Why No Reference Metrics for Image and Video Quality Lack Accuracy and Reproducibility

Margaret H. Pinson

***Abstract*—This article provides a comprehensive overview of no reference (NR) metrics for image quality analysis (IQA) and video quality analysis (VQA). We examine 26 independent evaluations of NR metrics (previously published) and analyze 32 NR metrics on six IQA datasets and six VQA datasets (new results). Where NR metric developers claim Pearson correlation values between 0.66 and 0.99, our measurements range from 0.0 to 0.63. None of the NR metrics we analyzed are accurate enough to be deployed by industry. Performance evaluations that indicate otherwise are based on insufficient data and highly inaccurate. We will examine development strategies, tools, datasets, root cause analysis, and our baseline metric for collaboration, Sawatch.**

***Index Terms*—Image quality, metric, no reference, NR, root cause analysis, RCA, Sawatch, video quality.**

I. Introduction

CONFLICTING assessments lead to dissenting opinions on the reliability of no-reference (NR) metrics for video quality assessment (VQA) and image quality assessment (IQA). NR metric developers often publish extremely favorable performance claims, such as 0.99 Pearson correlation coefficient between the NR metric and the mean opinion scores (MOS). But this is often just a single dataset. This sets unrealistic expectations based on insufficient data.

At the opposite extreme, industry assessments and discussions during the Video Quality Experts Group (VQEG) meetings often report poor performance for NR metrics. These assessments are typically unpublished and thus difficult to verify or replicate. Intel [1] evaluated six NR-IQA metrics on consumer content and reported that "those algorithms did not correlate well with human perceptual judgement of image quality." Shanghai Jiao Tong University used their Smartphone Camera Photo Quality Database (*SCPQD2020*) to analyze ten NR-IQA metrics and reported that "no current objective NR model works well" [2].

Part of the problem is a lack of communication between academic researchers and industry users. To address this issue, VQEG created and published industry requirements for NR metrics [3]. These requirements simplify into two assertions. First, to be exploitable, NR metrics must provide root cause analysis (RCA). Most industry applications for NR metrics involve identifying and mitigating specific impairments. Second, the external validity of an NR metric (outside the lab where it was developed) depends on its ability to assess camera capture impairments.

Another part of the problem is the lack of a comprehensive assessment of the current state of NR metric research for modern camera systems. Developers need this information to make the best decisions on where to focus future research. Industry needs this information to trust and deploy NR metrics.

We will begin by examining the accuracy and repeatability of subjective tests. We will consider the bias and noise associated with dataset design. We present a primary experiment design, for datasets that will be used to train or test NR metrics. This information creates an upper bound on NR metric performance.

We will then survey prior art and describe issues that create difficulties for NR metric research. We consider what "good quality" means to different users and how this impacts the datasets used to train NR metrics. We identify various strategies for developing NR metrics.

We compare statistics reported by NR metric developers with performance statistics from 26 independent evaluations on modern camera systems. We observe a concerning trend of training and testing on a single dataset. Comparisons among independent evaluations of NR metrics produce unstable statistics, because they used too few data points. We conclude that NR metrics must be trained and tested with at least ten datasets.

Guided by insights from prior research, we present a paradigm for NR metric development and describe NR metric *Sawatch*, which implements this paradigm. We split the research effort into separate algorithms that assess different impairments and that can be studied separately. These individual NR metrics combine to provide an overall quality estimation. A simple equation allows the end-user to adjust the weight of each impairment on the overall quality estimation, based on their unique requirements.

NR metric *Sawatch* leverages our open software framework for collaborative development of NR-IQA and NR-VQA metrics, called the *NR Metric Framework*. This framework provides the support tools necessary to begin research and avoid common mistakes. It also facilitates training and evaluating NR metrics on multiple datasets. Standard support tools will enable repeatable analyses and incremental improvements.

Using the *NR Metric Framework*, we evaluate the accuracy of 32 NR metrics. To ensure stability and reliability, our analysis uses twelve datasets that characterize different aspects of

Manuscript received 26 April 2022; revised 28 June 2022; accepted 29 June 2022. Date of publication 25 July 2022; date of current version 6 March 2023. This work was supported in part by the Public Safety Communications Research (PSCR) Division of the National Institute of Standards and Technology (NIST).

The author is with the National Telecommunications and Information Administration, Institute for Telecommunication Sciences, Boulder, CO 80305 USA (e-mail: mpinson@ntia.gov).

Digital Object Identifier 10.1109/TBC.2022.3191059

modern camera systems. We do not limit our analyses to the NR metric's intended scope, and we do not retrain machine learning algorithms. These analyses include both VQA and IQA metric research because, in the modern age of digital monitors, images are indistinguishable from still videos.

Where NR metric developers claim Pearson correlation coefficient values between 0.66 and 0.99, our measurements range from 0.0 to 0.63. Our analysis confirms the need for more research and development on NR metrics for modern camera systems. None of the NR metrics we analyzed are accurate and reliable enough for commercial applications.

We then present the performance of NR metric *Sawatch* on our twelve datasets. Caution must be exercised when comparing *Sawatch* with the other NR metrics in this paper, because *Sawatch*'s training data is the other metrics' testing data. Our final analysis uses the *Sawatch* RCA to reveal complex relationships between impairments, quality, industry use cases, and NR metric performance. These confounding factors may explain some of the instability we observe. The code, media, and data used by this report are available online at [4].

Our goal is to support the development of NR metrics that are accurate enough for commercial applications. The broadcast workflow would be able to detect quality problems and specific impairments in real-time broadcast streams. NR metrics could be used to optimize the encoding parameters of real-time video streams, similar to per-title encoding optimization for video on demand [5].

II. Glossary, Statistics, and Scatter Plots

A. Glossary

We will begin with a glossary. NR metrics and datasets of media with subjective ratings tend to have very long names that hinder readability. We will use abbreviations instead. Blue text indicates names and abbreviations that we created, when the author did not propose a name or abbreviation.

Table I and Table II list the NR metrics mentioned in this report. The column "#" identifies the number of datasets used to develop the NR metric. Roman numerals in the reference ("Ref.") column refer to sections of this paper. Table III lists the datasets mentioned in this report. The "Notes" column of Table III briefly summarizes the experiment design, using the following codes: "I" for images, "V" for videos, "C" for camera capture impairments, "T" for transcoding (possibly with rescaling), "S" for simulated impairments, "E" if the dataset uses an experiment method to rate the media, and "M" for miscellaneous (e.g., tonemapping, multiexposure fusion, image enhancement, and image blending algorithms).

Three of the NR metrics in Table I and Table II are modified from the author's original intent. *2stepQA* [6] is a two-step reduced reference (RR) metric. The first step, an NR metric that we refer to as *2stepQA-NR*, is an NR constrained variant of *SpEED-QA* [44]. *LBP* [29] was intended for texture classification. *HVS-MaxPol* includes four variants, *NSS* three variants, and *SpEED-NR* two variants. Since these variants yield similar results, we only examine *HVS-MaxPol natural 1*, *NSS* trained on *CID2013*, and *SpEED-NR SingleScale*.

TABLE I
NR Metric List

Abbrev.	Name	#	Ref.
2stepQA-NR	Two Step Quality Assurance, NR Constrained	1	[6]
ADMD	Assessment of Dermoscopy Images with Multiple Distortions	2	[7]
AGWN	Additive Gaussian White Noise	1	[8]
BLIINDS-II	Blind Image Integrity Notator using DCT Statistics II	1	[9]
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator	2	[10]
BTMQI	Blind Tone-Mapped Quality Index	2	[11]
C-DIVIINE	Complex Extension of DIIVINE	4	[12]
CPBD	Cumulative Probability of Blur Detection	1	[13]
CurveletQA	Curvelet Quality Assessment	2	[14]
DIIVINE	Distortion Identification-based Image Verity and Integrity Evaluation	2	[15]
dipIQ	Quality-Discriminable Imagepairs Inferred Quality	4	[16]
Entropy Noise	The authors do not provide a name	1	[17]
FRIQUEE	Feature Maps-Based Referenceless Image Quality Evaluation Engine	6	[18]
HVS-MaxPol	Human Visual System MaxPol	7	[19]
iAITech-NJIT	5 NR metrics that asses quality for computer vision applications	0	[20]
JNB	Just Noticeable Blur	1	[21]
JP2KNR	JPEG2000 No-Reference	1	[22]
Key Indicators	15 NR metrics that indicate quality impairments	5	[23]-[28]
LBP	Local Binary Patterns		[29]
Log-BIQA	Laplacian of Gaussian Blind Image Quality Assessment	3	[30]
MaxPol	Synthetic MaxPol	4	[31]
Munsell Red	Munsell Red	2	[32]
NIMA	Neural Image Assessment	3	[33]
NIQE	Natural Image Quality Evaluator	2	[34]
NIQE-K	*NIQE*-Kurtosis	3	[35]
NoRM	Non-reference Quality Metric	1	[36]
NR-IQA-CDI	NR-IQA Contrast-Distored Images	4	[37]
NR-mean	NR-IQA-CDI *Mean*	4	[37]
NR-std	NR-IQA-CDI *Standard Deviation*	4	[37]
NR-skew	NR-IQA-CDI *Skewness*	4	[37]
NR-kurtosis	NR-IQA-CDI *Kurtosis*	4	[37]
NR-entropy	NR-IQA-CDI *Entropy*	4	[37]
NR-PWN	No Reference Perceptually Weighted Noise	2	[38]
NSS	Natural Scene Statistics	3	[39]
OG-IQA	Oriented Gradients Image Quality Assessment	3	[40]
PIQE	Perception Based Image Quality Evaluator, also abbreviated "PIQUE"	3	[41]
QAC	Quality Aware Clustering	3	[42]
S-BlackLevel	*Sawatch Version 3*, Black Level	12	V
S-Blockiness	*Sawatch Version 3*, Blockiness	12	V
S-Blur	*Sawatch Version 3*, Blur	12	V
S-ColorNoise	*Sawatch Version 3*, Color Noise	12	V
S-dipIQ	*Sawatch Version 3*, *dipIQ* scaled to [0..1]	12	V
S-FineDetail	*Sawatch Version 3*, Fine Detail	12	V
S-Jiggle	*Sawatch Version 3*, Jiggle	12	V
S-PanSpeed	*Sawatch Version 3*, Pan Speed	12	V
S-Pallid	*Sawatch Version 3*, Pallid	12	V
S-SuperSat	*Sawatch Version 3*, Super Saturation	12	V
S-WhiteLevel	*Sawatch Version 3*, White Level	12	V
Sawatch	*Sawatch Version 3*	12	V

B. Groups of Datasets

Table III lists the 39 datasets mentioned in this paper. We will use twelve of these datasets to analyze the

TABLE II
NR METRIC LIST, PART 2

Abbrev.	Name	#	Ref.
SI and *TI*	Spatial Perceptual Information (SI) Temporal Perceptual Information (TI)	1	[43]
SpEED-NR	Spatial Efficient Entropic Differencing for Quality Assessment, NR Constrained	7	[44]
SSEQ	Spatial–Spectral Entropy-based Quality Index	1	[45]
TDME	Transform Domain Measure of Enhancement	1	[46]
TDMEC	Transform Domain Measure of Enhancement Color	1	[47]
TLVQM	Two Level Video Quality Model	4	[48]
V-BLIINDS	Video BLIINDS	2	[49]
VIQET	VQEG Image Qualtiy Evaluation Tool	1	[50], [51]
VIIDEO	Video Intrinsic Integrity and Distortion Evaluation Oracle	1	[52]

performance of various NR metrics. Nine of these datasets were designed for NR metric research. The remaining three datasets (*AGH/NTIA/Dolby*, CCRIQ and CCRIQ2 & VIME1) were not designed for NR metric research.

The first set contains six IQA datasets with camera impairments and user generated content (named IQA UGC). *BID* includes blur from a variety of causes with diverse subject matter. CCRIQ has photographs of the same subject matter taken with 23 cameras and displayed at two monitor resolutions (HD and 4K). The *CCRIQ2 & VIME1* dataset has two parts: *CCRIQ2* has extra photographs from CCRIQ, and *VIME1* has photographs of a city in Scotland. *CID2013* has a design similar to CCRIQ but one monitor resolution (HD) and limited scene composition. *ITS4S2* has a large variety of subject matter, cameras, and camera impairments. *LIVE-Wild* has a large variety of subject matter from mobile devices; these are 500 × 500 pixel images. The Table III "Notes" column marks these six datasets with **1**.

The second set contains three VQA datasets with camera impairments and user generated content (named *VQA UGC*). *ITS4S3* has simulated first responder content and a variety of cameras. *ITS4S4* has a mix of simulated camera pans and real camera pans; other impairments are avoided. *KoNViD-1K* contains a large variety of subject matter and camera impairments. The Table III "Notes" column marks these three datasets with **2**.

The third set contains three VQA datasets with transcoding impairments and broadcast content (named *VQA BC*). These datasets have subject matter, cameras, and bitrates suitable for broadcast applications. *AGH/NTIA/Dolby* contains *MPEG2*, *AVC*, and *HEVC*. ITS4S simulates a 720p adaptive bitstream ladder. Dataset *vqegHDcuts* reuses video files and MOSs from VQEG high definition (HD) tests, but each source video was cut whenever the content or camera motion changed. The Table III "Notes" column marks these three datasets with **3**.

The *vqegHDcuts* dataset was created using an unprecedented method, as described in [84]. Longer videos that contained temporal changes were divided into shorter segments that do not contain temporal changes. The original rating was assigned to each segment. The goal was to exclude temporal integration from the videos and the NR metric. This is an unprecedented method, so the magnitude of the error added to the MOSs is unknown. We created this faux dataset because few freely available VQA datasets combine broadcast content with the large variety of subject matter needed for NR metric research.

C. Notation and Statistics

Throughout this report, we will compare MOS to the estimated MOS from an NR metric ($\widehat{\text{MOS}}$). Our primary statistic is Pearson correlation coefficient (ρ) between MOS and $\widehat{\text{MOS}}$, because Pearson correlation coefficient is usually reported by prior publications. We use $\widehat{\text{MOS}}$ directly as output by the NR metric; we do not apply a logistic fit to the dataset's MOSs. If the NR metric fails for some media, those data points will be omitted from our calculations.

We will use "dataset" to refer to the data produced by a single experiment (i.e., a set of images or videos with individual subject ratings). No limits are placed on the number of subjects, media, labs, test environment, or rating method. Most of the datasets mentioned in this report were conducted with the 5-level Absolute Category Rating (ACR) method. We cannot currently recommend any techniques for combining multiple ACR datasets into a superset for NR metric research.

III. SUBJECTIVE TESTING

Ultimately, the accuracy of an NR metric depends on the internal validity of the datasets used for training and testing.

A. Accuracy Limitations and Increasing Data Requirements

A study of subject rating behaviors [87] shows that subjects' scoring is a random process. This is expected behavior that must be accepted; not a flaw or fault that can be eliminated.

The VQEG *MM2* dataset studied the impact of test environment on subject ratings [77]. Ten subject pools were collected from six labs under various environmental conditions. The analyses predicted lab-to-lab Pearson correlation coefficients in the range of 0.90 to 0.99 for 15 subjects (as per ITU-R Rec. BT.500) and in the range of 0.95 to 0.99 for 24 subjects (as per ITU-T Rec. P.913). The mode is 0.96 and 0.97 for 15 and 24 subjects, respectively. The mode decreases to ≈0.92 if the subjective test spans a narrow range of quality, due to the random error around each MOS.

