Full-Reference Image Quality Expression via Genetic Programming

Illya Bakurov, Marco Buzzelli, Raimondo Schettini, Mauro Castelli, and Leonardo Vanneschi

***Abstract*— Full-reference image quality measures are a fundamental tool to approximate the human visual system in various applications for digital data management: from retrieval to compression to detection of unauthorized uses. Inspired by both the effectiveness and the simplicity of hand-crafted Structural Similarity Index Measure (SSIM), in this work, we present a framework for the formulation of SSIM-like image quality measures through genetic programming. We explore different terminal sets, defined from the building blocks of structural similarity at different levels of abstraction, and we propose a two-stage genetic optimization that exploits hoist mutation to constrain the complexity of the solutions. Our optimized measures are selected through a cross-dataset validation procedure, which results in superior performance against different versions of structural similarity, measured as correlation with human mean opinion scores. We also demonstrate how, by tuning on specific datasets, it is possible to obtain solutions that are competitive with (or even outperform) more complex image quality measures.**

***Index Terms*— Image quality, full-reference image quality assessment, image similarity, SSIM, genetic programming.**

I. Introduction

NOWADAYS, digital imaging constitutes a whole new way of communication and, due to technological advancements, creating any sort of digital content and sharing it to a virtual community of millions of people is just a matter of a few clicks and seconds. Such ease of digital content creation also encourages the demand for its efficient management. In this sense, one of the key concepts underpinning numerous digital imaging management techniques is the one of image similarity. Notwithstanding, the concept of *similarity* is vague and multifaceted, therefore open to different interpretations and definitions: it can be analytically defined with objective criteria, but it can also refer to subjective interpretations. Image similarity can be used to quantify the perceived visual quality of digital images, typically employed in the field of signal compression and telecommunication, serve as the basis for image-based search and image clustering, both useful in the management of personal photo collections and in detecting cases of copyright infringement and use of proprietary content without proper credit. In this study, we focus on the objective means to quantify images' visual quality.

Many aspects influence the perceived visual quality of a digital image or video. For instance, an accidental shake of the capturing device during the acquisition process can blur the whole image. Likewise, poor lighting conditions can result in images with a low dynamic range and high noise levels. Also, raindrops adhering to a window or camera lens can significantly decrease the performance of scene recognition systems for automotive applications, etc. Moreover, several image processing steps usually occur before the final users can employ the digital imagery. Frequently, these are organized into a sequential set of steps, like digitization, compression, storage, transmission, and reproduction, and may result in a noticeable visual degradation of the output digital imagery. From the perspective of media quality perception, this might result in a sub-optimal viewing experience. In this sense, to better manage the ever-growing digital content (images in particular) and improve users' experience, it becomes necessary to efficiently quantify the perceived visual quality of images. For those types of applications in which humans ultimately view the images, the most appropriate method for images assessment is through the human visual system itself [1], i.e., by involving people to assess images' quality. However, in the context of a digital society where hundreds of millions of photos are being generated and uploaded to social media every day, subjective evaluation becomes impractical as it is time-consuming, expensive, and highly sensible to the experimental design [2]. Considering the aforementioned limitations, several researchers have proposed objective measures that can automatically (i.e., without human involvement) estimate the perceived visual quality as humans would do. These are called image quality assessment measures (IQAMs) and are frequently classified based on the availability of

Manuscript received 22 March 2022; revised 6 October 2022 and 26 December 2022; accepted 4 February 2023. Date of publication 17 February 2023; date of current version 28 February 2023. This work was supported by national funds through the FCT (Fundação para a Ciência e a Tecnologia) under the projects Algoritmos de Inteligência artificial no Consumo de crédito e conciliação de Endividamento (AICE) (DSAIPA/DS/0113/2019) and UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS. Mauro Castelli acknowledges the financial support from the Slovenian Research Agency (research core funding no. P5-0410). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rafal K. Mantiuk. *(Corresponding author: Marco Buzzelli.)*

Illya Bakurov and Leonardo Vanneschi are with the NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, 1099-085 Lisbon, Portugal (e-mail: ibakurov@novaims.unl.pt; lvanneschi@novaims.unl.pt).

