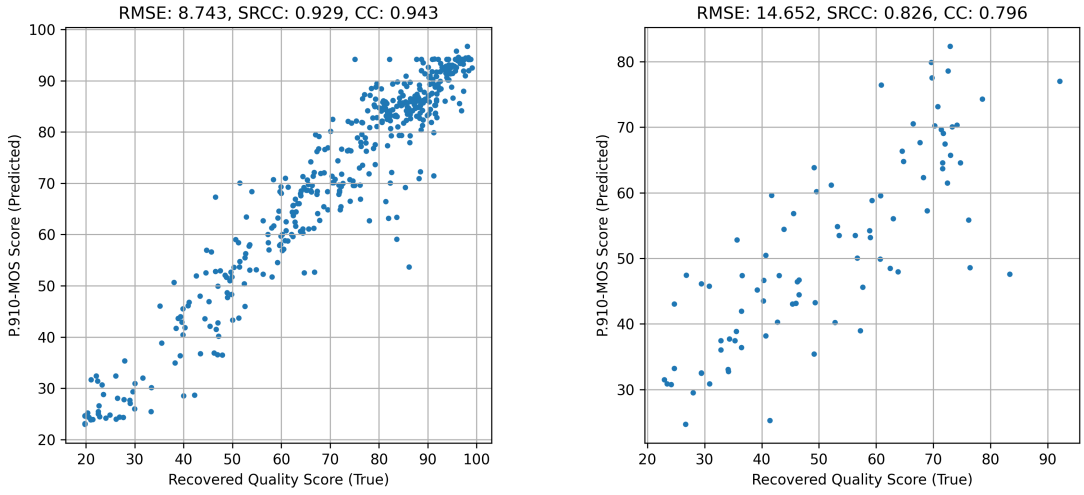
(a) PSNR **(b) SSIM** **(c) VMAF-NEG** **(d) VMAF**

Figure 4: Scatter plots of PSNR, SSIM, VMAF-NEG and VMAF vs. recovered quality scores and corresponding SRCC and CC for the subjective tests with the video coding algorithms.



(a) PSNR **(b) SSIM** **(c) VMAF-NEG** **(d) VMAF**

Figure 5: Scatter plots of PSNR, SSIM, VMAF-NEG and VMAF vs. recovered quality scores and corresponding SRCC and CC for the subjective tests with the video denoising algorithms.



(a) Encoding **(b) Denoising & upscaling**

Figure 6: Scatter plot of of SVR predicted scores vs recovered quality (true) scores for (a) Encoding ($\nu = 0.5$, $\gamma = 0.85$, $C = 1$) and (b) Denoising & upscaling ($\nu = 1.0$, $\gamma = 0.0025$, $C = 32.0$).

of the convex hull are then used for training the SVR. The input features to the SVR are scaled and normalized prior to training. The ν-SVR is optimized with a radial basis function (RBF) kernel and k-fold cross validation, with the number of folds equal to 4. A grid search is used to find the values for hyperparameters that jointly optimize the CC and SRCC, with $C \in [2^{-5}, 2^{15}]$ and $\gamma \in [2^{-15}, 2^{3}]$. The remaining 20% of sequences are then used for testing on the video coding and denoising applications under the testing conditions described in Sections 4.1 and 4.2 respectively.

4 EXPERIMENTS

We validate the proposed methodology for domain-specific fusion of multiple video quality metrics on two applications: video coding with and without neural-network based preprocessing, and neural-network based video denoising. In both cases, the utilized algorithms are perceptually-oriented, which causes discrepancy in performance assessment with different reference-based quality metrics.

4.1 Video Coding

One of the key objectives of optimized video streaming is the selection of the appropriate encoder and encoder preset that meets the desired execution time complexity and bitrate efficiency versus the state-of-the-art. While it is well known that bitrate efficiency increases with the use of encoding presets of increased complexity, e.g., x264 AVC preset=veryslow versus preset=fast, it is not clear what is the relationship between different encoders. The use of state-of-the-art preprocessing technologies for perceptual optimization [4, 10, 20, 37] complicates this further, as such preprocessing makes the encoding even more perceptually-tuned than the original perceptual tuning within the encoders & recipes themselves.