A more comprehensive analysis of 60 subjective tests appears in [88]. This report uses the Student's *t*-test to analyze statistical differences at the 95% confidence level. "Disagreement" incidents are defined as both labs concluding that media **A** and **B** have significantly different quality, but the MOSs are **A** > **B** for one lab and **A** < **B** for the other lab. The likelihood that two labs will disagree on the rank order of two media is ≤ 1%, for tests with at least 15 subjects. This report also measures the MOS confidence interval (ΔS_{CI}), which is defined as the difference in MOS values at which 95% of the pairs will be statistically different. The following relationships are trends for subjective tests that use the 5-level ACR scale:

1) 24 subjects: $\Delta S_{CI} \approx 0.5$ to 0.7

TABLE III
DATASET LIST

Abbrev.	Name	Notes	Ref.	Year	Resolution	Use Case	Size
AGH / NTIA / Dolby	Experiment design dataset with unrepeated scenes	V T **3**	[53]	2019	1920 × 1080	Broadcast	291
AVA	Aesthetic Image Analysis	I C	[54]	2012	Unknown	72 photography styles	963
BID	Blurred Image Database	I C **1**	[55]	2011	1920 × 1440	UGC (blur)	585
CCRIQ	Consumer Content Resolution and Image Quality	I C **1**	[56]	2015	1920 × 1080, 3840 × 2160	UGC	392
CID2013	Camera Image Database 2013	I C **1**	[57]	2015	1600 × 1200	UGC	475
CCRIQ2 & VIME1	CCRIQ dataset 2, Video and Image Models for Consumer Content Evaluation dataset 1	I C **1**	[58]	2020	1440 × 900	UGC	92 & 102
CVD2014	Camera Video Database	V C	[59]	2014	640 × 480, 1280 × 720	UGC	234
CVIQD2018	Compressed VR Image Database	I T	[60]	2018	360-degree	UGC	544
DIQA	Document Image Quality Analysis	I E	[61]	2016	1840 × 3264	Document scanning	176
ESPL-LIVE	Embedded Signal Processing Laboratory LIVE HDR Image Database	I M	[62]	2017	4288 × 2848	Multiexposure fusion and tonemapping	1811
ETRI-LIVE STSVQ	ETRI-LIVE Space-Time Subsampled Video Quality Database	V T	[63]	2022	3840 × 2160	Broadcast	437
FocusPath	Medical image database of digital pathology for natural sharpness	I C	[19]	2019	1080 × 1080	Medical: automatic slide focusing	864
ITS4S	ITS 4 Second Dataset	V T **3**	[32]	2018	1280 × 720	Broadcast, public safety	813
ITS4S2	2nd in *ITS4S* dataset series	I C **1**	[64]	2019	1920 × 1080	UGC and public safety	1473
ITS4S3	3rd in *ITS4S* dataset series	V C **2**	[65]	2019	1920 × 1080	Public safety	596
ITS4S4	4th in *ITS4S* dataset series	V C **2**	[66]	2019	1920 × 1080	Public safety (camera pans)	198
IQA UGC	Image Quality Analysis User Generated Content	I C	III.B	—	Various	UGC	6 datasets
KoNViD-1K	Konstanz Natural Video Database	V C **2**	[67]	2017	960 × 540	UGC	1200
KonIQ-10k	Konstanz Natural Image Quality Database of ≈10,000 images	I C	[68]	2020	224 × 224 & 1024 × 768	UGC	10373 & 10373
KonVid-150k	Konstanz Natural Video Database of ≈150,000 videos	V C	[69]	2021	960 × 540	UGC	1566 & 152265
LIVE-2006	2006 LIVE Image Quality Assessment Database	I T S	[70]	2006	≈768 × 512	Simulated UGC	779
LIVE 3D VR IQA	LIVE 3D Virtual Reality IQA Database	I S T	[71]	2020	7680 × 3840	UGC for 3D VR	465
LIVE-Qual	LIVE-Qualcomm Mobile In-Capture Video Quality Database	V C	[72]	2018	2048 × 1152	Broadcast	210
LIVE-VQC	LIVE Video Quality Challenge database	V C	[73]	2019	Various	UGC, mobile phones	585
LIVE-Wild	LIVE Public-Domain Subjective In the Wild Image Quality Challenge Database	I C **1**	[74]	2016	500 × 500	UGC, mobile phones	1162
LIVE-Wild-Compressed	LIVE Wild Compressed Picture Quality Database	I C T	[6]	2019	500 × 500	UGC	400
LIVE-YouTube-HFR	LIVE-YouTube High Frame Rate dataset	V T	[75]	2021	1920 × 1080, 3840 × 2160	Broadcast	480
MD-Derm	Multiple Distortiones Dermoscopy Images	I S	[7]	2016	Unknown	Medical, dermoscopy	450
Micro-Video	Micro-Video UGC Dataset	V C	[76]	2020	Various	Mobile phones	121
MM2	VQEG Multimedia 2	V T	[77]	2012	640 × 480	Broadcast	64
Panorama	Comparative Study of Algorithms for Realtime Panoramic Video Blending	V M	[78]	2018	4000 × 2000	Panoramic video blending for VR	42
SIQD	Surveillance Image Quality Database	I C	[79]	2018	Various	Surveillance	500
SRID	Super-resolution Reconstructed Image Database	I M	[80]	2017	Various	Super-resolution	360
UHD-HDR-WCG	Waterloo Ultra High Definition High Dynamic Range Wide Color Gamut Database	V T	[81]	2019	3840 × 2160	Broadcast	154
YouTube-UGC	YouTube User Generated Content Dataset	V C T	[82], [83]	2019	Various	UGC	1500
VQA BC	Video Quality Analysis of Broadcast Content	V T	III.B	—	Various	Broadcast	3 datasets
VQA UGC	Video Quality Analysis, User Generated Content	V C	III.B	—	Various	UGC	3 datasets
vqegHDcuts	vqegHDcuts	V T E **3**	[84]	2010	1920 × 1080	Broadcast	2145
Youku-V1K	Youku internet video quality assessment database	V C T	[85]	2021	1920 × 1080	Broadcast and UGC	1061

2) 15 subjects: $\Delta S_{CI} \approx 0.7$ to 1.0
3) 9 subjects: $\Delta S_{CI} \approx 1.1$ to 1.4
4) 6 subjects: $\Delta S_{CI} \geq 1.5$

These values provide a lower limit to expected performance, based on well-designed experiments conducted by the ITU and VQEG. Deviations from this ideal produce larger values of ΔS_{CI} for the given numbers of subjects, as may unknown factors.

The implication for NR metric training is that MOSs have limited accuracy. If the Pearson correlation coefficient between

MOS and $\widehat{\text{MOS}}$ is $0.96 < \rho \leq 1.0$, the NR metric is probably overtrained; and $0.90 < \rho \leq 0.96$ is an extraordinary claim that must be justified by overwhelming proof. These thresholds are informed by analyses of subject ratings in [77], [87], and [88].

To develop an NR metric, the researcher must design an experimental NR metric and compare the metric values to MOSs. The results of one trial feeds into the next. This cycle of multiple comparison tests steadily increases the likelihood of concluding that a defective idea has merit (i.e., type-1 error). To compensate, we must develop and evaluate NR metrics with a lot of subjective data. See [89] for details.

Ultimately, ρ cannot prove whether an NR metric behaves similarly to a subjective test; we cannot determine a minimum performance threshold. A solution is proposed in [88], where statistics gathered from 60 subjective tests and 90 lab-to-lab comparisons are used to conclude whether an NR metric is equivalent to a subjective test. The metric's confidence interval (CI) is computed, so that the user can make statistically significant decisions. We will not use these statistics, because they are not intended for comparisons between metrics. Code implementing these statistics is available at [4].

B. Impact of Dataset Design on NR Metrics

The ability of a dataset to characterize a media system depends on the subject matter depicted. Common subject matter selection strategies are convenience sampling, systematic selection, and maximum variety. Convenience sampling uses media conveniently available, which produces biased results (e.g., see the analysis of *VIME1* in [58]). Datasets that use convenience sampling are not always explicitly labeled as such. *AGH/NTIA/Dolby*, CCRIQ, and *CCRIQ2* use the systematic selection criteria from [86]. Variables include textures, shapes, colors, object size, in-scene motion, camerawork, lighting, focal distance, depth of field, camera viewpoint, and unusual characteristics (e.g., ramped color, multiple objects moving in an unpredictable manner). The maximum variety strategy leverages random chance and large pools of subject matter (e.g., *AVA*, *BID*, *KoNViD-1K*, *KonVid-150k*, *ITS4S2*, and *LIVE-Wild*). Some datasets combine convenience sampling with maximum variety (e.g., ITS4S, *ITS4S3*, and *vqegHDcuts*).

The media system itself must be characterized with equal care. Variables include the camera, encoder, transmission system, decoder, and monitor. A single software encoder cannot demonstrate the visual differences produced by encoders from different manufacturers. Software encoders and simulated impairments rarely match the visual response of hardware codecs and camera capture. Example strategies from worst to best in terms of external validity (ability to characterize real applications) are one software codec (ITS4S), convenience sampling of multiple cameras (*ITS4S3*, *ITS4S4*), and a systematic selection of cameras (CCRIQ).

The most impactful design decision is the use case. Most of the datasets in Table III contain user generated content (UGC) for entertainment purposes. Some datasets provide insights into other use cases, like optical character recognition (*DIQA*), medical (*FocusPath* and *MD-Derm*), public safety (*ITS4S3*), video surveillance (*SIQD*), and service quality for video on demand (*vqegHDcuts*, ITS4S, and *AGH/NTIA/Dolby*).

Discussions during VQEG meetings indicate that the use case with the highest demand but fewest datasets is live services for broadcast applications. For example, a professional broadcast studio produces high quality news or sporting event videos for live streaming. The studio production is typically high quality but could include some UGC content (e.g., remote news crews) or variability from weather, lighting, and bandwidth limitations from the field to the studio. High footage costs hinder academic research.

Datasets with conventional experiment designs, like *LIVE-2006* [70], avoid media with camera impairments. These experiment designs reflect the perspective that $\widehat{\text{MOS}}$ should only assess the quality of the transmission system. Thus, $\widehat{\text{MOS}}$ should ignore aesthetics, subject matter, and camera capture. Several impactful industry use-cases support this viewpoint (e.g., a quality feedback loop when transcoding broadcast videos). Consequently, research often begins with the supposition that a trustworthy NR metric can be developed from datasets that characterize the transmission system.

The opposing perspective is that $\widehat{\text{MOS}}$ must assess all impairments, so that $\widehat{\text{MOS}}$ tracks the user's ad-hoc assessments of quality. Users may reject an NR metric if $\widehat{\text{MOS}}$ does not reflect their intuition of the media's overall quality. Our knowledge of human factors supports this viewpoint. MOSs are influenced by aesthetics, subject matter, and camera impairments—especially at bitrates used by modern video systems where compression artifacts are subtle.

To complicate matters, different applications define "good quality" differently. Broadcasters ignore some impairments, or rather consider them to be artistic intent that must be retained—like muted color and dark night scenes. Our prior analysis of public safety content indicates that first responders place a higher than usual importance on vibrant colors [4].

Task specific concerns impact how first responders describe the quality of media [90] and by consequence may subtly impact MOSs. If a bodycam is more sensitive than the human visual system, this could be good sometimes (e.g., a remote viewer can understand events) and bad other times (e.g., a jury incorrectly concludes that the first responder saw events that were not visible at the time). Detectives can reach invalid conclusions if a video surveillance recording changes shapes, motion, or colors. First responders who participated in the *ITS4S3* and *ITS4S4* subjective tests told us that their primary concern was whether they could extract a high quality still frame, to serve as evidence.

Each dataset contains bias and noise from design decisions [91] around use case, subject matter, impairment creation, dataset size, and number of subjects. NR metrics inherit the bias and noise of their training datasets. This can cause an NR metric to respond very differently during training, testing, validation, and application (by third parties).

A mitigation strategy is to combine multiple datasets into a meta-dataset using anchor conditions and a reference test [92]. The *vqegHDcuts* dataset uses such a method to merge multiple VQEG datasets [84]. In addition to reducing

bias and noise, meta-datasets simplify the NR metric training process. However, [93] challenges the concept of anchor conditions or a "common set" in cross-lab experiments. Inclusion of a common set impacted both the evaluation of the common set and the evaluation of the media in the new experiment. We recommend supplementing meta-dataset analyses with analyses of the individual datasets.

C. Ideal Dataset for NR Metric Research

In [32], we describe discrepancies between the experiment designs commonly used for subjective tests and the needs of NR metric research. We conclude that the optimal dataset for training an NR metric for modern camera systems will:

- Contain a huge variety of subject matter
- Include camera impairments
- Portray all state-of-the-art camera applications
- Assess various display devices
- Implement an unrepeated scene design [94]
- Exclude outdated impairments
- Exclude temporal integration
- Exclude transmission errors
- Contain images or videos of 4 s duration

This experiment design is informed by ATIS, VQEG, and ITU validation tests of video quality metrics (see the author biography). In the unrepeated scene design, each subject views each source media once. The goal is to characterize the diverse responses of popularized media systems. For example, *ITS4S3* contains faux public safety media that demonstrate application specific problems, like camera jiggle and inclement weather. The subjects (a mixture of first responders and people in related fields) were able to express task specific requirements on the 5-level ACR scale, without the complexities and limitations of ITU-T Rec. P.912 recognition tests. This experiment design maximizes the variety of subject matter and impairments, which minimizes the likelihood that an NR metric will behave erratically when tested on new scenes or a different manufacturer's codec.

We recommend postponing study of excluded impairments. Outdated impairments are excluded because they could mislead machine learning. Temporal integration is excluded because it can be studied separately and applied as post processing (e.g., how to estimate the overall quality of a movie from immediate quality impressions gathered each second). The maximum video duration results logically from the exclusion of temporal integration. Subjects can comfortably rate 4 second videos, as demonstrated by the ITS4S dataset [32], but the pre-test subjects did not feel comfortable rating 3 second videos. Transmission errors are extremely challenging for full reference (FR) metrics. NR solutions may require supplementary network data or advanced support tools (e.g., object detection).

As datasets diverge from this ideal, the NR metric developer is increasingly likely to miss a critical factor. This can cause the NR metric to produce wildly inaccurate $\widehat{\text{MOS}}$ for subject matter and impairments that do not appear in their training data. For example, CCRIQ [56] reveals whether an NR metric correctly emulates the relative perceptual impact of HD and 4K monitors, because subjects rated images at both monitor resolutions. Most datasets do not model the small MOS difference between HD and 4K monitors. This difference is ≈0.2 MOS for high quality images and ≈0.0 MOS for low quality images [56].

IV. NR Metric Strategies and Impediments

A. NR Metric Development Strategies

We will now move from considerations of quality to a review of the various algorithm development strategies used by NR metric developers. Some NR metrics, like *VIQET* [50]–[51] and *Sawatch* [3], deploy multiple strategies and combine the outputs of multiple NR metrics.