Marco Buzzelli and Raimondo Schettini are with the Department of Informatics Systems and Communication, University of Milano–Bicocca, 20126 Milan, Italy (e-mail: marco.buzzelli@unimib.it; raimondo.schettini@unimib.it).

Mauro Castelli is with the NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, 1099-085 Lisbon, Portugal, and also with the School of Economics and Business, University of Ljubljana, 1000 Ljubljana, Slovenia (e-mail: mcastelli@novaims.unl.pt).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2023.3244662, provided by the authors.

Digital Object Identifier 10.1109/TIP.2023.3244662

a pristine reference image. When a measure assesses the similarity between a pristine reference image and its degraded variant, it is usually called full-reference IQAM (FR-IQAM). On the other hand, when the reference is not available, the measure is often called no-reference IQAM (NR-IQAM). Moreover, when the reference image is not completely provided (i.e., only some partial information is available, as a given set of the extracted features), the measure is frequently called reduced-reference IQAM (RR-IQAM). Historically, researchers handcrafted these measures based on their understanding of the human visual system, mathematics, and information theory. One of the most notable is the Structural Similarity Index Measure (SSIM) and its numerous variations [1]. The simplicity of the SSIM formulation resulted in high levels of human interpretability (each component amounting to the final SSIM value can be individually analyzed and assessed), as well as high efficiency.

Following a different direction from the development of handcrafted solutions, multiple attempts have been made to embed machine learning (ML) in the design process of image quality. This goes from learning the free parameters of handcrafted measures [3], [4], to exploiting deep learning for the training of an image similarity neural network [5], [6]. These approaches have successfully raised the bar in the effectiveness of image similarity. However, especially for deep learning methods, the resulting models tend to be particularly fine-tuned to a specific annotated dataset as they are notoriously data greedy. Furthermore, they are particularly computationally expensive, although hardware accelerators mitigate the problem.

In this paper, we present a different application of machine learning to the definition of image quality measures. Specifically, we explore a novel approach based on genetic programming (GP) to define a computationally-constrained formulation of FR-IQAMs. We use GP because GP is typically able to evolve solutions to complex problems without any predefined hypothesis on the shape of the model. Furthermore, GP is versatile, allowing the user to define the most appropriate language to code the evolving solutions. This is typically done by defining two sets: the set of primitive functions and the set of terminal symbols, that are used to construct the expressions representing the models. Last but not least, GP has reported a noteworthy number of practical successes in real-world applications in the last decade [7]. We explore different sets of terminals and primitive functions, all based on differentiable operators, thus making the final solutions potentially useful as loss functions in machine learning applications based on gradient backpropagation. We constrain the complexity of the tree of operations via hoist mutation during the optimization phase, and through mathematical simplification for the final application. We run our experiment in a cross-dataset scenario, compare selected individuals against existing image quality assessment measures, and provide an analysis of their structure and behavior.

The paper is organized as follows: Section II introduces the necessary theoretical background by providing an overview of SSIM and its variants, the conceptually different approaches to formulate novel IQAMs, and presents the reader with GP - the technique that sustains our approach. Section III describes the proposed GP-based approach to formulate novel FR-IQAMs based on SSIM. Section IV presents the research objectives, characterizes the datasets considered in our study, presents and discusses the hyper-parameters that were used, and shows the obtained results. Section V discusses the obtained experimental results. Finally, Section VI concludes the work and proposes ideas for future research.