The baseline results corresponding to this experiment are shown in the left four plots of Fig. 1. They have been generated for the AV2 CTC test content [34] (excluding sequences above 1080p resolution) and the optimized encoding recipes from recent work [32]. This work also describes the full set of test conditions with respect to encoding presets, multi-resolution encoding ladders and choice of crf values, which have been adopted for our experiments. In summary:

- resolutions span from 144p to 1080p;
- the crf ranges span from 18 to 42 (for x264 AVC) and from 22 to 63 for the svt-av1 AVC and vpxenc VP9;
- the utilized presets are shown in Fig. 1 and span the bulk of the complexity-quality tuning on offer by each encoder;
- convex-hull selection takes place for all encoders and the BD-rates for all surviving points in the convex hull are measured using slope-based integration and the Netflix libvmaf BD-rate calculator [14];
- execution time is measured using GNU parallel on an Intel CPU (in our case this was an Intel Xeon 8275CL 24-core CPU);
- VMAF, VMAF-NEG, SSIM and PSNR are computed using the Netflix libvmaf library [14] and in our case SSIM is rescaled so that it is more aligned to MOS using the FB-MOS rescaling [18];
- preprocessing comprises an optimized deep perceptual preprocessing model from recent work [4] that provides for perceptually-optimized preprocessing with a single model for all encoders, presets, resolutions and crf values.

The entire set of libvmaf JSON measurements is approximately 2.5 million files, which showcases the comprehensive nature of this test. As shown in Fig. 1, despite the extensive nature of the test, when focusing on a single metric like VMAF, SSIM or PSNR, one can reach very different conclusions with respect to the complexity–vs.–BD-rate efficacy of the different encoders and their presets. For example,

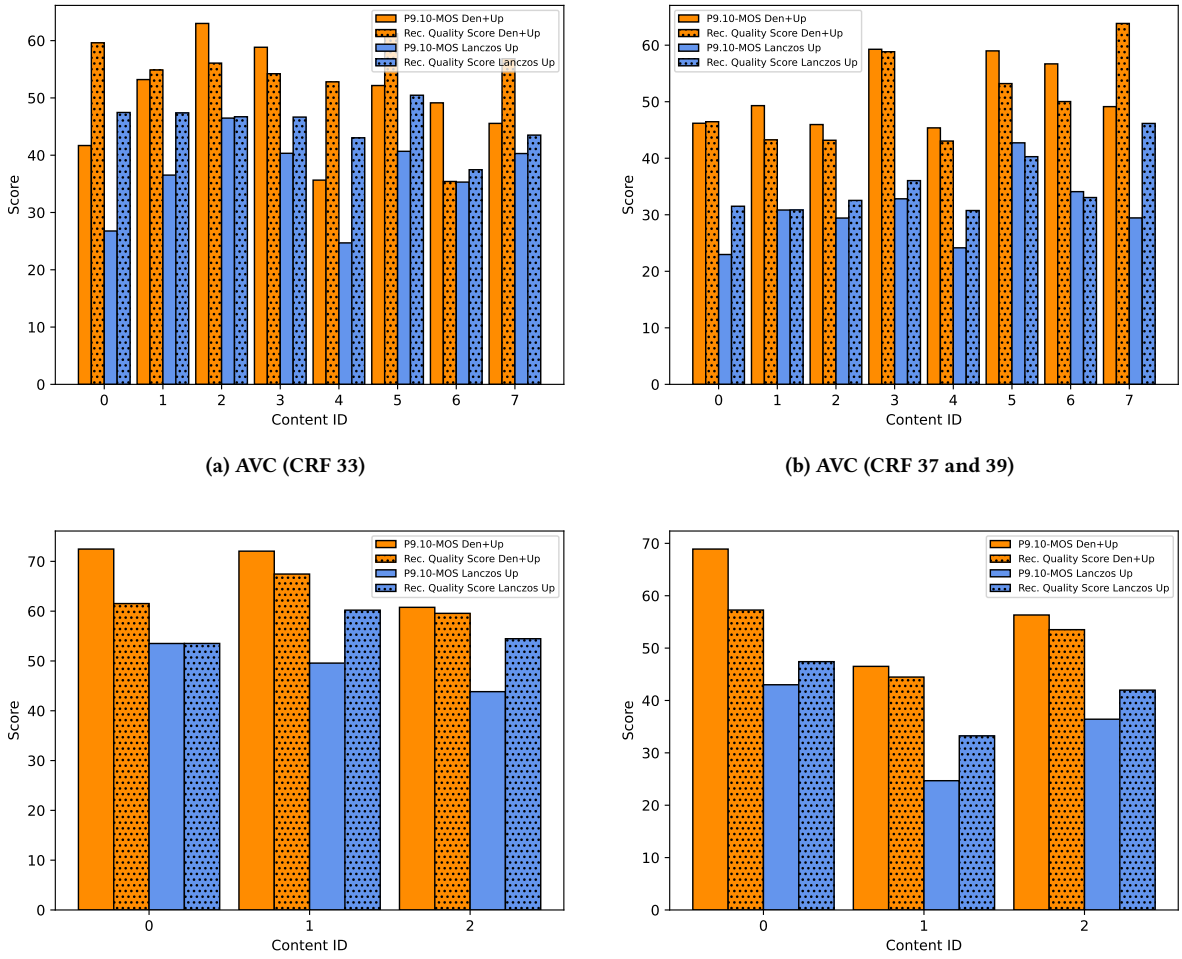
- when considering SSIM, the M12 preset of svt-av1 AV1 encoder is shown to be on-par to x264 AVC preset=fast with respect to execution time and BD-rate; on the other hand, they differ by more than 60% in BD-rate when considering VMAF;
- x264 AVC preset=veryslow with preprocessing is shown to be faster and almost on-par to vpxenc VP9 preset=cpu3 with respect to SSIM BD-rate, but is more than 15% worse when considering VMAF;
- the difference between the fastest and slowest presets of each encoder is much less pronounced in SSIM BD-rate versus when considering VMAF BD-rate;
- the effect of perceptually-optimized preprocessing is much more pronounced on all encoders when considering VMAF than when considering SSIM as the quality metric, and it is reversed when considering PSNR (since preprocessing changes the input to the encoder and all metrics are measured using the input video as reference).