The first strategy is to extract simple statistics from the media. We will refer to these as simple structural pattern (**SSP**) metrics. The most prominent **SSP** metrics are *SI* and *TI* [43], which characterize videos in a subjective test. Because industry continues to rely on *SI* and *TI*, VQEG is developing a proposal to update ITU-T Rec. P.910 to clarify *SI* and *TI* ambiguities that stem from recent technology advances. Other **SSP** metrics include *AGWN* [8], *Entropy Noise* [17], and *LBP* [29]. *NR-IQA-CDI* calculates five **SSP** statistics from the luma plane (mean, standard deviation, skewness, kurtosis, and entropy) but does not combine these into an overall quality estimate.

The second strategy applies the theory of natural scene statistics (NSS) from [95] to identify structural patterns or irregularities in the media that characterize compression or other artifacts. These metrics transform the image, extract statistics, and then apply machine learning. We will refer to these as machine learning NSS (**ML-NSS**) metrics, to avoid confusion with the *NSS* metric [39]. *NIQE* [34] uses a circularly-symmetric Gaussian weighting function and a multivariate Gaussian model. *2stepQA-NR* [6] and *NIQE-K* [35] combine *NIQE* with other algorithm components. *BRISQUE* [10] uses mean subtracted contrast normalized (MSCN) coefficients. *PIQE* [41] takes inspiration from *NIQE* and *BRISQUE*, using both circularly-symmetric Gaussian weighting function and MSCN. *SpEED-NR* [44] uses a Gaussian scale mixture (GSM) model. *ADMD* [7] and *JP2KNR* [22] use wavelets. *Log-BIQA* [30] uses Gradient Magnitude and Laplacian of Gaussian (LOG). *OG-IQA* [40] uses the gradient orientation and magnitude. *NSS* [39] uses the five statistics from *NR-IQA-CDI* [37].

The third strategy is to mimic characteristics of the human visual system (HVS). We will refer to these as **HVS** metrics. *CPBD* [13] models human perception of localized blur. *JNB* [21] relies upon heuristics obtained from a subjective test that characterizes the response of the human visual system to blurriness. *MaxPol* [31] and *HVS-MaxPol* [19] model the relative sensitivity of the human visual system to image blur, using a convolutional filter. *NR-PWN* [38] applies a perceptual noisiness model.

The fourth strategy is to detect a single impairment. We will refer to these as **RCA** metrics. Guidance on training **RCA** metrics appears in [3]. The **RCA** strategy is often used in conjunction with **HVS** or another strategy. Examples include *ADMD* (uneven illumination for dermoscopy images), *AGWN*

(noise), and *MaxPol* (blur). *TDME* [46] and *TDMEC* [47] use a discrete cosine transform (DCT) to detect contrast enhancement. *BTMQI* [11] detects tone-mapped images (i.e., converted from high dynamic range to low dynamic range). *NoRM* [36] detects 3D rendering artifacts. The *Key Indicators* [23]–[28] are a set of 15 **RCA** metrics that detect blackout (all picture content lost), blockiness, block loss, blur, contrast, exposure, flickering, freezing, interlacing, letter-boxing, noise, pillar-boxing, slicing, spatial activity, and temporal activity. *Sawatch* version 3 uses a set of eleven **RCA** metrics.

The fifth strategy is to train the NR metric using empirical data from which the relative ranking of two media can be inferred. We will refer to these as ranking (**RANK**) metrics. The resulting metric may have scope limitations, such as only allowing comparisons among different transcodings of a single media. Metric *dipIQ* [16] is trained on data from a FR metric, which was fed into a pairwise learning to rank (L2R) algorithm. The authors also propose performance assessment statistics.

The sixth strategy is to assess media quality based on success or failure of a specific task. Examples include the likelihood that computer vision (CV) will succeed or fail (*iAITech-NJIT*, *DIQA* dataset) and automatic focusing of digital pathology slides (*HVS-MaxPol*). We will refer to these as **TASK** metrics.

The authors of *HVS-MaxPol* [19] provide another perspective on NR metric development strategies. The authors evaluate 30 NR-IQA metrics published between 2002 and 2018 that detect sharpness vs blur. These **RCA** metrics are categorized by run speed and algorithm development approach (i.e., learning-based, gradient map, contrast map, wavelet, phase coherency, luminance map, total variance, and singular value decomposition). They observe that most of these NR metrics have acceptably high accuracy but unacceptably poor computational speeds.

B. Motivation for Scope Limitations

Researchers eliminate variables to focus their efforts, and this can increase the likelihood of success. NR-IQA metric research eliminates motion and requires fewer computing resources. The same modern cameras and displays are used to create and consume UGC, so NR-IQA metrics can in theory be extended to perform well for NR-VQA. Ad-hoc support for this theory can be found later in this paper, by comparing the performance of NR-IQA metrics on IQA UGC, *VQA UGC*, and *VQA BC*.

The most popular strategy is to limit the impairments. **RCA** metrics take this to the extreme of allowing only a single impairment. Numerous NR-IQA metrics limit their scope to the *LIVE-2006* dataset's [70] impairments, which are JPEG compression, JPEG2000 compression, and three simulated impairments—white noise, Gaussian blur, and a fast-fading Rayleigh channel (FF)—to simulate bit-errors during transmission over a wireless channel. This dataset was indispensable for early NR-IQA research.

All datasets become less relevant over time. For example, the *LIVE-2006* dataset [70] dataset has undesirable characteristics for ongoing NR metric research. White noise and Gaussian blur do not look like the noise and blur produced by camera capture. Modern transmission systems do not produce bit-errors. The image resolution (typically 768 × 512 pixels) is low by today's standards. Camera technology has advanced rapidly since 2006, so even the dataset's high-quality images may differ in subtle ways from high-quality images captured by modern cameras.

An alternate strategy is to limit the subject matter depicted. Sometimes, this is an unintentional consequence of training on a single dataset that contains limited subject matter. *VIQET* [50]–[51] contains four different NR-IQA metrics, one for each allowed subject matter: flat surface, landmark at night, landscape with good lighting, and still life. *VIQET* was trained on the CCRIQ dataset [56], which includes photographs from a variety of modern cameras (phones, tablets, compact cameras, and DSLR cameras).

Subject matter limitations may also reflect the needs of a specific use case. The *DIQA* [61] dataset contains scanned documents and simulated "ratings" that assess the likelihood that optical character recognition (OCR) will succeed, by comparing the original document with the text produced by OCR. *ADMD* [7] limits the scope to dermoscopy images (skin lesions). *NIQE-K* [35] models the opinion of radiologists when viewing ultrasound images. The *ITS4S3* dataset [65] depicts subject matter used by first responders: crime scenes, fireground, prison riots, search and rescue, and cityscapes.

Niche use cases have added challenges around privacy concerns, access to media, subject recruitment, and rating method (e.g., how to ask experts about the usability of images for their task). The tasks performed may have media quality requirements that differ from the default consumer camera settings. First responders and medical professionals could greatly benefit from NR-IQA and NR-VQA metrics that would let cameras understand and respond to these user requirements.

NR metrics with limited scopes could theoretically be updated with an expanded scope. Retraining is particularly important for **ML-NSS** metrics, and MATLAB offers tools to re-train *NIQE* [34] and *BRISQUE* [10].

Users wantonly ignore scope limitations. Thus, the perceived accuracy of an NR metric depends on its response to both in-scope and out-of-scope media. Users expect the NR metric's performance to degrade gracefully as the media stray increasingly beyond the intended scope. We expect $\widehat{\text{MOS}}$ to become less accurate, but random values are unacceptable.

C. NR Metrics Analyzed on Modern Cameras

Table IV, Table V, and Table VI summarize the accuracy of NR metrics for modern camera systems, as reported in a variety of publications. These analyses usually appear as a side comment within a publication that announces a new dataset or NR metric.

The first two columns contain the NR metric's name and the Pearson correlation coefficient (ρ) or range of coefficients reported by the metric developer. See Table I and Table II for these references.

TABLE IV
ACCURACY OF *NR METRICS* FOR MODERN CAMERA SYSTEMS, PART 1

Developer		Evaluator Assessment			
NR Metric	**ρ**	**ρ**	**Dataset**	**Ref**	**Notes**
BLIINDS-II	0.93	0.11	*KonVid-10K → LIVE-Wild*	[69]	Retrain
		0.18	*LIVE-2006 → LIVE-Challenge*	[18]	Retrain
		0.21	*SIQD*	[79]	Fit
		0.38	*CID2013*	[1]	Fit
		0.45	*ESPL-LIVE*	[62]	Retrain Fit
		0.60	*KonVid-10K*	[69]	Retrain
BRISQUE	0.90, 0.94	0.11	*YouTube-UGC*	[83]	
		0.20	*vqegHDcuts*	[84]	
		0.23	6 datasets	[84]	
		0.25	*BID*	[99]	Fit
		0.27	*Panorama*	[100]	Fit
		0.33	*LIVE-2006 → LIVE-Challenge*	[18]	Retrain
		0.33	*SIQD*	[79]	Fit
		0.34	*ETRI-LIVE STSVQ*	[63]	Fit
		0.34	*ETRI-LIVE STSVQ*	[63]	Fit Retrain
		0.36	*UHD-HDR-WCG*	[81]	Fit
		0.38	*LIVE-YouTube-HFR*	[75]	Retrain
		0.40	*YouTube-UGC*	[85]	Retrain
		0.44	*ESPL-LIVE*	[62]	Retrain Fit
		0.45	*CID2013*	[57]	Misc
		0.48	*CID2013*	[99]	Fit
		0.49	*CID2013*	[57]	
		0.50	*ITS4S*	[84]	
		0.57	*CVD2014*	[59]	Fit
		0.57 0.60	*3 datasets*	[48]	
		0.58	*Live-Qualcomm*	[72]	Retrain
		0.59	*3 datasets*	[84]	Retrain
		0.60	*KonVid-10K → LIVE-Wild*	[69]	Retrain
		0.62	*CID2013*	[1]	Fit
		0.64	*LIVE-VQC*	[73]	Retrain Fit
		0.64	*LIVE-VQC*	[85]	Retrain
		0.66	*KoNViD-1K*	[85]	Retrain
		0.67	*SRID*	[80]	
		0.71	*KonVid-10K*	[69]	Retrain
		0.76	*CVIQD2018*	[60]	Fit
		0.78	*Youku-V1K*	[85]	Retrain
		0.83	*LIVE 3D VR IQA*	[71]	Retrain Fit
		0.90	*LIVE-Wild-Compressed*	[6]	Retrain Fit
C-DIIVINE	0.88–0.94	0.44	*ESPL-LIVE*	[62]	Retrain Fit
		0.47	*LIVE-2006 → LIVE-Challenge*	[18]	Retrain
CPBD	0.91	0.20	*SIQD*	[79]	Fit
		0.26	*CID2013*	[1]	Fit
		0.27	*BID*	[99]	Fit
		0.39	*CVD2014*	[59]	Fit

TABLE V
ACCURACY OF *NR METRICS* FOR MODERN CAMERA SYSTEMS, PART 1I

Developer		Evaluator Assessment			
NR Metric	**ρ**	**ρ**	**Dataset**	**Ref**	**Notes**
DIIVINE	0.92	0.22	*SIQD*	[79]	Fit
		0.23	*CID2013*	[57]	Misc.
		0.33	*BID*	[99]	Fit
		0.37	*LIVE-2006 → LIVE-Challenge*	[18]	Retrain
		0.43	*SRID*	[80]	
		0.48	*ESPL-LIVE*	[62]	Retrain Fit
		0.48	*KonVid-10K → LIVE-Wild*	[69]	Retrain
		0.53	*CID2013*	[99]	Fit
		0.61	*KonVid-10K*	[69]	Retrain
FRIQUEE	0.63–0.97	0.33	*Panorama*	[100]	Fit
		0.47	*BID*	[99]	Fit
		0.60	*CID2013*	[99]	Fit
		0.70	*LIVE-VQC*	[85]	Retrain
		0.71 0.75	*3 datasets*	[48]	
		0.74	*LIVE-Qualcomm*	[72]	Retrain
		0.75	*KoNViD-1K*	[85]	Retrain
		0.76	*YouTube-UGC*	[85]	Retrain
		0.76	*3 datasets*	[98]	Retrain
		0.85	*Youku-V1K*	[85]	Retrain
NIQE	0.91	0.09	*CVD2014*	[59]	Fit
		0.10	*vqegHDcuts*	[84]	
		0.11	*YouTube-UGC*	[83]	
		0.12	*ESPL-LIVE*	[49]	Retrain
		0.15 0.19	*MD-Derm*	[7]	
		0.21	*ETRI-LIVE STSVQ*	[63]	Fit
		0.21	*Panorama*	[100]	Fit
		0.22	*CID2013*	[57]	Misc
		0.25	*LIVE-YouTube-HFR*	[75]	
		0.28	*YouTube-UGC*	[85]	
		0.32	*6 datasets*	[84]	
		0.34	*CID2013*	[1]	Fit
		0.35 0.41	*3 datasets*	[48]	
		0.38	*CID2013*	[57]	
		0.38	*ETRI-LIVE STSVQ*	[63]	Fit Retrain
		0.46	*BID*	[99]	Fit
		0.47	*LIVE 3D VR IQA*	[71]	Fit
		0.48	*SIQD*	[79]	Fit
		0.48	*3 datasets*	[98]	
		0.52	*SRID*	[80]	
		0.53	*CVIQD2018*	[60]	Fit
		0.53	*UHD-HDR-WCG*	[81]	Fit
		0.54	*ITS4S*	[84]	
		0.55	*KoNViD-1K*	[85]	
		0.58	*LIVE-VQC*	[73]	Fit
		0.60	*Youku-V1K*	[85]	
		0.63	*LIVE-VQC*	[85]	
		0.66	*CID2013*	[99]	Fit
		0.68 0.68	*LIVE-Qualcomm*	[72]	Fit
		0.84	*LIVE-Wild-Compressed*	[6]	Retrain Fit

The next four columns contain information from independent assessments of the NR metrics. Column "ρ" is the Pearson correlation coefficient from the reference noted in column "Ref." Column "Dataset" identifies the dataset used for the analysis, or the number of datasets if more than one dataset is used. Occasionally, the authors retrain the metric using dataset $\boldsymbol{A}$ and test on dataset $\boldsymbol{B}$. We show this as ($\boldsymbol{A} \rightarrow \boldsymbol{B}$). Our preliminary analysis [84] uses six UGC datasets that mix IQA and VQA: *BID*, CCRIQ, *CCRIQ2&VIME1*, *CID2013*, *KoNViD-1K*, and *LIVE-Wild*. Similarly, [48] uses three datasets (*KoNViD-1K*, LIVE-Qualcomm, and *CVD2014*) and [98] uses three datasets (*KoNViD-1K*, *LIVE-VQC* and *YouTube-UGC*), which they refer to as *UGC-VQA*.