II. Background and Related Literature

A. Structural Similarity

The Structural Similarity Index Measure (SSIM) compares a reference pristine image x and a corrupted version y, based on three components that are independently evaluated: luminance similarity, contrast similarity, and structure similarity [1]. The so-called "suggested usage" by Wang et al. [8] requires the grayscale conversion of color images, therefore always considering x and y as 2-dimensional matrices. The luminance information is represented by each image's average (μ), thus the luminance-based similarity is:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{1}$$

where $C_1 = (0.01 \cdot 255)^2$ is a small constant for numerical stability, as are $C_2 = (0.03 \cdot 255)^2$ and $C_3 = \frac{C_2}{2}$ in the following equations for the other components. Contrast [9] is represented through the use of standard deviation (σ), and consequently the contrast-based similarity is:

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2}$$

Finally, structure is computed by normalizing each image by the corresponding mean and variance. These are then compared with the inner product, computed through their covariance σ_{xy}:

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{3}$$

All statistics $\mu_{\{x,y\}}$, $\sigma_{\{x,y\}}$ and σ_{xy} are computed locally with a Gaussian weighting function. The three components l, c and s are then combined into an overall similarity map as:

$$SSIM(x, y) = \left[l(x, y)\right]^{\alpha} \cdot \left[c(x, y)\right]^{\beta} \cdot \left[s(x, y)\right]^{\gamma} \tag{4}$$

where exponents α, β and γ can be tuned to regulate the impact of the individual components in the overall similarity. When these exponents are all set equal to 1, the expression is simplified as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{x,y} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

Finally, the SSIM map is spatially averaged in order to produce a single output scalar similarity.

Wang et al. [10] pointed out that SSIM considered at a single scale (SS-SSIM) would be appropriate only for specific viewing conditions related to viewing distance and display size. To remedy this drawback, they proposed Multi-Scale SSIM (MS-SSIM) by aggregating intermediate similarity indexes from a range of different image resolutions:

$$MS\text{-}SSIM(x, y) = [l_M(x, y)]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j} \tag{6}$$

Moving from scale j to $j+1$, a low-pass filter followed by a down-sampling operation with a factor of 2, is applied over the reference-distortion pair. The luminance similarity, denoted by $l_M(x, y)$ is computed only at scale M due to its scale invariance.

The high popularity of SSIM and MS-SSIM has lead to numerous implementations, with varying performance. In 2021 Venkataramanan et al. [11] conducted a comparative analysis of such implementations, and by considering a variety of design choices formulated Enhanced SSIM (ESSIM), which defines guidelines on input color representation, window size and stride, and image rescaling.

Furthermore, a number of SSIM-inspired metrics have been proposed through the years. These will be presented in the next section, along with other hand-crafted measures.

B. Hand-Crafted FR-IQAMs

Among the most simple approaches for quality assessment is the use of non-image-specific measures such as Peak Signal-to-Noise Ratio (PSNR) or the Root Mean Square Error (RMSE). These are known to be suboptimal and poorly correlated to the response of the human visual system [10], and are commonly used as a baseline for comparison.