When combining all four different metrics (PSNR, SSIM, VMAF-NEG and VMAF) into a single P.910-MOS metric via the proposed methodology and its instantiation, we obtain the derived SVR scatter plot of Fig. 6(a). Comparing with Fig. 4, the SVR plot shows increased correlation coefficients (SRCC and CC) versus using any metric independently, which indicates a better fit to the recovered quality scores.

We can now evaluate the BD-rate performance of preprocessing vs encoding complexity directly with our P.910-MOS metric, as illustrated in the rightmost plot of Fig. 1. The derived BD-rates of P.910-MOS fall between those of SSIM, VMAF-NEG and VMAF, and provide for a directly-interpretable way of cross-comparing different encoding technologies and their settings than an ad-hoc weighted averaging of BD-rates of different metrics. Importantly, the obtained BD-rates are also in good agreement with the BD-rates obtained by slope-based merging of recovered quality scores for the example cases of x264 AVC preset=veryslow and vpxenc VP9 preset=cpu0. Specifically, when BD-rate of preprocessing is measured vs. encoding of the same AVC and VP9 presets for the AV2 CTC results of the central part of Fig. 1, we obtained -11% and -9% (resp.), while for the slope-based merging of recovered quality scores we obtained -12% and -6% (resp.).

4.2 Video Denoising and Upscaling

In the second application, we aim to denoise and upscale video by learning to recover from compression-induced artifacts. The inference architecture tested comprises a single-frame input model



(a) AVC (CRF 33)

(b) AVC (CRF 37 and 39)

(c) VP9 (CRF 52)

(d) VP9 (CRF 61)

Figure 7: Recovered quality scores vs. SVR-based P.910-MOS for Lanczos upscaling and neural network denoising and upscaling.

with only 5 residual blocks in its inference. The first layer is designed by a strided convolution (downscaling the input resolution by a factor of 2) with LeakyRelu non-linearity thereafter, followed by 5 residual blocks and two Depth2Space (pixelshuffle) layers for upscaling. Each convolutional feature map has 64 channels. Finally, a convolutional layer is used to predict the output.