TABLE VI
ACCURACY OF *NR METRICS* FOR MODERN CAMERA SYSTEMS, PART 1II

Developer		Evaluator Assessment			
NR Metric	**ρ**	**ρ**	**Dataset**	**Ref**	**Notes**
QAC	0.86	0.11	*MD-Derm*	[7]	
		0.17	*CID2013*	[99]	Fit
		0.32	*BID*	[99]	Fit
		0.87	*CVIQD2018*	[60]	Fit
SSEQ	0.94	0.05	*Panorama*	[100]	Fit
		0.29	*KonVid-10K* → *LIVE-Wild*	[69]	Retrain
		0.61	*KonVid-10K*	[69]	Retrain
TLVQM	0.77–0.85	0.29	*LIVE-YouTube-HFR*	[75]	Retrain
		0.42	*ETRI-LIVE STSVQ*	[48]	Retrain Fit
		0.56	*Micro-Video*	[76]	
		0.66	*YouTube-UGC*	[85]	Retrain
		0.68	*KONvID-150K*	[69]	Retrain
		0.77	*KoNViD-1K*	[85]	Retrain
		0.78	*Youku-V1K*	[85]	Retrain
		0.80	*LIVE-VQC*	[85]	Retrain
V-BLIINDS	0.88, 0.75	0.12	*CVD2014*	[59]	Fit
		0.40	*LIVE-YouTube-HFR*	[75]	Retrain
		0.58	*YouTube-UGC*	[85]	Retrain
		0.64 0.66	*3 datasets*	[48]	
		0.66	*3 datasets*	[98]	Retrain
		0.67	*Live-Qualcomm*	[72]	Retrain
		0.68	*KonVid-150k*	[69]	Retrain
		0.72	*LIVE-VQC*	[73]	Retrain Fit
		0.72	*LIVE-VQC*	[85]	Retrain
		0.78	*Youku-V1K*	[85]	Retrain
VIIDEO	0.65	0.05	*vqegHDcuts*	[84]	
		0.10	*Live-Qualcomm*	[72]	Fit
		0.14	*LIVE-VQC*	[73]	Fit
		0.15	*YouTube-UGC*	[83]	
		0.15	*YouTube-UGC*	[85]	Retrain
		0.21	*LIVE-VQC*	[85]	Retrain
		0.26	*Micro-Video*	[76]	
		0.30	*KoNViD-1K*	[85]	
		0.32	*ITS4*	[84]	
		0.41	*Youku-V1K*	[85]	
		0.62	*Panorama*	[100]	Fit

Column "Notes" summarizes any procedures used other than simply correlating $\widehat{\text{MOS}}$ to MOS. "Retrain" means a machine learning metric was retrained and analyzed on the dataset (e.g., with an 80/20 split). "Fit" means $\widehat{\text{MOS}}$ was fitted to MOS using a non-linear mapping. "Misc." refers to other miscellaneous processing. Information could be missing from this column; some publications did not describe their test procedures clearly.

Additional NR metric assessments can be found in the documents cited in Table IV, Table V, and Table VI. These tables focus on NR metrics that are analyzed by multiple publications.

Most of these assessments use a single dataset. Likewise, most of the NR metrics are trained on a single dataset (see Table I). This results in a huge range of ρ values. For example, *BRISQUE* analyses ranges from 0.11 to 0.90, and *NIQE* analyses ranges from 0.09 to 0.84. These examples make it clear that ρ for any single dataset cannot be interpreted as an indicator of ρ outside that dataset.

One of the few evaluations that uses many datasets appears in [19]. This paper compares *HVS-MaxPol* to seven other sharpness vs blur metrics. The authors use four datasets with synthetic blur and three datasets with camera capture blur. Their meticulous analysis includes a table that allows easy comparisons among the four synthetic datasets (*LIVE-2006* and three others) and the three camera capture datasets (*BID*, *CID2013*, and *FocusPath*).

The primary issue we observe is insufficient data—development and evaluation based on a single dataset or datasets that are too similar to each other. These very narrow results can then establish unrealistic expectations for more general NR metric performance. Evaluators analyze NR metrics on tiny "proof of concept" datasets and imply that their results (good or bad) will extend to a broader evaluation of modern media systems. Derivative issues follow—brilliant ideas discarded, erroneous ideas pursued, and widespread misinformation about the accuracy of NR metrics.

The choice to train or test on a single dataset cannot be justified. Better, faster, and more reliable results can be obtained with multiple datasets—some in-scope, to improve internal validity, and some out-of-scope, to ensure external validity. Many datasets are now freely available: 25 datasets from LIVE (see [96]), 9 datasets from the Universität Konstanz (see [97]), 37 datasets on the Consumer Digital Video Library (https://www.cdvl.org), etc. The metric's internal validity can be improved by including datasets that focus on a specific application. Each of the following datasets emulates a different use case: *AVA*, *DIQA*, *FocusPath*, *ITS4S3*, *LIVE-YouTube-HFR*, *Panorama*, *SIQD*, *SRID*, *UHD-HDR-WCG*, and *YouTube-UGC*.

A secondary issue is measurement noise from differences in analysis methods. The impact can be observed by comparing results for the *CID2013* dataset from different papers. Different publications report different ρ values for the *CID2013* dataset (e.g., *BRISQUE* [0.45 to 0.62], *NIQE* [0.22 to 0.66], and *DIIVINE* [0.23 to 0.53]). This makes it very difficult to reach any viable conclusions.

The most common method variants are fitting and retraining. The choice to fit $\widehat{\text{MOS}}$ to MOS is influenced by VQEG validation tests. The VQEG validation tests are designed for high performing metrics that have a linear response to MOS. The logistic fit removes subtle nonlinearities associated with the subjective dataset. However, NR metrics are much less accurate. The logistic fit disguises the NR metric's nonlinearity problems, which is undesirable. We recommend against fitting functions when analyzing NR metrics.

Retraining is a confounding factor because each evaluator retrains the NR metric differently. Retraining requirements may hinder the adoption of an NR metric. Evaluators should either analyze the NR metric exactly as provided by the developer or provide two analyses—first without retraining and second with retraining. The first analysis would provide baseline statistics for comparisons between datasets. The second analysis would demonstrate the NR metric's potential improvement for the new dataset.

Since no single publication provides us with stable accuracy measurements for NR metrics applied to modern camera

systems, we will infer a threshold using the average accuracy across multiple tests. If an author provides multiple estimates, these will be averaged. Statistics from developers are ignored; these are usually the metric's performance on the training data. Taking the average of correlation values (denoted $\overline{\rho}$) is suspect from a mathematic theory standpoint, but we have no viable alternative.

For *BRISQUE*, $\overline{\rho} = 0.48$ overall and $\overline{\rho} = 0.42$ when retraining is eliminated. For *NIQE*, $\overline{\rho} = 0.39$ overall and $\overline{\rho} = 0.38$ when retraining is eliminated. For the NR metrics in Table VI, $\overline{\rho} = 0.41$ overall and $\overline{\rho} = 0.34$ when retraining is eliminated. Finally, for the seven blur/sharpness NR metrics in [19], $\overline{\rho} = 0.51$. This estimate includes the author's prior work (*MaxPol*) but omit *HVS-MaxPol*, as it was trained on these three modern camera datasets.

Most of these experiments were conducted by universities or our department. *YouTube-UGC* [82], [83] by Google®provides an independent industry assessment of NR metrics for the UGC use case. *YouTube-UGC* contains 1,500 videos that were selected from 1.5 million YouTube videos. Their analyses of *BRISQUE*, *NIQE*, and *VIIDEO* approach the minimum reported accuracy. The NR metric with the best accuracy is *NIMA* [33] with $\rho = 0.53$. Google attributes some of the decreased performance of NR metrics on *YouTube-UGC* to aesthetic quality problems that are outside the NR metrics' intended scope [83]. *Youku-V1K* has a similar design—1,072 videos from the Youku service—but much higher correlations.

V. NR Metric Sawatch: Baseline for Collaboration

We must now interrupt our overview of NR metrics to describe NR metric *Sawatch Version 3* and the RCA metrics upon which *Sawatch* is built. We will use these NR metrics to expose the differences among datasets and the repercussions of these differences for NR metrics.

We begin with the supposition that NR metrics must be trained on a minimum of ten datasets that characterize a variety of modern camera systems and camera capture impairments. This is not an exact calculation. Most researchers must depend on openly available datasets. Ten datasets should ensure a judicious variety of principal investigator, use case, subject matter, experiment design, noise, and bias.

We use functional programming to split the research effort into independent algorithms, each providing RCA for a single impairment. These can be developed separately and replaced with improved algorithms. *Sawatch* is provided as a baseline metric for collaboratively developing NR metrics using this paradigm. Code is available in the *NRMetricFramework* repository [4]. *Sawatch Version 3* can be used for any purpose, commercial or non-commercial. However, *Sawatch Version 3* calls *dipIQ*, which is only freely available for research.

The Sawatch mountain range in central Colorado contains eight of the 20 highest peaks in the Rocky Mountains. Similarly, the *Sawatch* metric is a collection of NR metrics and RCA parameters. Mountain climbers tackle increasingly difficult mountains. Similarly, NR metric development is a difficult challenge, and our goal is steady improvement until we achieve the highest levels of performance.

A. Background

Sawatch builds upon the development methods we used from 1989 to 2011 to develop FR metrics that can be implemented as reduced reference (RR) metrics. The best known of those are *Video Quality Metric* (*VQM*) [102] from ITU-T Rec. J.244 (2004) and ITU-R Rec. BT.1683 (2004) and *Video Quality Metric for Variable Frame Delay* (*VQM-VFD*) [103].

Our FR/RR design strategy was to develop several different metrics using the **HVS** and **RCA** strategies. These metrics were motivated by the human visual system and provide limited RCA. $\widehat{\text{MOS}}$ is a linear equation that takes these individual metrics as input parameters. *VQM* was trained on 11 datasets and *VQM-VFD* was trained on 79 datasets. Our training leveraged both per-dataset analyses and meta-dataset analyses. The large number of datasets and **RCA/HVS** strategy produced metrics that are resilient to advances in video technology, as demonstrated by [104].

B. Design Principles

NR metrics typically assess overall quality ($\widehat{\text{MOS}}$), but companies tell us that NR metrics must also provide RCA that explains why the quality is bad [3]. Companies want to use NR metrics to detect and respond to problem in real time—adjust camera settings, apply post-processing to remove the impairment, select appropriate encoder settings, change to a more appropriate computer vision algorithm, etc.

Instead of a "one size fits all" solution, industry wants an NR metric that can be easily adjusted—like a muffin recipe that tells the chef how to adjust the recipe for nut muffins, chocolate chip muffins, blueberry muffins, or cheese muffins. NR metrics must provide RCA and, if possible, a simple way for a lay person to adjust the impact of each measured impairment on MOS.

Sawatch is a versioned series of NR metrics that provide RCA, open source, and moderate to fast run speed. The intention is that *Sawatch* will be updated regularly instead of remaining a fixed, static algorithm. *Sawatch* is intended for a broad range of modern camera systems, video content, photography problems, and camera capture impairments. $\widehat{\text{MOS}}$ is calculated as a weighted sum of the other NR metrics, each assessing a single impairment. This equation can easily be adjusted to omit an impairment that users do not wish to be penalized.

To simplify development, we accept the following constraints. First, *Sawatch version 3* cannot assess transmission errors or temporal integration, as per Section III–C. These can be studied separately and applied as post-processing. Second, *Sawatch* assesses the quality of the image or video after scaling to a monitor for display. That is, the added value of a 40 megapixel (MP) photograph over an 8 MP is irrelevant when both are displayed to a 1080 × 1920 pixel monitor.

TABLE VII
RCA PARAMETERS IN NR METRIC *SAWATCH* VERSION 3

Parameter	w_p	Description
S-Black Level	0.75	Black level is too high (pale image)
S-Blockiness	2.40	Inordinate vertical and horiztonal edges
S-Blur	2.40	Entire image is too blurry
S-ColorNoise	1.05	Color problems including sampling noise, color clipping, and post-processing
S-dipIQ	1.80	Compression artifacts
S-Fine Detail	1.50	Fine details have been lost
S-Jiggle	1.50	Camera jiggle
S-PanSpeed	2.40	Camera pans too quickly
S-Pallid	0.15	Colors are too unsaturated
S-SuperSaturated	0.15	Colors are too saturated
S-WhiteLevel	0.75	White level is too low (dark image)

Sawatch version 3 is a linear combination of eleven NR metrics. We will refer to these as parameters. Each parameter analyzes one impairment to provide RCA. *Sawatch* results are on a one to five scale as per the 5-level ACR method. Due to relative differences between datasets and error in the RCA metrics, $\widehat{\text{MOS}}$ is sometimes above or below this range. Each parameter is on a zero to one scale, where zero indicates no impairment and one is a nominal upper limit for the maximum impairment. *Sawatch* has the form:

$$\widehat{MOS} = 6.2 - \sum_{p=1}^{11} \left(w_p x_p\right) \tag{1}$$

where w_p is the weight for parameter p, and x_p is the value of parameter p. The influence of the nth parameter can be removed from $\widehat{\text{MOS}}$ by setting w_n to zero. The expected rating behavior is thus retained: media with few or no impairments will have $\widehat{\text{MOS}} \approx 5.0$. Table VII lists the parameters and their weights.

The constant (6.2) is derived observationally, from our twelve training datasets (IQA UGC, *VQA UGC*, and *VQA BC*). Extensions above five and below one occur whenever multiple datasets are mapped to a single scale. For example, when the six VQEG HD datasets are mapped to a single scale, the MOSs range from 0.82 to 5.26 [101].

We could not use linear regression to determine these weights. Each training dataset yields very different values for w_p, due to differences in the frequency and severity of impairments among datasets. Imperfections in the RCA metrics create dataset dependencies, and the low accuracy of all NR metrics leaves us hesitant to trust meta-data analyses. Instead, we manually adjusted one weight at a time and examined how the accuracy of *Sawatch* changed for each dataset.

For some applications, $\widehat{MOS}$ estimation accuracy is more important than flexibility. Machine learning could be used to replace (1) with an optimal combination of the RCA metrics for general use. This strategy might let us model the complex interactions between impairments that we note in Section VII.

Several factors influence the upper and lower bounds for *Sawatch* $\widehat{MOS}$ on the 12 training datasets (5.1 to 0.0). $\widehat{MOS} < 1$ tend to be outliers but can also be caused by the relative nature of MOSs (i.e., subjects adjust their use of the rating scale to the media in the dataset). The distribution of MOSs is influential: subjects are reluctant to assign a perfect 5.0 MOS to any media, and datasets tend to have few media with MOS < 2. Some impairments cannot be detected (e.g., jerky motion, lens distortion, lens flare, flickering, freezing, and ghosting). Each of the eleven parameters in Table VII has a limited accuracy. *Sawatch* tends to produce values in the middle of the range (roughly 2.6 to 3.8); values at either extreme (near 5.0 or 1.0) are unlikely. As the overall accuracy of *Sawatch* improves with future versions, we expect the distribution of $\widehat{MOS}$ to flatten.

C. Assumptions and Filters

The parameters adhere to the following design specifications. Calculations occur in the YCbCr color space with 8-bit pixel depth. Thus, the luma (Y) plane spans [0..255]. Parameters are scaled to [0..1], where zero indicates no impairment and one indicates maximum impairment. Images and videos are scaled to the monitor resolution prior to beginning calculations.

Spatial impairments are defined for images (photographs) and calculated for each video frame separately. Temporal impairments are calculated on sequential pairs of video frames; images are replicated to create a still video. Per-frame video results are aggregated into a single value, typically the mean of all frames. This aggregation can be replaced with an improved temporal integration algorithm later.

Some parameters divide images into subregions that contain $\approx$1% of the pixels. The results for each subregion are combined into a single estimate, typically focusing on the worst case (high impairment levels) or the best case (low impairment levels). This technique allows us to avoid the impact of confounding visual patterns (e.g., intentionally blurred backgrounds look blurry but may not impact MOS).