In addition, a wide variety of hand-crafted measures for FR-IQA has been developed through the years. Authors Damera-Venkata et al. [12] propose to decouple IQA through the independent evaluation of a Noise Quality Measure (NQM) and a Distortion Measure (DM). NQM models noise and contrast variations, accounting for spatial frequencies and masking effects, based on Peli's contrast pyramid. DM estimates the frequency of highest distortion and compares it with a no distortion reference taking into account the human contrast sensitivity function. The Visual Information Fidelity measure (VIF) [13] models image quality assessment as a distortion channel communication, based on a two-stage process: they first quantify the information loss from reference image to distorted image and, subsequently, correlate this quantity to perceived image quality. This approach was developed as an extension of the Information Fidelity Criterion (IFC) [14], which relies upon natural scene statistics in the context of the distortion channel to model the shared information as a proxy for image quality. The Riesz-transform based Feature SIMilarity metric (RFSIM) [15] computes 1st and 2nd order Riesz transform coefficients of the input images, which are derivative filters obtained as a vector-valued extension of the Hilbert transform. The coefficients in key locations identified by strong edges are then compared in pairs via harmonic mean, and combined via multiplication. Information-content Weighted SSIM (IW-SSIM) [16] is based on the observation that the pooling operation of local similarity should take into account the information content of the corresponding region, to be estimated using statistical models of natural images. The authors apply a similar concept to various measures, introducing consistent improvements. Authors of Most Apparent Distortion (MAD) [17] exploit the intuition that two different macro-categories of distortion levels can be identified and treated differently. Specifically, images with close-to-imperceptible distortions are modeled via local luminance and contrast masking, while extremely-low-quality images are characterized with spatial-frequency components. Feature SIMilarity (FSIM) [18] models the local image similarity by exploiting phase congruence to identify the significance of different image regions for overall quality assessment and gradient magnitude to account for contrast information in the reference-distortion comparison. Authors of the Gradient Similarity Measure (GSM) [19] propose an IQAM that exploits structure contrast and luminance similarity in an SSIM-inspired fashion, although incorporating mechanisms to model masking effects and modeling the relative importance of the similarity components with an adaptive weight approach. Similarly inspired, Gradient Magnitude Similarity Deviation (GMSD) [20], which was designed with a focus on computational efficiency, is obtained by extracting and comparing the pixel-wise gradient magnitude information from the input image pair, and by pooling the resulting map via standard deviation. Sparse Feature Fidelity (SFF) [21] is based on encoding input images into sparse representations that are designed to mimic the mechanisms of the primary visual cortex, after a training phase performed on natural images. Comparison between reference and distorted image is then computed both in terms of structural and brightness differences, a strategy common throughout several IQAMs. Pei et al. [22] propose the computation of several IQMs, including SSIM, on a version of the image encoded with Difference of Gaussians (DOG) at multiple frequency bands. The resulting DOG-SSIM metric is then obtained by non-linear combination of the SSIM-like comparison of DOG bands, based on a trained random forest regression model. Authors of Normalized Laplacian Pyramid Distance (NLPD) [23] propose decomposing the input image pair using a Laplacian pyramid representation and normalizing the resulting representation by a local amplitude estimate. The root mean squared error of the normalized features is then used as a proxy for image similarity. The Structural Contrast-Quality Index (SC-QI) [24] models FR-IQA by comparing a representation of the input images inspired by the structural contrast index (SCI). Based on the discrete cosine transform, SCI is used to estimate the perceptual complexity of image texture patterns, which in turn informs the property of image regions to be more or less sensitive to distortion.

The variety of SSIM-inspired metrics, as well as the high performance achieved by appropriate parameters selection as shown by ESSIM, are suggestive of the potential value of an SSIM-based approach. This motivates us to propose a novel solution for FR-IQA that is also based on SSIM components, however leveraging machine learning to optimize the combination of such components in a data-driven fashion.

Our approach can potentially approximate other solutions such as IW-SSIM, as it is based on the same underlying components (referred to as terminals in our framework). The combination of other terminals, such as gradient descriptions of the input images as done in GSM, or DCT-based features as done in SC-QI, is also theoretically possible in our framework, although we consider it as a direction for future developments, aiming for the generation of lower-complexity solutions.

C. Learning-Based FR-IQAMs

In more recent years, the attention of the scientific community has gravitated towards deep learning-based methods for image quality assessment.

DeepSim [5] exploits a pre-trained convolutional neural network (CNN) to extract layers activations of the input image pair at different levels. Local similarities are computed between the corresponding activation maps, and subsequently pooled to produce an overall quality score. Mid-level representations after Rectified Linear Units (ReLUs) and max-pooling operations were found to be the most useful for image quality assessment. DeepIQA [6] is a neural network specifically designed for image quality assessment, which can be trained end-to-end for quality score regression of either image pairs (FR-IQA) or single images (NR-IQA). The architecture is devised so as to allow the inspection of learned relative importance of local quality contributions. PieAPP [25] is a deep-neural-network-based perceptual quality metric. The underlying neural model is trained on comparative labels in a pair-wise learning, i.e. during training the objective is not to explicitly regress a similarity value for the input reference-distortion image pair, but to predict the preference between two distorted images. Authors of the Learned Perceptual Image Patch Similarity (LPIPS) [26] compared the effectiveness of neural features extracted from different deep architectures, as commonly done by so-called "perceptual loss functions" in the domain of image synthesis. They thus trained a neural model that assesses image similarity, akin to full-reference image quality assessment, in a patchwise fashion, tested with both both traditional distortions and CNN-based distortions. The Deep Image Structure and Texture Similarity index (DISTS) [27] was developed to explicitly embed tolerance to texture resampling (i.e. repositioning image patches). The underlying neural model first builds a multi-scale overcomplete representation of the input images. The correlations of spatial averages in the resulting feature maps are then combined to correlations of the feature maps themselves to model the overall image quality. Lukin et al. [28] approached the task of image quality assessment as a stacking problem: they apply multiple FR-IQAMs and combine the resulting assessments through a neural network, for which several configurations have been investigated.

