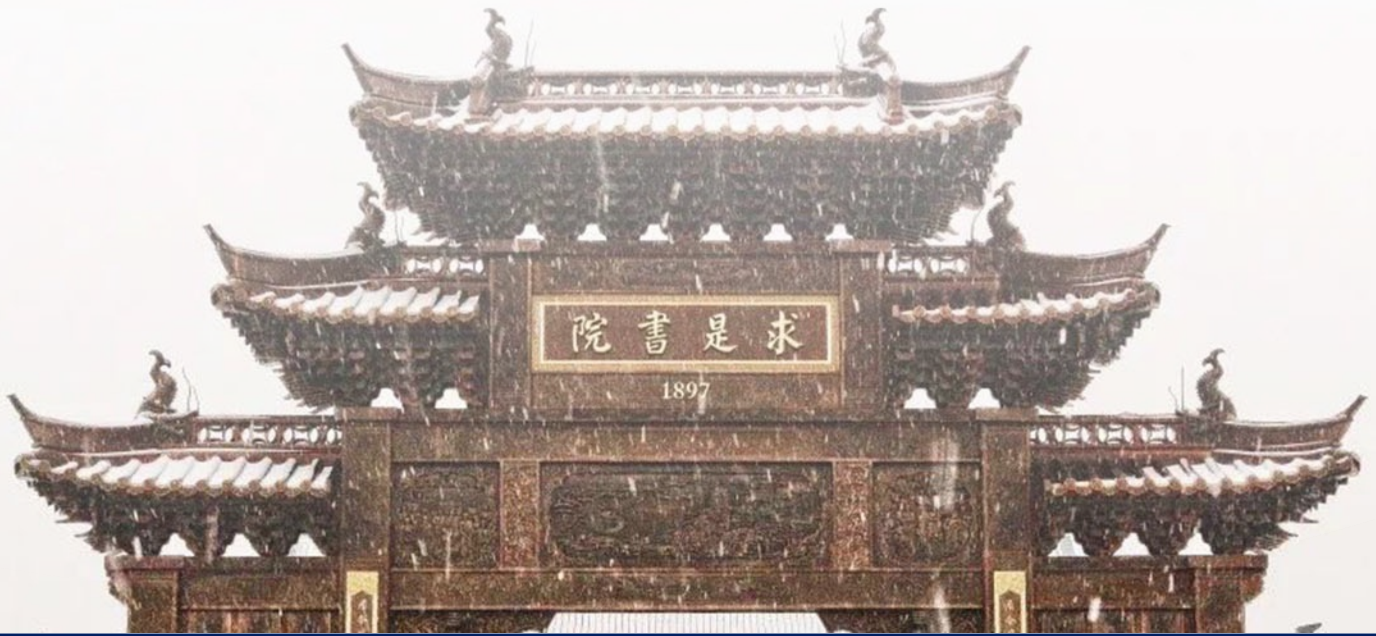


AHG17/AHG4: A Generic Full-Reference Objective Quality Assessment Method for Compressed Video

JVET-AP0186

Jiayi Wang, Xiaoqi Zhuang, Jiaqi Zhang, Lu Yu (ZJU)

Presenter: Lu Yu
Organization: Zhejiang University

Related Work & Resources

Related Proposals

This work follows and extends a series of proposals submitted to MPEG AG 5. For a comprehensive understanding, please refer to the following supplementary documents:

MPEG-m73451 (June 2025): Technical details for FDIM

MPEG-m76136 (April 2026) & m76136 (October 2025):
Evaluation Results for FDIM
m76934 (April 2026) provides cross check

Open Source Code

The full implementation of the FDIM metric is publicly available as open-source software on GitHub. We welcome the codec community to use, test, modify, and contribute to the project.

GitHub Repository
github.com/MCL-ZJU/FDIM

Agenda

01

Motivation & Problem Statement

The challenges in video quality assessment.

02

Proposed Method: FDIM

The hybrid architecture and training strategies.

03

Experimental Validation

Comprehensive results across multiple datasets.

04

Conclusion & Proposal

Summary of contributions and future recommendations.

01 Motivation & Problem Statement

Emerging Challenges

Shift from Block-based Artifacts

Advances in encoding technologies introduce content-varying distortions, differing fundamentally from traditional blocking/blurring.

HDR/WCG Perceptual Complexity

distinct perceptual characteristics that challenge existing models.