Several parameters refer to the spatial information (SI) filter, which forms the core of *VQM* [102] and *VQM-VFD* [103]. We will refer to this edge detection filter as si5 for a 5×5 edge filter, si11 for 11×11, and si15 for 15×15. These are bandpass filters, where each row or column is identical. Like the Sobel filter, the SI filter applies separate horizontal and vertical filters and combines them using Euclidian distance (i.e., square, sum, square root). Larger edge filters, like si15, are fairly impervious to small edges and shot noise. Like Sobel, the SI filter has a $\times 4$ edge magnitude multiplier.

The horizontal and vertical filtered images can be used to compute a more robust calculation of edge angle than is possible with the 3×3 Sobel filter. This angle estimation is used to separate the SI pixels into horizontal and vertical edges (HV) and diagonal edges (HVbar), using an angle threshold, Θ. For more information, see filter_si_hv_adapt.m [4].

D. RCA Metric Training

Our training data consists of the twelve datasets described in Section II–B: IQA UGC, *VQA UGC*, and *VQA BC*. We chose six IQA dataset and six VQA datasets as a compromise between the ideal (more datasets) and the reality of computation resources (storage and computation speed). Our primary

challenge in training **RCA** metrics is that these datasets provide MOSs, not RCA. We used the following RCA metric training strategies. Other strategies are proposed in [3].

Our first strategy is to create a challenge dataset—a set of images or videos that demonstrate a single impairment, while avoiding others. This strategy is used by [55] for **RCA** metrics that detect blur. The authors begin with a dataset of synthetically blurred images and then verify their results using *BID* which contains naturally blurred images. Similarly, *ITS4S4* includes camera pans with different speeds, frame rates, and subject matter. While other impairments could not be fully eliminated, their influence was minimized. The *ITS4S4* dataset was used to train the *S-PanSpeed* metric in *Sawatch Version 3*.

A challenge dataset simplifies algorithm development, because MOS is highly correlated with the quality of the chosen impairment. The disadvantage is that challenge datasets will probably be small and may lack external validity. For example, **RCA** metrics for noise, like *AGWN*, seem to have trouble with unforeseen photographs that contain fine details. Similar dataset design problems cause facial recognition to fail on people wearing certain t-shirts [105]. The **RCA** metric must be verified using other datasets. The expense of creating a challenge dataset limits the viability of this strategy.

Our second and more commonly used strategy is to visually examine scatter plots. Differences in the impairment's prevalence and severity can cause the scatter plots from different datasets to look very different. However, we expect the MOS and $\widehat{\text{MOS}}$ scatter plots for multiple datasets to cover a similar area and depict similar shapes. Multiple impairments influence MOSs, so there is considerable noise around the MOS vs $\widehat{\text{MOS}}$ fit line of an **RCA** metric and we expect low ρ values. Pearson correlation coefficient assumes that the data should form a scattering of points around a fit line. This assumption is only true when the impairment is very common, either in general (like blurriness) or because it is the main impairment of a particular dataset (e.g., blur for *BID* or pan speed for *ITS4S4*).

Our RCA development cycle was as follows. We chose an impairment, brainstormed algorithms with low complexity and fast run speed, and calculated the algorithms for one dataset that contains the impairment. Our analysis included examining statistics (ρ), examining MOS vs $\widehat{\text{MOS}}$ scatter plots, and visually inspecting media, to see whether the algorithm detects the intended impairment. Promising algorithms were iteratively improved, applied to other datasets, and compared to *Sawatch*'s residuals. The iterative improvement cycle is computationally efficient, because [4] provides a mechanism to save and investigate intermediate results of the NR metric calculation. Scatter plots heavily influenced these decisions.

Where possible, we evaluate **RCA** metrics on datasets that include low levels of the impairment. This indicates the **RCA** metric's false positive error rate. For example, the goal of *ADMD* [7] is to detect uneven illumination, but our analysis of datasets without uneven illumination indicates that *ADMD* detects an infrequent characteristic of high-quality media.

The scatter plots for low impairment datasets may have no obvious pattern and misleadingly low ρ. The fit line can change direction (positive correlation to neutral or negative correlation). Thus, low impairment datasets can only be understood in the context of other datasets' scatter plots. This does not indicate a problem if the range of $\widehat{\text{MOS}}$ values is in the range associated with "no impairment" for datasets with low levels of the impairment. For example, *Sawatch*'s *White Level* has ρ values between 0.00 and 0.08 for the three video compression datasets, because the videos were produced by professional videographers who correctly set the camera's white level.

As a final verification step, we visually inspected media. Only by viewing media with high and low $\widehat{\text{MOS}}$ can we know whether the metric assesses the intended impairment.

E. Sawatch Parameters

Let us now examine the nine parameters associated with *Sawatch Version 3*. We describe each parameter at a high level. Our goal is to identify underlying characteristics of the human visual system, not the quirks of a scene, camera, or codec. Omitted algorithm details, such as scaling factors and clipping levels, can be found in [4]. This repository contains scatter plots and additional statistics for each parameter. Each RCA metric is prefixed with "*S-*" to denote the association with *Sawatch*.

S-BlackLevel estimates whether the black level is too high, based on the standard deviation of Y (the luma image). *S-BlackLevel* only triggers when the mean of Y is above mid-level grey.

S-Blockiness analyzes the angle of small edges in the luma plane, using an si5 filter with $\Theta = 0.01$ radians. Put simply, *S-Blockiness* triggers if the entire image has higher than expected HV edge energy, relative to the HVbar edge energy. HV pixels adjacent to HVbar pixels are omitted (set to zero), because the measured edge angle is unreliable there. The image is divided into $\approx$100 subregions. For each subregion, we compute the average HV magnitude divided by the average HVbar magnitude. The denominator is clipped to prevent low magnitude noise from amplifying the ratio. *S-Blockiness* is the average of the low value subregions; this eliminates intentional horizontal and vertical lines (e.g., news feed banner, faux picture frame, picture-in-picture border).

S-Blur analyzes the delta that an Unsharp filter would add to the image. The image is divided into $\approx$100 subregions, and each subregion's average magnitude is divided by the range of filtered values. *Unsharp* averages the high value subregions (i.e., areas with the sharpest, most in-focus edges). A divisor normalizes for differences between low and high contrast content—think lion vs zebra fur patterns. *S-Blur* has a correction factor for 4K monitors.

S-ColorNoise uses quirks of the YCbCr color space to detect color problems. The Cb and Cr color planes do not align to how people think and talk about colors. Thus, we expect edges in the Cb plane to also appear in the Cr plane. Put simply, *S-ColorNoise* triggers when the Cb and Cr planes are too dissimilar. This flags colorful camera noise from low light environments, abnormal colors (e.g., the camera responded poorly to very bright light), and some manual color enhancements.

We apply the si11 filter to the Cb and Cr planes, divide these into ≈100 subregions, and calculate ρ between the si11 filtered Cb and Cr. *S-ColorNoise* averages the high value subregions (giving the benefit of doubt to Cb/Cr differences being legitimate) and clips at an experimentally determined upper limit (Cb/Cr similarity meets or exceeds expectations). Color noise cannot be computed for color deficient images.

Sawatch is the *dipIQ* metric, linearly scaled to [0..1]. We will refer to this simply as *dipIQ* in our plots and tables. As mentioned previously, *dipIQ* uses an L2R algorithm and truth data calculated from an FR metric [16]. Our analysis indicates *dipIQ* is well suited as an RCA metric for compression artifacts. *dipIQ* performs best for ITS4S and *AGH/NTIA/Dolby*, which closely match the scope where FR metrics work best (e.g., professional footage, compression only impairments).

S-FineDetail is Pearson correlation coefficient squared (ρ^2) between the si5 and si15 filtered luma planes. High values (near one) indicate that all small edges are pieces of larger edges. *S-FineDetail* identifies up-sampling, too aggressive noise filtering, and low bit-rate compression that erases fine details.

S-Jiggle estimates camera jiggle. For each pair of frames, we divide the frame into ≈100 subregions to estimate horizontal and vertical motion. The camera jiggle for each frame is computed as the spread of estimates for each subregion. These separate estimates are combined at different levels of temporal granularity to avoid the influence of frame repeats from 3/2 pulldown and frame rate conversions.

S-PanSpeed was trained on dataset *ITS4S4*, which includes pan speeds from very slow to the background crossing the monitor in ≈0.33 s (e.g., bodycams and security cameras). For each pair of frames, we use the ≈100 horizontal and vertical motion estimates from *S-Jiggle*. These separate estimates are combined at different levels of granularity to obtain an overall estimate for motion that is more influenced by horizontal motion than vertical motion. *S-PanSpeed* demonstrates the viability of the challenge dataset strategy.

S-Pallid identifies images that have too little pigmentation (i.e., deficient in color). Artists choose black-and-white media for a variety of reasons, but subject ratings indicate a small but consistent preference for colorful media. The Cb and Cr planes are divided into ≈100 subregions, and *S-Pallid* is the fraction of regions that contain little variation in Cb or Cr, based on the standard deviation of Cb and Cr. *S-Pallid* has an unusually well-defined upper-triangle shape for ITS4S, *ITS4S2*, and *ITS4S3*, which evaluate media quality for public safety use cases. This seems to indicate that color deficiency is an impairment that hinders first responder applications.

S-SuperSaturated detects media whose color saturation was manually boosted beyond typical values. We calculate the fraction of pixels where either Cb or Cr have larger magnitudes than commonly observed in cameras. *S-SuperSaturated* may be associated with a drop in quality, as demonstrated by dataset *KonVid-1K*. However, the other datasets neither support nor convincingly reject this conclusion. The need for additional training data is reflected in a low weight, w_p.

S-WhiteLevel is the 98^{th} percentile of luma values, when dark border regions are ignored. *S-White Level* is undefined if the entire image is dark, because many videos include intentionally black frames. *S-WhiteLevel* is clipped at an experimentally determined upper threshold, where training data indicates quality stops rising.

VI. NR Metric Accuracy For Modern Camera Systems

Previously published analyses of NR metrics contain exaggerations, ambiguities, and inaccuracies. We conclude that NR metrics must be developed and evaluated with at least an order of magnitude more data (i.e., at least 10 datasets). To address these concerns, we will now present our analyses of NR metrics for modern camera systems. Algorithm discrepancies may occur unintentionally. Note that we:

- Do not retrain machine learning algorithms
- Do not apply a non-linear fit to $\widehat{\text{MOS}}$
- Ignore the NR metric's intended scope
- Use freely available NR metric code if possible
- Compare to diverse media from modern camera systems

This protocol emulates an industry user who wants plug-and-play convenience. NR metrics with very slow computation speeds are omitted as impractical for industry use cases.

A. Our Evaluation Methods and Datasets

Our analysis uses the same twelve datasets that were used to train *Sawatch Version 3:* IQA UGC, *VQA UGC*, and *VQA BC* (see Section II–B). Before running the NR metric, the images and video frames are scaled to the monitor resolution, to replicate the subjects' viewing conditions. NR-IQA metrics are applied to videos by calculating per-frame values and then taking the average over all frames. We expect this to be a tolerable strategy for 8 s videos where the quality may change over time (like *KoNViD-1K* and *AGH/NTIA/Dolby*) and an excellent strategy for shorter video with consistent quality over time (like ITS4S, *ITS4S3*, *ITS4S4*, and *vqegHDcuts*).

Our analyses use 90% of media from each dataset; the remaining 10% of media are held in reserve for verifying the performance of future NR metrics. This 90/10 split was performed once and is recorded in the *NRMetricFramework*. We recommend the same 90/10 split be used for all future training and evaluation. Thus, **ML-NSS** metrics would sub-divide the 90% for training and testing.

A few of the NR metrics evaluated in this section were trained on one or two of our twelve evaluation datasets. *Munsell Red* was trained on ITS4S. *HVSMaxPol* was trained on *BID* and *CID2013*. *NSS* was trained on *CID2013*; the other two variants of *NSS* yield similar performance.

Pearson correlation coefficient will not detect undesirable data distribution patterns, like one value of $\widehat{\text{MOS}}$ spanning the full range of MOSs. Therefore, we will also perform visual examinations of MOS vs $\widehat{\text{MOS}}$ scatter plots. A broad scattering of points around a line is always desirable. If the NR metric detects an infrequently occurring impairment, then we would expect a lower triangle (i.e., narrow range for high quality, wide range for low quality). If the NR metric detects an infrequently occurring characteristic of high-quality media, then we

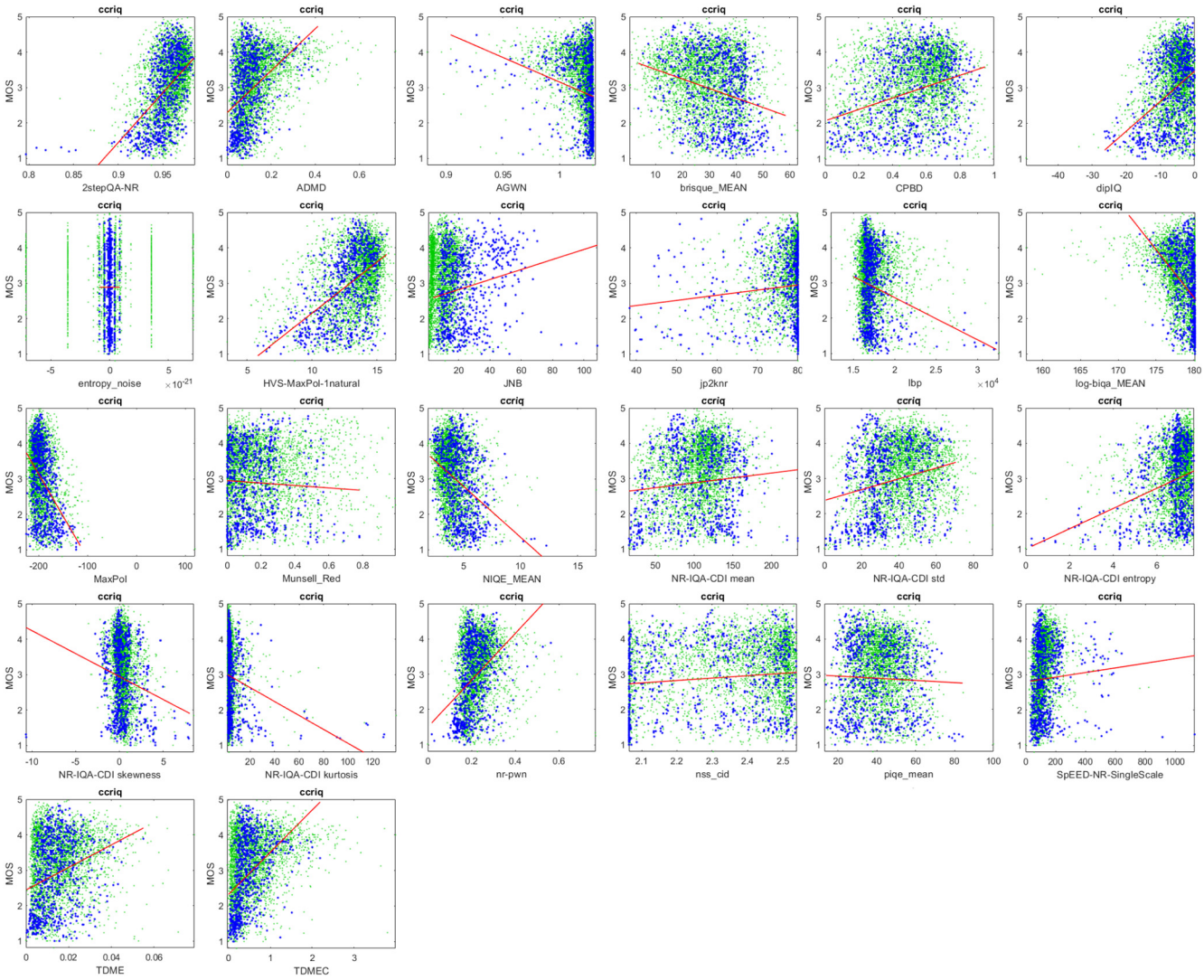

Fig. 1. Scatter plots depict the response of previously published NR metrics to the CCRIQ dataset (blue dots) within the context of the six IQA UGC datasets (green dots). The red line shows the linear fit for the CCRIQ dataset.

would expect an upper triangle (i.e., wide range for high quality, narrow range for low quality). See Figure 1 to visualize the triangle patterns: *TDMEC* is an upper triangle shape, and *dip-IQ* is a lower triangle shape.