Our solution to measuring image quality is a learning-based approach that does not involve deep neural models, relying instead on low-level features that are properly combined according to a tree of operations, optimized by targeting a combination of effectiveness and efficiency.

D. Genetic Programming

Genetic Programming is a population-based stochastic iterative search algorithm proposed and popularized by J. R. Koza [29], that extends genetic algorithms to explore the search space of computer programs. Like other evolutionary meta heuristics, GP evolves a set of candidate solutions (the population) by mimicking the basic principles of Darwinian evolution. The evolutionary process involves an iterative application of a fitness-based selection of the candidate solutions and their variation throughout genetically-inspired operators, such as crossover and mutation [29]. If abstracted from some implementation details, GP can be seen as a genetic algorithm in which the initialization and the variation operators were specifically adjusted to work on computer programs, typically (but not necessarily) represented as trees. In this form of representation, the evolving programs are constructed by composing elements belonging to two specific, predefined, sets: a set of primitive functions F, which appear as the internal nodes of the trees, and a set of terminals T, which represent the leaves of the trees. In the context of supervised ML problem-solving, the trees encode data-driven predictive models, often represented as mathematical expressions in the so-called Polish prefix notation, in which the operators (a.k.a. primitive functions) precede their operands (a.k.a. terminals).

Typically, GP is used with the so-called subtree mutation and swap crossover [29]. The latter exchanges two randomly selected subtrees between two different parent individuals. The former randomly selects a subtree in the structure of the parent individual and replaces it with a new, randomly generated tree.

E. Evolutionary Approaches to IQA

The application of genetic programming and other evolutionary techniques to the field of image processing and computer vision has been conducted for several decades, as documented in a recent survey by Khan et al. [30]. The authors present the different GP techniques applied to a wide variety of fields, such as image enhancement, compression, segmentation, classification, registration, retrieval and object detection. The last two categories, in particular, are tightly related to the task of image similarity, as they require the definition of a matching function between a query and elements in a dataset.

Most of the existing literature is, in fact, focused on image similarity for content-based image retrieval systems, exploiting genetic algorithms for image matching, and genetic programming to search a proper image similarity function.

Joshi and Tapaswi [31] propose an image retrieval solution that formulates image similarity at the level of content, whereas we focus on the level of signal and image quality. The authors devise an unsupervised segmentation technique to decompose the image into regions, which are described in terms of color and texture features, and compared using a genetic algorithm applied to feature selection. Syam and Rao [32] conduct image retrieval by extracting color, shape, texture and contourlet features from the database and query images. They then resort to genetic algorithms to search the database for the best matching items, using image indices as chromosome genes, and relying on the square euclidean distance to measure the fitness of the corresponding precomputed features.