While the target for this application is high-crf encoding, and training is done using crf $\in \{35, .., 42\}$ for AVC and crf $\in \{58, .., 63\}$ for VP9 encodes, our tests involve the use of both low, medium and high-crf content in order to test the architecture in out-of-distribution cases and ensure it does not lead to worse results than standard upscaling. In particular, for inference tests we used 720p and 1080p content from the AV2 CTC dataset that has been downscaled to 360p and 540p (resp.) and compressed with AVC crf $\in \{23, 33, 37, 39\}$ and VP9 crf $\in \{33, 52, 61\}$ using x264 AVC preset=veryslow and vpxenc VP9 preset=cpu5. The produced downscaled and compressed output is then:

- upscaled with FFmpeg Lanczos (with setting param5=0);
- denoised and upscaled with the neural network architecture under consideration

As with the video coding application, we first combine the four individual metrics from the Netflix libvmaf library (VMAF, VMAF-NEG, PSNR and SSIM) into a single P.910-MOS metric that predicts the recovered quality scores obtained after SUREAL processing. In this case, the derived scatter plot of Fig. 6(b) shows similar but improved correlation coefficients to VMAF and VMAF-NEG in Fig. 5(c)+(d). Out of the utilized libvmaf metrics, only VMAF is close to the subjective scores, and PSNR and SSIM can safely be disregarded when it comes to visual quality assessment of the utilized video denoising and upscaling architecture.

Given our learned P.910-MOS metric, we assess the performance of the denoising and upscaling architecture versus Lanczos upscaling on a subset of sequences from the AV2 CTC dataset in Fig. 7, for mid CRF and high CRF ranges and on both AVC and VP9. As expected, the difference between the two methods increases as the

noise levels (i.e., crf values) increase. Additionally, the derived P.910-MOS metric appears to follow the true recovered quality scores for most cases, thereby providing an automated way to assess the impact of neural denoising and upscaling versus standard linear upscaling at scale.

5 CONCLUSION

We address the growing challenge of how to assess perceptually-oriented video signal processing techniques more reliably and in a manner that can scale to large datasets. Our results on video coding and video denoising & upscaling show that even psychovisually-tuned metrics like Netflix libvmaf's video multi-method assessment fusion (VMAF) and structural similarity (SSIM) most often lead to a mixed picture with respect to the tangible benefits of advanced coding or denoising methods. While this can be resolved with large-scale subjective tests, a thorough exploration of the parameter and content space quickly becomes infeasible. Instead we propose the significantly condense the application and parameter space with a pseudorandom sampling strategy that is guided by metrics, and then carry out limited subjective testing in order to fit the fusion of multiple metrics to the recovered quality scores of each application of interest. Validation in video coding and video denoising and upscaling experiments showed that this can provide for a better way to combine standard libvmaf metrics towards a fused objective score that can be applied at scale. Future work can investigate the extension of this methodology to further objective metrics and also other applications.

6 ACKNOWLEDGEMENTS

The work of the authors from iSIZE was supported in part by a grant from Innovate UK, project SEQUOIA (#96984)

REFERENCES

[1] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. 2019. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 221–231.

[2] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. 2018. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing* 12, 2 (2018), 355–362.

[3] Yochai Blau and Tomer Michaeli. 2018. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6228–6237.

[4] Aaron Chadha and Yiannis Andreopoulos. 2021. Deep Perceptual Preprocessing for Video Coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14852–14861.

[5] Chih-Chung Chang and Chih-Jen Lin. 2002. Training v-support vector regression: theory and algorithms. *Neural computation* 14, 8 (2002), 1959–1977.

[6] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. 2018. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3155–3164.

[7] Ross Cuttler. 2022. *P.910 Crowd*. Retrieved April 7, 2022 from https://github.com/microsoft/P.910

[8] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728* (2020).

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[10] Christian R Helmrich, Sebastian Bosse, Mischa Siekmann, Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. 2019. Perceptually optimized bit-allocation and associated distortion measure for block-based image or video coding. In *2019 Data Compression Conference (DCC)*. IEEE, 172–181.

[11] Recommendation ITU-T. 2008. ITU-t P. 910. *Subjective video quality assessment methods for multimedia applications* (2008).

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[13] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4681–4690.

[14] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock. 2018. VMAF: The journey continues. *Netflix Technology Blog* 25 (2018).

[15] Zhi Li, Christos G Bampis, Lucjan Janowski, and Ioannis Katsavounidis. 2020. A simple model for subject behavior in subjective experiments. *Electronic Imaging* 2020, 11 (2020), 131–1.