B. Pearson Correlation Coefficient Comparisons

Table VIII reports the accuracy of NR metrics when assessing MOS. Column "ω" indicates the intended outcome of the metric: ***M*** for MOS, ***R*** for RCA, ***C*** for CV failure rate, and ***K*** for **RANK** metrics that order media. By failure rate, we mean the NR metric predicts the likelihood that CV will fail due to media quality problems. The intended outcome of the *TDME* and *TDMEC* metrics is ambiguous (RCA or MOS). Table VIII emphasizes RCA metrics because we intentionally sought this type of NR metric.

Column "Shape" indicates the shape of the scatter plots. "//" indicates a scattering of data around a fit line. "▼" indicates an upper triangle. "▲" indicates a lower triangle. "▒" indicates a random scattering with no obvious pattern. "ε" means the scatter plot has severe outliers that must be investigated, or the code produced errors for some media.

Column "ρ" is the Pearson correlation coefficient reported by the NR metric's developer. Column "IQA UGC" reports $\overline{\rho}$ for our six IQA UGC datasets. Column "VQA UGC" reports $\overline{\rho}$ for our three *VQA UGC* datasets. Column "VQA BC" reports $\overline{\rho}$ for our three *VQA BC* datasets. The symbol "$\mathfrak{S}$" indicates values could not be computed because the code runs too slowly to be practical (e.g., 10 m to 4 h per video).

We cannot compute $\overline{\rho}$ for *CurvletQA* because the code produced errors for too many media and some scatter plots depict data scattered around two or more fit lines. Despite this problem, *CurveletQA* is one of the more promising NR metrics based on the underlying shape (see [4] for plots).

The evaluations shown in Table VIII always compare $\widehat{\text{MOS}}$ to MOS but some NR metrics do not produce MOS estimates (see column ω). The mismatch explains some of the decrease in $\overline{\rho}$. This mismatch is most severe for *dipIQ*, which produces rankings instead of MOS estimates. Other statistics and new methods are needed to properly analyze NR metrics that

TABLE VIII
COMPARISON BETWEEN PEARSON CORRELATION COEFFICIENT OF $\widehat{\text{MOS}}$ VS MOS, FROM DEVELOPERS AND EVALUATORS

Developer			Our Evaluation			
NR Metric	**ω**	**ρ**	**IQA UGC**	**VQA UGC**	**VQA BC**	**Plot**
2stepQA-NR	M	—	0.55	0.49	0.51	// ε
ADMD	R	0.84..0.97	0.33	֍	֍	▼
AGWN	R	0.98	0.24	0.24	0.11	▼ ε
BRISQUE	M	0.94	0.17	0.29	0.41	▼ ε
CPBD	R	0.91	0.27	0.27	0.29	▼
CurveletQA	M	0.93	ε	—	—	// ε
dipIQ	K	0.89..0.96	0.33	0.42	0.63	// ▲
Entropy Noise	R	—	0.00	0.00	0.00	
HVS-MaxPol	R	0.67..0.93	0.47	0.54	0.25	▲
JNB	R	0.88..0.93	0.20	0.10	0.13	▼ ε
JP2KNR	M	0.92	0.20	0.09	0.22	▲▼
LBP	R	—	0.14	0.26	0.31	▲
Log-BIQA	M	0.92..0.95	0.40	0.31	0.10	▼ ε
MaxPol	R	0.94..0.97	0.35	0.43	0.12	▲ ε
Munsell Red	R	0.16, -0.07	0.06	0.14	0.09	▲▼
NIQE	M	0.91	0.33	0.28	0.38	// ε
NR-mean	M	—	0.23	0.22	0.09	
NR-std	M	—	0.21	0.23	0.06	
NR-entropy	M	—	0.29	0.29	0.05	▲
NR-skew	M	—	0.21	0.18	0.04	ε
NR-kurtosis	M	—	0.16	0.18	0.04	▲ ε
NR-PWN	R	0.80..0.98	0.27	0.16	0.30	▼// ε
NSS	M	0.66..0.91	0.15	0.24	0.06	ε
PIQE	M	0.86..0.90	0.16	0.32	0.56	//
SpEED-NR	M	0.63..0.91	0.16	0.17	0.12	▼ε
TDME	R	—	0.23	֍	֍	▼
TDMEC	R	0.99	0.36	֍	֍	▼

produce rank orders or predict CV success rates. The developers of the *iAITech-NJIT* NR metrics note differences between human perception and their CV use case [20]. The developers of *dipIQ* propose three statistics for evaluating the ability of **RANK** metrics [16].

For RCA metrics, an upper or lower triangle indicates that the NR metric could plausibly detect the intended impairment. However, the media must be visually examined to ensure that the correct impairment is detected with increasing sensitivity in response to changes in $\widehat{\text{MOS}}$. We did not perform this visual examination for the NR metrics in Table VIII.

Table IX reports the accuracy of NR metric *Sawatch Version 3* and its parameters for its training datasets. *S-BlackLevel* produces zero (0) for all media in the six VQA datasets. The GitHub repository provides scatter plots for each dataset (MOS vs $\widehat{\text{MOS}}$). Lab-to-lab differences have been retained (i.e., the MOSs are not mapped to a single scale).

C. Scatter Plots

A deeper understanding of NR metric performance requires visual examination of scatter plots of MOS vs $\widehat{\text{MOS}}$. Differences in the impairment's prevalence and severity can cause the scatter plots from different datasets to look very different. However, we expect the scatter plots for multiple datasets to cover a similar area and depict similar shapes.

Figure 1 plots the NR metrics from Table VIII for the CCRIQ dataset, within the context of the other five IQA UGC datasets. *CurveletQA* is omitted due to the aforementioned

TABLE IX
ACCURACY OF SAWATCH ON TRAINING DATA

NR Metric	Goal	IQA UGC	VQA UGC	VQA BC	Shape
Sawatch Version 3	MOS	0.62	0.62	0.57	//
S-BlackLevel	RCA	0.09	—	—	▲
S-Blockiness	RCA	0.13	0.22	0.36	▲
S-Blur	RCA	0.57	0.45	0.24	//
S-ColorNoise	RCA	0.28	0.03	0.15	▲
S-FineDetail	RCA	0.46	0.36	0.27	▼
S-Jiggle	RCA	—	0.38	0.07	▲
S-PanSpeed	RCA	—	0.42	0.63	▲ //
S-Pallid	RCA	0.12	0.17	0.10	▼
S-SuperSaturation	RCA	0.08	0.08	0.06	▲
S-WhiteLevel	RCA	0.28	0.22	0.03	▲
dipIQ	RCA	0.33	0.42	0.63	▲ //

problems. Figure 2 shows these same NR metrics for the ITS4S dataset within the context of the other *VQA UGC* and *VQA BC* datasets. Each metric's values span a different range and larger values could have either a positive or negative connotation.

Figure 3 and Figure 4 show the same plots for *Sawatch*. Each RCA metric spans a similar range (zero to one), where one indicates maximum impairment. We want RCA metrics to produce a vertical line at zero if the impairment is not present (e.g., the IQA UGC datasets lack motion impairments).

Each of these scatter plots shows the response of one NR metric on one dataset (blue dots) within the context of several other datasets (green dots). The x-axis is $\widehat{MOS}$ and the y-axis is MOS. The red line shows a linear fit for the current dataset (blue dots). To simplify comparisons between plots, the *LIVE-Wild* MOSs have been linearly mapped from its native [0,100] scale to [1,5]. This mapping does not fully account for differences between how subjects use the 5-level and 100-level ACR scales. More scatter plots are available at [4].

The NR metric scatter plots produce one of four shapes. From most desirable to least desirable, these are a scattering of data around a fit line, a lower triangle, an upper triangle, or no apparent pattern. An upper or lower triangle is undesirable if the NR metric predicts overall quality (MOS). A lower triangle is desirable for an RCA metric that detects a characteristic that appears infrequently in low quality media (e.g., noise or coding artifacts). An upper triangle is desirable for an RCA metric that detects a characteristic of some (but not all) high quality media (e.g., sharpness, colorfulness, or good composition). However, if the NR metric is supposed to detect an impairment associated with low quality, an upper triangle probably means the RCA metric detects something other than the intended impairment.

D. Analysis

Table VIII shows a significant drop in accuracy from the developer's ρ to our $\overline{\rho}$. Most of these NR metrics produce a scatter plot shape that is undesirable when estimating MOS (i.e., upper triangle, lower triangle, or no discernable pattern). Therefore, users may perceive a random relationship between $\widehat{\text{MOS}}$ and their ad-hoc assessments of MOS.

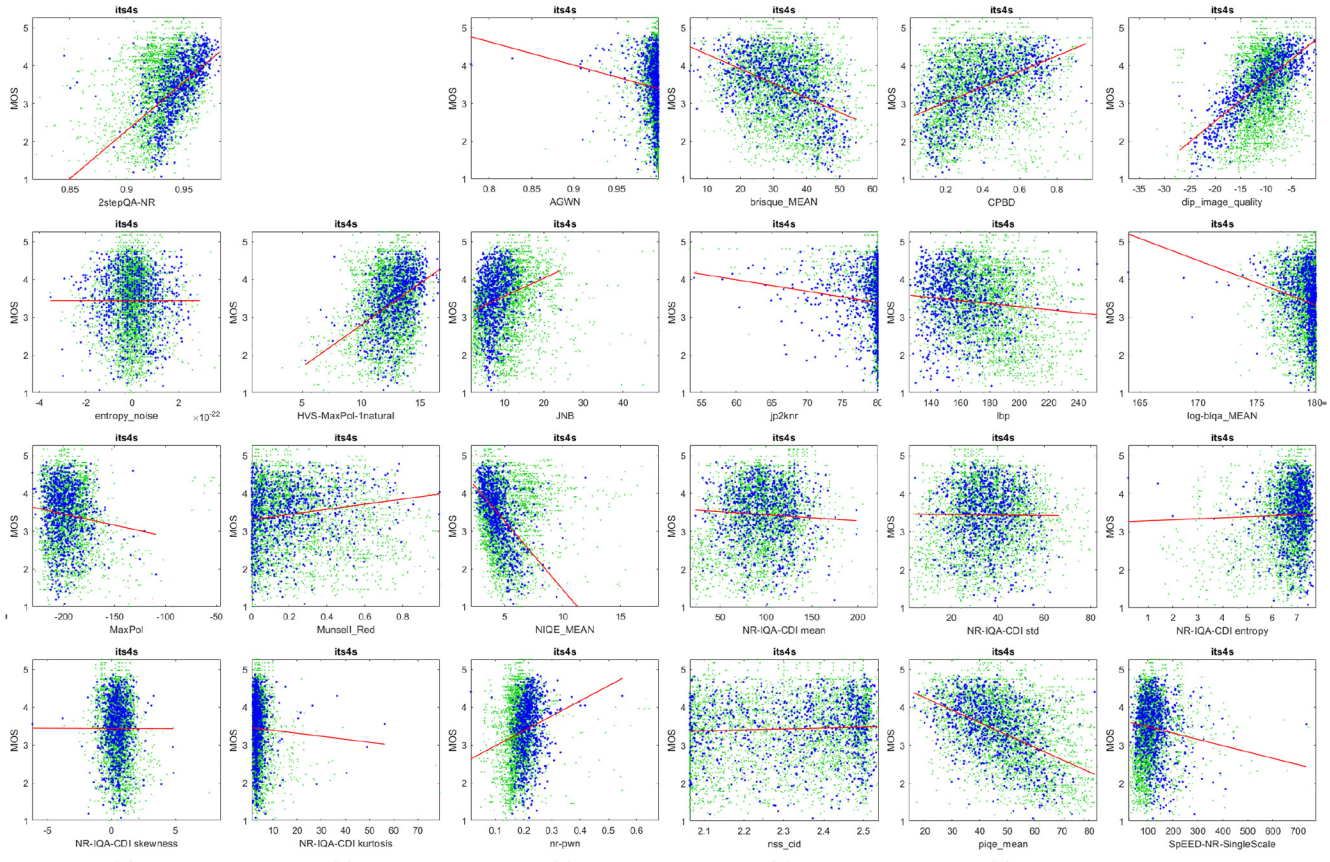

Fig. 2. Scatter plots depict the response of previously published NR metrics to the ITS4S dataset (blue dots) within the context of the six VQA datasets (green dots). The red line shows the linear fit for the ITS4S dataset.

Only *NIQE* and *2stepQA-NR* portray a spread of data around a line and both had severe outliers (recall that *2stepQA-NR* calls *NIQE*). *2stepQA-NR* and *HVS-MaxPol* have the best $\overline{\rho}$, but $\overline{\rho}$ was too low to support industry deployment. *dipIQ* portrays a spread of data around a line for ITS4S and *AGH/NTIA/Dolby* but not for the other ten datasets.

Sawatch Version 3 portrays a spread of data around a line for IQA UGC (see Figure 3) and *VQA UGC / VQA BC* (see Figure 4). The *Sawatch Version 3* RCA metric $\overline{\rho}$ in Table IX is often low. This does not necessarily indicate that the RCA metric is inaccurate. The ideal RCA metric should detect a single impairment and not respond to other impairments. We normalize RCA metric response to the [0,1] range (i.e., each x_p in (1)). We want $\widehat{MOS} > 0.8$ for the severe levels of the intended impairments. We want a vertical line at $\widehat{MOS} = 0.0$ if the impairment is not present. We observe this behavior in Figure 3 for *S-Jiggle* and *S-PanSpeed* and in Figure 4 for *S-Blockiness*. For RCA metrics, we want to see consistent response across 10+ datasets and scatter plot shapes that match the expected metric behavior. Media near $\widehat{MOS} \approx 1.0$ must be visually examined to confirm the presence of the impairment.

When the impairment is extremely rare, the fit line will be nearly random. The scatter plots can only be understood within the context of scatter plots from many other datasets, as per *S-Blockiness* in Figure 4. Only *KoNViD-1K* has enough super saturated media to properly analyze *S-SuperSaturated*. More datasets with super saturated colors and black balance problems are needed to further develop these NR metrics and ensure they provide proper RCA.