Torres et al. [33] proved the effectiveness of genetic programming applied to image similarity. The authors focus on shape-based image retrieval, therefore exploiting a set of domain-specific terminals and fitness functions. In particular, they precompute similarity over beam angle statistics descriptors, multiscale fractal dimensions, Fourier decomposition, and moment invariants. These similarity values constitute the terminal set, and are combined using a function set composed of multiplication, addition, division, and square rooting. They exclude subtraction to prevent generation of negative values. In our case, we do not limit our terminals to precomputed similarities, and as such we consider the subtraction function a fundamental element in developing a new similarity measure. In a related work by Ferreira et al. [34], the authors extend the application of GP-based image similarity to color and texture similarity, thus introducing a larger set of precomputed similarity terminals. They also introduce mechanisms for relevance feedback in their image retrieval objective, and consequently reformulate the fitness functions to another set of domain-specific solutions. In our paper, we focus on image similarity for quality assessment, and our fitness function minimizes the correlation with human-provided judgements (mean opinion scores) on a number of datasets. Calumby et al. [35] also present a framework for image retrieval with relevance feedback. The authors use genetic programming to effectively learn a similarity perception function specific to the user interacting with the system, based on the provided feedback on a given set of proposed retrieved items. Similarly to previous works, they combine multiple pre-computed image similarity measures, relying on color and texture descriptors, and additionally incorporate textual similarity measures to include human-written labels in the retrieval loop.

Bakurov et al. [3], [4] focused on the existing single-scale SSIM formulation. They exploited several evolutionary computation techniques to estimate the best combination of luminance, contrast and structure components, through the search of parameters α, β and γ in Equation 4, as well as the optimal sliding window size used for processing. In a recent work [36], the authors also proposed a reformulation of SSIM that exploits strided convolutions and different filter sizes as a proxy for multi-scale analysis, optimizing their selection and combination through a set of genetic operators. By contrast, in this paper we completely deconstruct the original SSIM measure into its building components, defined at different levels of complexity and abstraction. Together with the introduction of a larger set of operators, as described in Section III, this enables the exploration of a wider space of SSIM-inspired similarity functions, which we extensively evaluate on a multi-dataset experimental setup.

III. Proposed Search for Image Quality Measures

In this study we propose to exploit GP to design novel IQAMs, inspired by the foundations of SSIM and the underlying theory of image quality assessment.

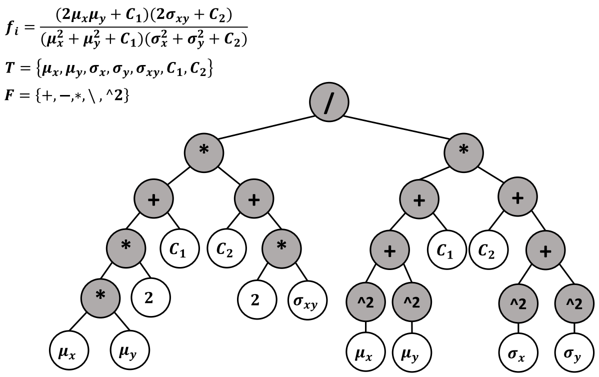

$f_i = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$
$T = \{\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy}, C_1, C_2\}$
$F = \{+, -, *, \backslash, \text{^}2\}$
/
*
*
+
+
+
+
*
C_1
C_2
*
+
C_1
C_2
+
*
2
2
σ_{xy}
^2
^2
^2
^2
μ_x
μ_y
μ_x
μ_y
σ_x
σ_y

Fig. 1. Tree-based representation of SS-SSIM for $\alpha = \beta = \sigma = 1.0$.

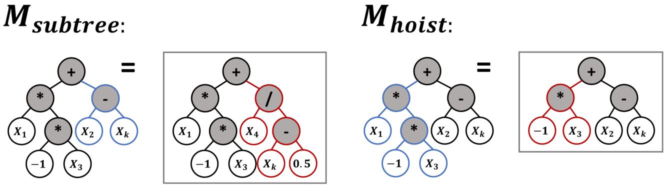


Fig. 2. Representation of how subtree and hoist mutations work.

Figure 1 shows with an example how single-scale SSIM (SS-SSIM) can be represented as a tree of program elements. Note that the SS-SSIM presented in this figure is a simplified version of the one presented in Equation 4 for $\alpha = \beta = \sigma = 1.0$ [1].