[16] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* 21, 12 (2012), 4695–4708.

[17] Maxime Oquab, Pierre Stock, Daniel Haziza, Tao Xu, Peizhao Zhang, Onur Celebi, Yana Hasson, Patrick Labatut, Bobo Bose-Kolanu, Thibault Peyronel, et al. 2021. Low bandwidth video-chat compression using deep generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2388–2397.

[18] Shankar L Regunathan, Haixiong Wang, Yun Zhang, Yu Ryan Liu, David Wolstencroft, Srinath Reddy, Cosmin Stejerean, Sonal Gandhi, Minchuan Chen, Pankaj Sethi, et al. 2020. Efficient measurement of quality at scale in Facebook video ecosystem. In *Applications of Digital Image Processing XLIII*, Vol. 11510. SPIE, 69–80.

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

[20] Heiko Schwarz, Muhammed Coban, Marta Karczewicz, Tzu-Der Chuang, Frank Bossen, Alexander Alshin, Jani Lainema, Christian R Helmrich, and Thomas Wiegand. 2021. Quantization and entropy coding in the versatile video coding (VVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 10 (2021), 3891–3906.

[21] H.R. Sheikh and A.C. Bovik. 2006. Image information and visual quality. *IEEE Transactions on Image Processing* 15, 2 (2006), 430–444. https://doi.org/10.1109/TIP.2005.859378

[22] H.R. Sheikh, A.C. Bovik, and G. de Veciana. 2005. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing* 14, 12 (2005), 2117–2128. https://doi.org/10.1109/TIP.2005.859389

[23] Zeina Sinno and Alan C Bovik. 2018. Large scale subjective video quality study. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 276–280.

[24] Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14, 3 (2004), 199–222.

[25] Pranav Sodhani. 2021. *Evaluate videos with the Advanced Video Quality Tool*. Retrieved April 7, 2022 from https://developer.apple.com/videos/play/wwdc2021/10145/

[26] Marc Sullivan, James Pratt, and Philip Kortum. 2008. Practical issues in subjective video quality evaluation: Human factors vs. psychophysical image quality evaluation. In *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*. 1–4.

[27] Abhinau K Venkataramanan, Cosmin Stejerean, and Alan C Bovik. 2022. FUNQUE: Fusion of Unified Quality Evaluators. *arXiv preprint arXiv:2202.11241* (2022).

[28] Abhinau K Venkataramanan, Chengyang Wu, Alan C Bovik, Ioannis Katsavounidis, and Zafar Shahid. 2021. A hitchhiker's guide to structural similarity. *IEEE Access* 9 (2021), 28872–28896.

[29] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*. 0–0.

[30] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[31] Z. Wang, E.P. Simoncelli, and A.C. Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, Vol. 2. 1398–1402 Vol.2. https://doi.org/10.1109/ACSSC.2003.1292216

[32] Ping-Hao Wu, Ioannis Katsavounidis, Zhijun Lei, David Ronca, Hassene Tmar, Omran Abdelkafi, Colton Cheung, Foued Ben Amara, and Faouzi Kossentini. 2021. Towards much better SVT-AV1 quality-cycles tradeoffs for VOD applications. In *Applications of Digital Image Processing XLIV*, Vol. 11842. International Society for Optics and Photonics, 118420T.

[33] Ping-Hao Wu, Volodymyr Kondratenko, Gaurang Chaudhari, and Ioannis Katsavounidis. 2021. Encoding Parameters Prediction for Convex Hull Video Encoding. In *2021 Picture Coding Symposium (PCS)*. IEEE, 1–5.

[34] XIPH Media. 2021. *AV2-CTC Test Set*. Retrieved April 7, 2022 from https://media.xiph.org/video/aomctc/test_set/

[35] Fan Zhang, Angeliki V Katsenou, Mariana Afonso, Goce Dimitrov, and David R Bull. 2020. Comparing VVC, HEVC and AV1 using objective and subjective assessments. *arXiv preprint arXiv:2003.10282* (2020).

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[37] Yun Zhang, Linwei Zhu, Gangyi Jiang, Sam Kwong, and C-C Jay Kuo. 2021. A Survey on Perceptually Optimized Video Coding. *arXiv preprint arXiv:2112.12284* (2021).