S-Black Level, *S-Blockiness*, and *S-White Level* detect infrequent impairments and so either portray a lower triangle or a vertical line around $\widehat{MOS} \approx 0.0$, depending on the dataset. *Blur* portrays a loose scattering of data around a line, indicating this is a dominant impairment for all 12 datasets. *S-PanSpeed* portrays a scattering of data around a line for *ITS4S4*, where pan speed is the dominant impairment, $\widehat{MOS} \approx 0.0$ for the motionless IQA datasets, and a lower triangle otherwise. *S-Color Noise*, *S-Pallid*, and *S-Super Saturation* have less well-defined plot shapes, indicating these impairments are infrequent or less influential.

Several of the NR metrics in Table VIII show potential for RCA. Some of these were intended for RCA: *CPBD*, *HVS-MaxPol*, *JNB*, and *MaxPol* for blur/sharpness; *TDME* and *TDMEC* for contrast enhancement; and *NR-PWN* for noisiness. Other metrics were intended for MOS estimation but show potential for RCA based on the scatter plot shapes: *NR-IQA Entropy*, *NR-IQA Kurtosis*, *OG-IQA,* and *SpEED-NR*. These algorithms would need to be trained on more data, to ensure resiliency, and visual inspection must be performed, to ensure these metrics detect a specific impairment.

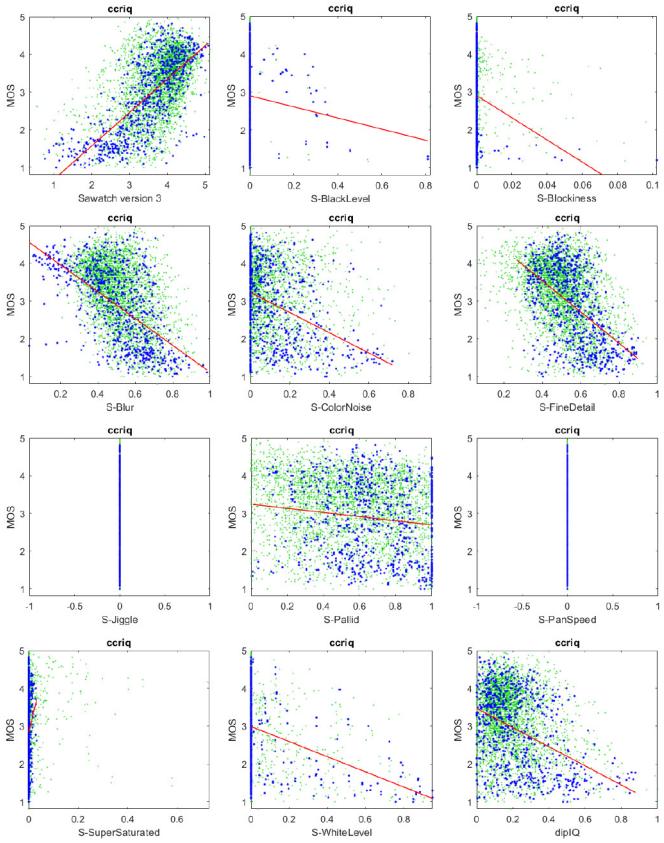

Fig. 3. Scatter plots depict the response of *Sawatch* and its constituent parameters to the CCRIQ dataset (blue dots) in the context of the six IQA UGC datasets (green dots). The red line shows the linear fit for the CCRIQ dataset.

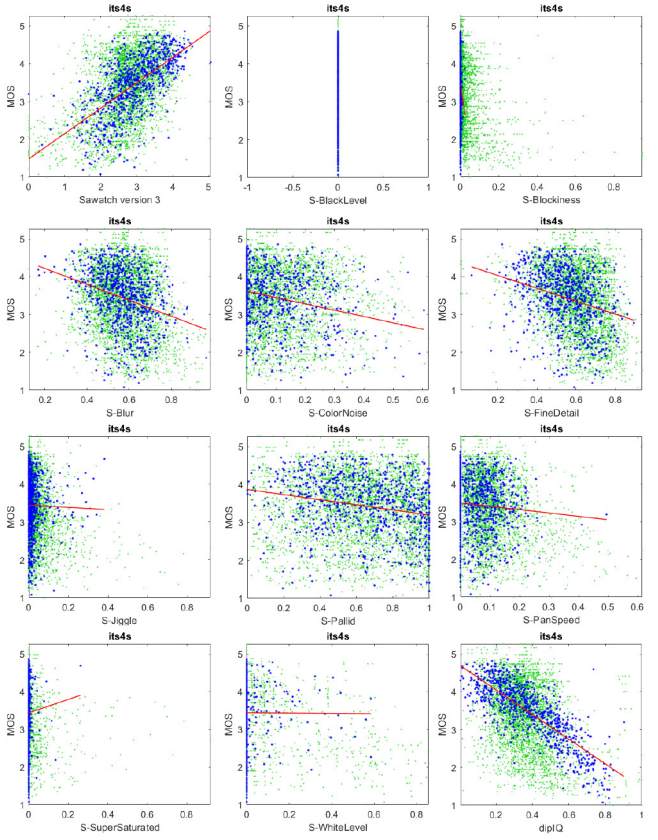

Fig. 4. Scatter plots depict the response of *Sawatch* and its constituent parameters to the ITS4S dataset (blue dots) within the context of the three *VQA BC* datasets (green dots). The red line shows the linear fit for the ITS4S dataset.

The two NR metrics in Table VIII that seem to have the best potential for RCA respond differently for the UGC and BC use cases. *HVS-MaxPol* has a lower triangle shape, $\overline{\rho} = 0.49$ for the nine UGC datasets, and $\overline{\rho} = 0.25$ for *VQA BC*. This is consistent with an impairment that is less prevalent for the broadcast use case. *dipIQ* has a lower triangle shape and $\overline{\rho} = 0.36$ for the nine UGC datasets, but a scatter around a fit-line and $\overline{\rho} = 0.63$ for the three BC datasets. This is consistent with an impairment that is less prevalent for the UGC use case.

The NR metrics in Table VIII exhibit problematic behaviors caused by insufficient training data. About half of them had problems (noted by ε) that must be addressed before the NR metric could be incorporated into an automated system. *NR-PWN* had particularly divergent responses to different datasets, with ρ ranging from 0.01 to 0.55. *JNB* responded poorly to the CCRIQ images displayed on a 4K monitor. *CPBD* had an undesirable scatter plot shape and fit but relatively high $\overline{\rho}$. Generally, we conclude that the metrics in Table VIII need to be trained on more datasets before they will mature into accurate, reliable, and deployable algorithms.

By contrast, *Sawatch* has demonstrated consistency across multiple datasets but needs to be supplemented with more RCA metrics (to assess missing impairments) and must be validated on unforeseen datasets. Examples of missing impairments include banding, mosquito noise, ringing, lens distortion, sun flare, ghosting, scaling errors, slicing, motion blur, flickering, jerky motion, de-interlacing artifacts, and panorama stitching artifacts.

VII. Caveats and Complications

Some datasets include similar impairments but at different levels of severity. Other datasets may omit an impairment entirely. *ITS4S3* emphasizes camera jiggle and lens flare because these are common problems for first responders. Camera jiggle and lens flare are missing from *VQA BC* because professional videographers avoid these impairments. White balance and black balance problems are common in UGC content that mostly comes from phones, tablets, and compact cameras. These problems do not appear in broadcast footage, where professional videographers manually set the camera's white balance and black balance.

Different use cases can change the relative impact of an impairment on MOS, or even invert the relationship between the impairment and MOS. Professional videographers slowly pan to create a pleasant visual appearance during the pan. Conversely, video surveillance users and drone operators pan and zoom very quickly to minimize travel time from one view another area. For this task, the pan quality may be irrelevant. When digital pathology (DP) slide imaging systems are not adjusted properly, the automatic focal system produces blurry DP images [19]. Conversely, professional videographers use blur to create pleasing aesthetics.

TABLE X
PEARSON CORRELATION COEFFICIENT RESPONSE TO BITRATE FOR ITS4S DATASET

Bitrate	***S-FineDetail***	***S-PanSpeed***	***S-ColorNoise***
Original	0.10	0.27	0.21
2.340 Mbps	0.21	0.41	0.27
1.732 Mbps	0.05	0.22	0.24
1.256 Mbps	0.21	0.01	0.26
0.915 Mbps	0.35	0.05	0.27
0.512 Mbps	0.30	0.06	0.35
All	0.30	0.10	0.25

We cannot predict or fully explain the relationship between the end user's use case and the perceptual impact of various impairments. Unexpected factors will make the NR metric appear to be more accurate or less accurate. This is a particular problem for proprietary NR metrics and **ML-NSS** metrics. The lay person has no way to understand or explain the NR metric's unexpected response to their use case.

We can infer a complex relationship between MOS, quality, and impairments by examining Table X, which shows the relationship between the ITS4S dataset [32] and RCA metrics. ITS4S emulates the bitrate ladder of broadcast video streaming service at 720p 24fps using an unrepeated scene experiment design, where each media contains similar content (e.g., different segments of a dance video). The Pearson correlation coefficient values in Table X are noisy and inexact, because a different the set of videos is used for each bitrate.

Table X shows three of the *Sawatch Version 3* parameters. *S-FineDetail* is more accurate for lower bitrates than higher bitrates. *S-PanSpeed* is more accurate for high bitrates than low bitrates. *S-ColorNoise* is minimally influenced by bitrate. Our point is that impairments may have greater or lesser impact on MOS in response to resolution, compression bitrate, or other unknown factors.

We know that market expectations and prior experience influence MOSs. Packet loss is commonplace for subjects with low Internet connectivity at home but may seem out of place for subjects with high-speed networks. Older datasets (like LIVE-2006) need to be deprecated or analyzed separately.

We suspect that gender, age, hobbies, and culture influence MOSs. However, [87] indicates that analyzing these factors would be prohibitively expensive (e.g., 200 subjects in a lab environment). Demographic differences may contribute to our difficulties when comparing datasets.

VIII. CONCLUSION

Based on this overview of prior research and our independent analysis of NR metric performance for modern camera systems, we conclude that none of the NR metrics we analyzed are accurate enough to be deployed by industry. Performance evaluations that indicate otherwise are based on insufficient data and are highly inaccurate.

All datasets have limitations that impact NR metric research. Datasets with MOSs are inherently size limited, due to constraints on how many media a subject can rate. The relationship between subject matter, impairment, and industry use-case is extremely complex. Analyses of a single dataset yield unstable performance statistics and lack external validity. Therefore, there is a high risk for any dataset that it does not meaningfully demonstrate the relationship between media, impairments, and MOS. This problem has three consequences.

First, NR metrics must be developed and evaluated with much more data. Based on our experience, we recommend a minimum of ten datasets with diverse characteristics. To have external validity, the dataset design must match an industry use case (e.g., variety of modern cameras, realistic impairment creation process). Datasets with unprecedented or unrealistic elements, like simulated impairments or limited subject matter, should be balanced by more realistic datasets.

Second, NR metrics should provide RCA. We began with an assertion from industry that impactful use cases require RCA, not MOS (see [3]). The complex relationship between industry use-case, impairments, and MOS means that NR metrics will rarely satisfy the industry end-user's exact requirements. The NR metric must justify $\widehat{\text{MOS}}$ by identifying specific impairments (i.e., explain why the quality is bad). This actionable information will allow industry users to bridge the gap between the NR metric design and their use case.

Third, NR metrics must be trained on a broad scope of all modern camera systems. Similar conclusions appear in [72] and [73]. Most NR metric research builds on the unstated hypothesis that media with limited impairments can be used to develop NR metrics that are accurate enough for industry. Our and other people's evaluations of NR metrics for modern camera systems reject this hypothesis. NR metric research based on limited impairments provides a rich and impactful foundation for future research—but has not by itself yielded viable solutions.

We believe the path to eventual maturity, standardization, and industry acceptance of NR metrics will require modular construction, collaboration, and devotion to incremental improvements. We propose a paradigm for collaboratively developing NR metrics that uses functional programming to split the research effort into independent algorithms, each providing RCA for a single impairment. These can be developed separately and replaced with improved algorithms. We provide a baseline NR metric, *Sawatch Version 3*, to kick start NR metric research that uses this paradigm. We encourage researchers to leverage the tools from the *NRMetricFramework* repository. An interactive demo [106] lets users to run *Sawatch* on their own images.

We propose an experiment design for datasets that will be used to develop and evaluate NR metrics. We organize datasets into subsets to understand the likely range of responses for common use cases (e.g., UGC videos). Our recommended initial scope includes camera capture impairments and compression but excludes temporal integration, transmission errors, and outdated impairments. Extending an NR metric's scope involves three steps: 1) gather datasets with the new impairments, 2) make sure the existing NR metric does **not** respond to the new impairments, and 3) develop new algorithms that **only** predict the quality impact of the new impairments (e.g., residuals $\widehat{\text{MOS}} - MOS$). When we split the research into independent algorithms, each providing RCA,

the NR metric should inherently not respond to other impairments. Thus, we expect most of the effort to fall within the first and third step.

The information and ideas in this report, while occasionally discouraging, are necessary to enable a future where industry deploys NR metrics as trusted components of innovative new media services.

REFERENCES

[1] M. A. Saad, P. Corriveau, and R. Jaladi, "Revealing the dark side of a subjective study: Learnings from noise and sharpness ratings," in *Proc. 7th Int. Workshop Qual. Multimed. Exp.*, 2015, pp. 1–5.

[2] W. Zhu *et al.*, "A multiple attributes image quality database for smartphone camera photo quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 2990–2994.

[3] M. H. Pinson, P. J. Corriveau, M. Leszczuk, and M. Colligan, "Open software framework for collaborative development of no reference image and video quality metrics," in *Proc. Int. Symp. Electron. Imag. Human Vis. Electron. Imag.*, Jan. 2020, p. 92-1.

[4] "National Telecommunications and Information Administration, Institute for Telecommunication Sciences, NR Metric Framework." [Online]. Available: https://github.com/NTIA/NRMetricFramework (Accessed: Nov. 24, 2020).

[5] I. Katsavounidis. "Digital Optimizer—A Perceptual Video Encoding Optimization Framework." Netflix Tech Blog. Mar. 2018. [Online]. Available: https://netflixtechblog.com/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f (Accessed: Jul. 19, 2022).

[6] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, "Predicting the quality of images compressed after distortion in two steps," *IEEE Trans. Image Process.*, vol. 28, pp. 5757–5770, 2019.

[7] F. Xie, Y. Lu, A. C. Bovik, Z. Jiang, and R. Meng, "Application-driven no-reference quality assessment for dermoscopy images with multiple distortions," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 6, pp. 1248–1256, Jun. 2016.

[8] C. Lim and R. Paramesran, "Blind image quality assessment for color images with additive Gaussian white noise using standard deviation," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, 2014, pp. 39–41.

[9] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, pp. 3339–3352, 2012.

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

[11] K. Gu *et al.*, "Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 432–443, Mar. 2016.

[12] Y. Zhang, A. K. Moorthy, D. Chandler, and A. Bovik, "C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes," *Signal Process. Image Commun.*, vol. 29, no. 7, pp. 725–747, May 2014.

[13] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *Proc. Int. Workshop Qual. Multimed. Exp.*, 2009, pp. 87–91.