Limitations

Poor Generalization

Current VQA metrics often lack robustness and fail to maintain accuracy across diverse codecs, dynamic ranges and content types.

Subjective Testing Bottlenecks

While accurate, human subjective testing is resource-intensive and slow.

The Ideal Metric

- **Generic:**

  Works for both traditional & learning-based codecs

- **Robust:**

  Models SDR & HDR/WCG distortions accurately

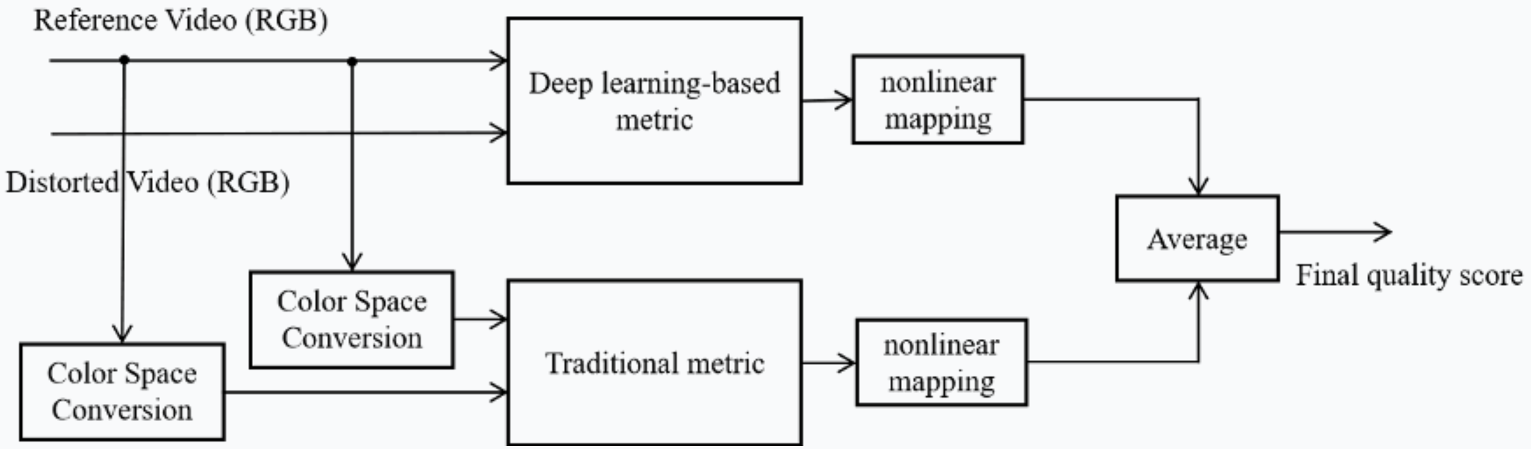
- **Perceptual-Aligned:**

  High correlation with human judgment

FDIM is designed to meet these goals.

02 Proposed Method: FDIM

FDIM integrates the strength of traditional handcrafted features and deep learning-based features.
Nonlinear mappings are calibrated on training data with mean opinion score ratings (1-5 scale).



Hybrid Architecture

Deep Learning component: Performs content-adaptive feature-distance modeling with multi-scale feature fusion to capture content-varying distortions (CNN-based).

Hand-crafted component: Provides stable, complementary cues for conventional distortions (VMAF).

Training Data

- Containing 16k+ compressed sequences
- Encoded by both traditional block-based codecs and end-to-end perceptually optimized neural video codecs
- DCR method with a 7-point scale & crowdsourcing and controlled laboratory subjective tests

Training Strategy

Ranking-inspired learning over both identical and distinct source contents

03 Experimental Validation

A Comprehensive Evaluation

We validate FDIM's performance against representative metrics across diverse scenarios.

Datasets

• **7 Datasets:** Covering diverse scenarios.
-**MPEG CVQM HD/UHD:** Key standard datasets for codec evaluation.
-**Four SDR Datasets:** MCML 4K, AVT-VQDB-UHD-1, WaterlooIVC 4K, BVI HD.
-**LIVE-HDR:** For HDR performance assessment.

Evaluation Metrics & protocals

• PLCC (Pearson Linear Correlation) & SROCC (Spearman Rank Order Correlation)

• Per-sequence evaluation & All-sequence evaluation

Competitors

PSNR • MS-SSIM • XPSNR • VMAF

wPSNR • DeltaE100 • PSNRL100

Key Results: MPEG CVQM Datasets

The MPEG CVQM datasets are the most relevant for standardization, containing both traditional and neural codec samples.