[14] L. Liu, H. Dong, H. Huang, and A. C. Bovik, "No-reference image quality assessment in curvelet domain," *Signal Process. Image Commun.*, vol. 29, no. 4, pp. 494–505, 2014.

[15] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, pp. 3350–3364, 2011.

[16] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "DipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, pp. 3951–3964, 2017.

[17] M. Rakhshanfar and M. A. Amer, "No-reference image quality assessment for removal of processed and unprocessed noise," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 2179–2183.

[18] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, 2017.

[19] M. S. Hosseini, Y. Zhang, and K. N. Plataniotis, "Encoding visual sensitivity by MaxPol convolution filters for image sharpness assessment," *IEEE Trans. Image Process.*, vol. 28, pp. 4510–4525, 2019, doi: 10.1109/TIP.2019.2906582.

[20] H. Shi and C. Liu, "An innovative video quality assessment method and an impairment video dataset," in *Proc. IEEE Int. Conf. Imag. Syst. Techn.*, Aug. 2021, pp. 24–26.

[21] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Trans. Image Process.*, vol. 18, pp. 717–728, 2009.

[22] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.*, vol. 14, pp. 1918–1927, 2005.

[23] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir, "Perceptual quality assessment for H.264/AVC compression," in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, 2012, pp. 597–602, doi: 10.1109/CCNC.2012.6181021.

[24] M. Leszczuk, "Assessing task-based video quality—A journey from subjective psycho-physical experiments to objective quality models," in *Proc. Int. Conf. Multimedia Commun. Services Security*, 2011, pp. 91–99.

[25] L. Janowski and Z. Papir, "Modeling subjective tests of quality of experience with a generalized linear model," in *Proc. Int. Workshop Qual. Multimedia Exp.*, 2009, pp. 35–40.

[26] E. Cerqueira, S. Zeadally, M. Leszczuk, M. Curado, and A. Mauthe, "Recent advances in multimedia networking," *Multimedia Tools Appl.*, vol. 54, no. 3, pp. 635–647, 2011.

[27] M. Leszczuk, M. Hanusiak, M. C. Q. Farias, E. Wyckens, and G. Heston, "Recent developments in visual quality monitoring by key performance indicators," *Multimedia Tools Appl.*, vol. 75, no. 17, pp. 10745–10767, 2016.

[28] "AGH Video Qualtiy of Experience (QoE Team) Indicators." [Online]. Available: https://qoe.agh.edu.pl/indicators/ (Accessed: Aug. 20, 2021).

[29] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[30] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Trans. Image Process.*, vol. 23, pp. 4850–4862, 2014.

[31] M. S. Hosseini and K. N. Plataniotis, "Image sharpness metric based on maxpol convolution kernels," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 296–300, doi: 10.1109/ICIP.2018.8451488.

[32] M. H. Pinson, "ITS4S: A video quality dataset with four-second unrepeated scenes," NTIA, Washington, DC, USA, document TM-18-532, Feb. 2018.

[33] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, pp. 3998–4011, 2018.

[34] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 209–212, Mar. 2013.

[35] M. Outtas, L. Zhang, O. Deforges, W. Hammidouche, A. Serir, and C. Cavaro-Menard, "A study on the usability of opinion-unaware no-reference natural image quality metrics in the context of medical images," in *Proc. Int. Symp. Signal Image Video Commun.*, 2016, pp. 308–313.

[36] R. Herzog *et al.*, "NoRM: No-reference image quality metric for realistic image synthesis," *Comput. Graph. Forum*, vol. 31, no. 2, pp. 545–554, 2012.

[37] I. T. Ahmed, C. S. Der, N. Jamil, and B. T. Hammad, "Analysis of probability density functions in existing no-reference image quality assessment algorithm for contrast-distorted images," in *Proc. IEEE 10th Control Syst. Graduate Res. Colloquium (ICSGRC)*, Shah Alam, Malaysia, 2019, pp. 133–137.

[38] T. Zhu and L. Karam, "A no-reference objective image quality metric based on perceptually weighted local noise," *EURASIP J. Image Video Process.*, vol. 2014, p. 5, Jan. 2014.

[39] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 838–842, Jul. 2015.

[40] L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, "Blind image quality assessment by relative gradient statistics and adaboosting neural network," *Signal Process. Image Commun.*, vol. 40, no. 1, pp. 1–15, Jan. 2016.

[41] N. Venkatanath, D. Praneeth, B. M. Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. 21st Nat. Conf. Commun. (NCC)*, Piscataway, NJ, USA, 2015, pp. 1–6.

[42] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 995–1002.

[43] *Subjective Video Quality Assessment Methods for Multimedia Applications*, Rec. P.910, Int. Telecommun. Union, Geneva, Switzerland, Apr. 2008.

[44] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

[45] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process. Image Commun.*, vol. 29, no. 8, pp. 856–863, Sep. 2014.

[46] A. Samani, K. Panetta, and S. Agaian, "Transform domain measure of enhancement—-TDME—For security imaging applications," in *Proc. IEEE Int. Conf. Technol. Homeland Security (HST)*, Nov. 2013, pp. 265–270.

[47] K. Panetta, A. Samani, and S. Agaian, "A robust no-reference, no-parameter, transform domain image quality metric for evaluating the quality of color images," *IEEE Access*, vol. 6, pp. 10979–10985, 2018.

[48] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, pp. 5923–5938, 2019, doi: 10.1109/TIP.2019.2923051.

[49] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, pp. 1352–1365, 2014.

[50] "VQET Image Quality Evaluation Tool (VIQET)." [Online]. Available: https://github.com/VIQET (Accessed: Aug. 20, 2021).

[51] S. Katsigiannis *et al.*, "Interpreting MOS scores, when can users see a difference? Understanding user experience differences for photo quality," *Qual. User Exp.*, vol. 3, no. 1, p. 6, May 2018.

[52] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, pp. 289–300, 2016.

[53] L. Janowski, L. Malfait, and M. Pinson, "Evaluating experiment design with unrepeated scenes for video quality subjective assessment," *Qual. User Exp.*, vol. 4, p. 2, Jun. 2019.

[54] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2408–2415.

[55] A. Ciancio, A. L. N. T. T. da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Process.*, vol. 20, pp. 64–75, 2011.

[56] M. A. Saad *et al.*, "Impact of camera pixel count and monitor resolution perceptual image quality," in *Proc. Colour Visual Comput. Symp. (CVCS)*, Aug. 2015, pp. 1–6.

[57] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "CID2013: A database for evaluating no-reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 24, pp. 390–402, 2015.

[58] J. Nawala, M. H. Pinson, M. Leszczuk, and L. Janowski, "Study of subjective data integrity for image quality data sets with consumer camera content," *J. Imag.*, vol. 6, no. 3, p. 7, 2020.

[59] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014—A database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process*, vol. 25, pp. 3073–3086, 2016, doi: 10.1109/TIP.2016.2562513.

[60] W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu, and G. Zhai, "A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process.*, 2018, pp. 1–6.

[61] X. Peng, H. Cao, and P. Natarajan, "Document image quality assessment using discriminative sparse representation," in *Proc. 12th IAPR Workshop Document Anal. Syst. (DAS)*, 2016, pp. 227–232.

[62] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-scale crowdsourced study for tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, pp. 4725–4740, 2017.

[63] D. Y. Lee *et al.*, "A subjective and objective study of space-time subsampled video quality," *IEEE Trans. Image Process.*, vol. 31, pp. 934–948, 2022.

[64] M. H. Pinson, "ITS4S2: An image quality dataset with unrepeated images from consumer cameras," NTIA, Washington, DC, USA, document TM-19-537, Apr. 2019.

[65] M. H. Pinson, "ITS4S3: A video quality dataset with unrepeated videos, camera impairments, and public safety scenarios," NTIA, Washington, DC, USA, document TM-19-538, Apr. 2019.

[66] M. H. Pinson and S. Elting, "ITS4S4: A video quality study of camera pans," NTIA, Washington, DC, USA, document TM-20-545, Dec. 2019.

[67] V. Hosu *et al.*, "The Konstanz natural video database (KoNViD-1k)," in *Proc. 9th Int. Conf. Qual. Multimed. Exp.*, 2017, pp. 1–6.

[68] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.

[69] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild," *IEEE Access*, vol. 9, pp. 72139–72160, 2021.

[70] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, pp. 3440–3451, 2006.

[71] M. Chen, Y. Jin, T. Goodall, X. Yu, and A. C. Bovik, "Study of 3D virtual reality picture quality," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 89–102, Jan. 2020.

[72] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2061–2077, Sep. 2018.

[73] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, pp. 612–627, 2019, doi: 10.1109/TIP.2018.2869673.

[74] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, pp. 372–387, 2016.

[75] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108069–108082, 2021.

[76] Y. Hu, Y. Zhang, Z. Liu, Z. Chen, and S. Liu, "Subjective study of perceptual quality for micro-video applications," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2020, pp. 229–232.

[77] M. H. Pinson *et al.*, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE J. Sel. Top. Signal Process.*, Vol. 6, no. 6, pp. 640–651, Oct. 2012.

[78] Z. Zhu *et al.*, "A comparative study of algorithms for realtime panoramic video blending," *IEEE Trans. Image Process.*, vol. 27, pp. 2952–2965, 2018, doi: 10.1109/TIP.2018.2808766.

[79] W. Zhu, G. Zhai, C. Yao, and X. Yang, "SIQD: Surveillance image quality database and performance evaluation for objective algorithms," in *Proc. IEEE Visual Commun. Image Process. (VCIP)*, 2018, pp. 1–4, doi: 10.1109/VCIP.2018.8698737.

[80] G. Wang, L. Li, Q. Li, K. Gu, Z. Lu, and J. Qian, "Perceptual evaluation of single-image super-resolution reconstruction," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3145–3149, doi: 10.1109/ICIP.2017.8296862.

[81] S. Athar, T. Costa, K. Zeng, and Z. Wang, "Perceptual quality assessment of UHD-HDR-WCG videos," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1740–1744, doi: 10.1109/ICIP.2019.8803179.

[82] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC dataset for video compression research," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, Aug. 2019, pp. 1–5.

[83] J. G. Yim, Y. Wang, N. Birkbeck, and B. Adsumilli, "Subjective quality assessment for YouTube UGC dataset," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 131–135, doi: 10.1109/ICIP40778.2020.9191194.

[84] M. H. Pinson, "Analysis of no-reference metrics for image and video quality of consumer applications," NTIA, Washington, DC, USA, document TM-20-547, Jan. 2020.

[85] J. Xu, J. Li, X. Zhou, W. Zhou, B. Wang, and Z. Chen, "Perceptual quality assessment of Internet videos," in *Proc. 29th ACM Int. Conf. Multimedia (MM)*, Oct. 2021, pp. 1248–1257.

[86] M. H. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2D and 3D subjective video quality tests," *EURASIP J. Image Video Process.*, vol. 2013, p. 50, Aug. 2013, doi: 10.1186/1687-5281-2013-50.

[87] L. Janowski and M. H. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2210–2224, Dec. 2015.

[88] M. H. Pinson, "Confidence intervals for subjective tests and objective metrics that assess image, video, speech, or audiovisual quality," NTIA, Washington, DC, USA, Rep. TR-21-550, Oct. 2020,.

[89] K. Brunnström, S. Tavakoli, and J. Søgaard, "Compensating for type-I errors in video quality assessment," in *Proc. 7th Int. Conf. Qual. Multimed. Exp*, 2015, pp. 1–2.
[90] M. H. Pinson, "Technology gaps in first responder cameras," NTIA, Washington, DC, USA, document TM-17-524, May 2017.
[91] D. Kahneman, O. Sibony, and C. R. Sunstein, *Noise: A Flaw in Human Judgement*, Hachette Book Group, New York, NY, USA, 2021.
[92] M. N. Garcia and A. Raake, "Normalization of subjective video test results using a reference test and anchor conditions for efficient model development," in *Proc. 2nd Int. Workshop Qual. Multimed. Exp.*, 2010, pp. 88–93, doi: 10.1109/QOMEX.2010.5517960.
[93] P. Pérez, L. Janowski, N. García, and M. H. Pinson, "Subjective assessment experiments that recruit few observers with repetitions (FOWR)," *IEEE Trans. Multimedia*, vol. 24, pp. 3442–3454, 2022, doi: 10.1109/TMM.2021.3098450.
[94] L. Janowski, L. Malfait, and M. H. Pinson, "Evaluating experiment design with unrepeated scenes for video quality subjective assessment," *Qual. User Exp.*, vol. 4, p. 2, Jun. 2019.
[95] S. Sebastian, J. Abrams, and W. S. Geisler, "Constrained sampling experiments reveal principles of detection in natural scenes," *Proc. Nat. Acad. Sci.*, vol. 114, no. 28, pp. E5731–E5740, 2017.
[96] (Univ. Texas, Austin, TX, USA). *Laboratory for Image & Video Engineering*. Accessed: Sep. 3, 2021. [Online]. Available: https://live.ece.utexas.edu/research/Quality/index.htm
[97] (Universität Konstanz, Konstanz, Germany). *The Konstanz Visual Quality Databases*. Accessed: Sep. 20, 2021. [Online]. Available: http://database.mmsp-kn.de/
[98] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.
[99] Y. Liu, K. Gu, S. Wang, D. Zhao, and W. Gao, "Blind quality assessment of camera images based on low-level and high-level statistical features," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 135–146, Jan. 2019.
[100] Z. Zhu, H. Liu, J. Lu, and S. -M. Hu, "A metric for video blending quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 3014–3022, 2020.
[101] "VQEG HDTV Test. Vqeghd1." [Online]. Available: https://www.cdvl.org/members-section/view-file/?id=3019 (Accessed: Jun. 21, 2022).
[102] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004, doi: 10.1109/TBC.2004.834028.
[103] S. Wolf and M. H. Pinson, "Video quality model for variable frame delay (VQM_VFD)," NTIA, Washington, DC, USA, document TM-11-482, Sep. 2011.
[104] C. G. Bampis, Z. Li, and A. C. Bovik, "Enhancing temporal quality measurements in a globally deployed streaming video quality predictor," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 614–618.
[105] A. Lee. "This Ugly T-Shirt Makes You Invisible to Facial Recognition Tech." Wired. Nov. 2020. [Online]. Available: https://www.wired.co.uk/article/facial-recognition-t-shirt-block (Accessed: Jul. 19, 2022).
[106] "Sawatch Demo." [Online]. Available: https://vqwt.its.bldrdoc.gov/login.php (Accessed: Jul. 19, 2022).

Margaret H. Pinson received the B.S. and M.S. degrees in computer science from the University of Colorado at Boulder, Boulder, CO, USA, in 1988 and 1990, respectively.

Since 1988, she has been with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, Boulder. She is an internationally recognized expert with 30 years of experience developing improved methods for assessing video quality. Her research includes algorithm development, human testing, and international standards. She contributed to eight national and international efforts of ATIS and the Video Quality Experts Group to independently validate video quality metrics. She led the effort to create ITU-T Rec. P.913, which describes improved subjective test methods for modern video systems. She has written 79 publications. She contributes to ITU Recommendations and has led several efforts to independently validate video quality metrics, which is a necessary step of the standards development process. Her current research focuses on NR metrics that predict what people would say is the quality of an image or video.

Mrs. Pinson is a VQEG Co-Chair, administers the Consumer Digital Video Library, and makes all of her algorithms openly available. She helped design and conduct three prize challenges.