Metric	HD (PLCC / SROCC) All-seq. Correlation	UHD (PLCC / SROCC) All-seq. Correlation	HD (PLCC / SROCC) Per-seq. Correlation	UHD (PLCC / SROCC) Per-seq. Correlation
FDIM	**0.955 / 0.956**	**0.881 / 0.904**	0.971 / 0.961	**0.926** / **0.948**
VMAF	0.943 / 0.941	0.828 / 0.843	**0.973** / 0.964	0.920 / 0.938
PSNR(AVG)	0.735 / 0.756	0.499 / 0.536	0.948 / 0.964	0.888 / 0.929
XPSNR(AVG)	0.895 / 0.918	0.756 / 0.808	0.948 / 0.964	0.887 / 0.922
MS-SSIM(AVG)	0.720 / 0.733	0.629 / 0.647	0.959 / **0.966**	0.892 / 0.924

FDIM exhibits strong generalization capability for both conventional compression artifacts from traditional hybrid codecs and generative distortions introduced by NN-based codecs.

Key Results: Four SDR Public Datasets

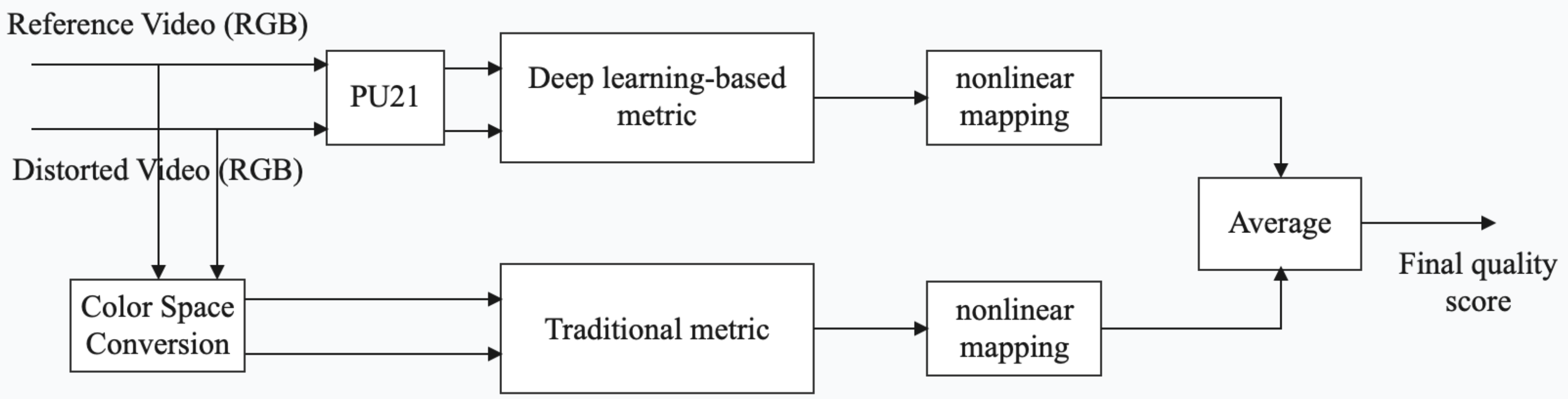
We have further evaluated FDIM on four well-established SDR datasets with a wide range of content and encoders, ensuring the results are robust and cross-dataset consistent. We **average** the correlation results the across four datasets.

Metric	PLCC All-seq. Correlation	SROCC All-seq. Correlation	PLCC Per-seq. Correlation	SROCC Per-seq. Correlation
FDIM	**0.862**	**0.870**	**0.909**	**0.915**
VMAF	<u>0.795</u>	<u>0.799</u>	0.891	0.891
PSNR(AVG)	0.612	0.640	0.880	0.876
XPSNR(AVG)	0.769	0.802	0.897	<u>0.896</u>
MS-SSIM(AVG)	0.555	0.674	<u>0.895</u>	<u>0.896</u>

FDIM yields competitive and stable all/per-sequence quality prediction across all tested datasets, demonstrating reliable generalization to diverse content characteristics and compression distortion patterns.

Key Results: LIVE-HDR Dataset

We compared the FDIM metric against three HDR objective quality metrics used in JVET CTC.
PU21 transform is applied to achieve a perceptually uniform mapping as a preprocessing module for HDR content.
LIVE-HDR dataset provides subjective scores collected under **two ambient illumination conditions (dark/bright)**.



Metric	bright (PLCC / SROCC) All-seq. Correlation	dark (PLCC / SROCC) All-seq. Correlation	bright (PLCC / SROCC) Per-seq. Correlation	dark (PLCC / SROCC) Per-seq. Correlation
FDIM	<u>0.874 / 0.922</u>	<u>0.845 / 0.882</u>	**0.977 / 0.977**	**0.973 / 0.977**
FDIM+PU21	**0.885 / 0.933**	**0.865 / 0.900**	<u>0.975 / 0.975</u>	<u>0.973 / 0.973</u>
wPSNR	0.636 / 0.689	0.562 / 0.621	0.946 / 0.960	0.941 / 0.960
DeltaE100	0.650 / 0.698	0.600 / 0.648	0.943 / 0.949	0.920 / 0.937
PSNRL100	0.724 / 0.729	0.681 / 0.678	0.962 / 0.965	0.948 / 0.959

Performance of 4K Downsampling

Downsampling for Efficiency

- **Objective:** To reduce computational complexity while maintaining high-quality prediction accuracy.
- **Process:** Input 4K (2160p) video sequences are spatially downsampled to 1080p before being fed into the FDIM model.

Performance on CVQM-UHD

Dataset	All-Sequence PLCC	All-Sequence SROCC	Per-Sequence PLCC	Per-Sequence SROCC
FDIM(4k downsampling)	**0.896**	**0.907**	**0.938**	**0.949**
FDIM	0.881	0.904	0.926	0.948

Key Results: LIVE-HDR Dataset

We evaluate the total runtime of FDIM and the computational complexity of the deep component of FDIM in terms of GFLOPs. The deep component of FDIM has **12M parameters**.

	GFLOP (deep component)	GPU		CPU	
		Runtime (s)	FPS	Runtime (s)	FPS
1280x720	94.2	0.068	14.779	5.047	0.198
1920x1080	211.6	0.107	9.330	11.064	0.090
3840x2160	844.9	0.383	2.613	42.807	0.023

FDIM is computationally feasible for real-world deployment with GPU acceleration, and the low-complexity implementation offers a great balance between performance and efficiency.

Key Factors for Superior Performance

Synergy of Hand-crafted & Deep Features

Deep Feature:
Extracts multi-scale representations, capturing structural, textural, and semantic distortions, with particular sensitivity to NVCs' generative artifacts.

Hand-crafted Feature:

Provides stable and interpretable assessments based on established perceptual models.

Effective Training Strategy

Large-scale & Diversity:
trained on a large-scale dataset of 16k sequences encoded by both traditional block-based codecs and end-to-end perceptually optimized neural video codecs

Ranking-inspired Learning:
adopted over both identical and distinct source

Moderate Model Capacity

Capacity Design:
designed with a moderate capacity to match the available training data

Efficiency Trade-off: avoid unnecessary complexity while preserving sufficient representation capability

04 Conclusion & Proposal

Summary of Contributions

FDIM: A generic Metric

A generic full-reference quality metric specifically designed for evaluating compressed video streams.

Broad Applicability

Supports a wide range of scenarios, including traditional & neural codecs, SDR & HDR/WCG content, and multiple video resolutions.

Superior Performance

Achieves higher correlation with human subjective scores compared to state-of-the-art metrics like PSNR, VMAF, XPSNR, and MS-SSIM across most of tested scenarios.

Practical Deployment

To facilitate real-world usage, a low-complexity implementation of FDIM is provided, ensuring it can be efficiently integrated into existing workflows.

Proposal

We propose that JVET consider **FDIM as a candidate quality metric beyond PSNR**.

Perceptual Coding Research

As a primary evaluation tool for new coding algorithms to better align with human visual perception.

CfP Proposal Evaluation

To assess the performance of proposals submitted to JVET and ensure high-quality contributions.

Standardized Performance Comparison

Providing a more reliable basis for codec comparisons.

We believe FDIM represents a significant step forward in video quality assessment and can greatly benefit the standardization process.

Thanks for your attention!
JVET-AP0186

Presenter: Lu Yu
Organization: Zhejiang University