Given our interest in finding simple and human-interpretable similarity expressions, we exploit novel strategies to effectively initialize GP's population, and the conceptual differences between GP's mutation operators. Specifically, we propose to use the Evolutionary Demes Despeciation Algorithm (EDDA) initialization technique [37], [38], known for its ability to generate small and high-quality initial solutions, and split the evolutionary procedure in two: the first half uses subtree mutation, due to its exploratory stoutness, whereas the second half uses hoist mutation, due to its pruning properties, thus simultaneously encouraging smaller and high-quality solutions. Hoist mutation randomly selects a subtree in the parent individual's structure and replaces it with a subtree within itself; therefore, it is guaranteed that the mutant will be smaller in size. Notice that, in the case of the subtree mutation, the mutant's size may decrease or increase. Figure 2 provides a visual intuition to distinguish between the subtree and the hoist mutation operators. Subtree mutation produces an offspring that is more complex than its parent. Contrarily, the offspring produced by hoist mutation is smaller.

The pseudocode of our proposed method is shown in Algorithm 1: for a given terminal set T_i, we execute a first run of $g/2$ optimization generations using subtree mutation, and use the resulting population P to start a second run of $g/2$ optimization generations using hoist mutation.

To evaluate the proposed GP-driven formulation of novel FR-IQAMs, we use the sum of the absolute values of Spearman's rank correlation coefficient (SRCC) and Pearson's

Algorithm 1 Pseudocode of the Proposed GP-Driven Method for the Formulation of Image Similarity Expressions

```
Let $T = [T_L,\ T_{M_A}, T_{M_S},\ T_H]$ be a collection of terminals' sets.
Let KADID10K$_{train}$ and KADID10K$_{valid}$, respectively be the training and validation datasets.
Let $IQA\_DB$=[TID2008, TID2013, CSIQ, LIVE2, LIVEM, VDID2014, PieAPP] be a collection of unseen IQA datasets.
Let $g$ be the number of generations, $C$ the crossover operator and $M = [M_{subtree}, M_{hoist}]$ a collection of mutation operators.
• for $T_i$ in $T$:
    – generate $T_i$ for KADID10K$_{train}$ and KADID10K$_{valid}$
    – generate $T_i$ $\forall$ dataset in $IQA\_DB$
• split KADID10K into training and validation partitions: KADID10K$_{train}$ and KADID10K$_{valid}$, respectively;
• for $T_i$ in $T$:
    – initialize GP's population using EDDA technique;
    – execute GP for $g/2$ generations with $T_i$ and $M_{subtree}$ using KADID10K$_{train}$ and KADID10K$_{valid}$ as the training and validation datasets;
    – extract the final population $P$;
    – with $P$ as the initial population, execute GP for $g/2$ generations with $T_i$ and $M_{hoist}$, using KADID10K$_{train}$ and KADID10K$_{valid}$;
    – extract the best similarity expression $f_{ssim}$;
    – calculate SRCC of $f_{ssim}$ $\forall$ dataset in $IQA\_DB$;
```

linear correlation coefficient (PLCC), both for optimization and for final assessment. Preliminary experiments showed that optimizing for both objectives leads to generally simpler and more compact solutions with respect to only optimizing for rank correlation, while at the same time reaching comparable performance. The computed image quality scores of a given candidate solution, for a given IQA dataset, are correlated with the respective subjective evaluation provided by the human observers, such as mean opinion score (MOS) and, in the case of VDID2014, differential MOS (DMOS).

We explore four different terminal sets, leveraging the SSIM components at various levels of detail. Sections III-A, III-B, III-C and III-D motivate and describe in detail each set, along with the corresponding functions. In addition, all optimized similarity expressions are further processed with a sigmoid function S:

$$S(x) = \frac{1}{1+e^{-x}}, \tag{7}$$

where x here represents the output of our genetically-optimized expression for image quality assessment. Preliminary experiments showed that this leads to solutions that are, on average, characterized by superior correlation with MOS (an effect that we attribute to the introduced non-linearity) in addition to the practical advantage of producing outputs that are constrained in a finite interval. The original SSIM and MS-SSIM formulations apply Equation 4 and 6 to spatially-varying information extracted with local statistics, thus obtaining a map of local similarity, which is eventually averaged:

$$*\text{-}SSIM(x, y) = mean(f(g(x, y))) \tag{8}$$

where f represents the generic application of any combination (such as Equations 4 and 6) of the local statistics maps extracted with g (such as μ and σ). In this work, we are in fact optimizing the structure of combination function f itself. Exploring the space of combination functions with full-size maps from g is here considered to be beyond practical computational constraints: these functions would be computed and processed on-the-fly at each optimization iteration, thus extending the duration of the optimization phase, and/or be stored for memoization, thus bloating memory occupation. For these reasons, we opted to reformulate the similarity computation as:

$$*\text{-}SSIM(x, y) = f(mean(g(x, y))) \tag{9}$$

that is, we first aggregate by average the extracted local statistics g, corresponding to the terminals described in the following sections, and then we apply the combination function f, optimized through GP. In the experimental section we will also explore the impact that this shift has on the original SSIM formulation, for all the terminals set defined in the following.

A. Low-Level Terminals

The terminal set with the lowest level of abstraction involves the application of a filtering function G to process two input images X and Y (standing for the reference and the distortion, respectively), using a Gaussian kernel with a standard deviation of 1.5 and window-size of 11 × 11 pixels:

- $T_{L1} := G(X)$
- $T_{L2} := G(X^2)$
- $T_{L3} := G(Y)$
- $T_{L4} := G(Y^2)$
- $T_{L5} := G(X \cdot Y)$,

where the power and the multiplication operators are point-wise functions. These terminals, from now on called *low-level* (T_L), serve as the essential building-blocks of SSIM, and can be further used to formulate more abstract components such as the luminance, contrast and structural similarity.

The set of functions strictly necessary to reach the full SSIM formulation from these terminals include subtraction, multiplication, and exponentiation to the power of 2, as highlighted in Section III-B. We extend this function set with the addition and division operators, in order to enlarge the space of explored solutions.

TABLE I

SUMMARY FEATURES OF THE CONSIDERED IQA DATASETS. NOTICE THAT THE COLUMN D/H_i RELATES TO VIEWING DISTANCE, WHILE #REF., #DIST., AND #PAIRS REFER TO THE NUMBER OF REFERENCE IMAGES, DISTORTION TYPES, AND RESULTING REFERENCE-DISTORTION PAIRS

Name	Resolution (px)	D/H_i	#Ref.	#Dist.	#Pairs
KADID10K [39]	512×384	varying	81	25	10125
TID2008 [40], [41]	512×384	3	25	17	1700
TID2013 [42]	512×384	3	25	24	3000
LIVE2 [43], [44]	480 (min) ÷ 768 (max)	3÷3.75	29	5	779
LIVEM [45]	1280×720	4	15	3	405
VDID2014 [46]	768×512, 512×512	4, 6	8	4	160
CSIQ [47], [48]	515×515	5	30	6	866
PieAPP [25]	256×256	varying	200	75	20280

B. Middle-Level Terminals (Asymmetric)

The terminals presented in Section III-A can be translated into more abstract terms that, in the scope of the GP search process, can be potentially combined into SSIM itself:

- $T_{M_A 1} := \mu_X = T_{L1}$
- $T_{M_A 2} := \mu_Y = T_{L3}$
- $T_{M_A 3} := \mu_X^2 = T_{L1}^2$
- $T_{M_A 4} := \mu_Y^2 = T_{L3}^2$
- $T_{M_A 5} := \sigma_X^2 = T_{L2} - T_{L1}^2$
- $T_{M_A 6} := \sigma_Y^2 = T_{L4} - T_{L3}^2$
- $T_{M_A 7} := \sigma_X \sigma_Y = \sqrt{(T_{L2} - T_{L1}^2) \cdot (T_{L4} - T_{L3}^2)}$
- $T_{M_A 8} := \sigma_{XY} = T_{L5} - T_{L1} \cdot T_{L3}$

Note that a solution (similarity expression) computed from this terminal set, or from the low-level terminal set, will not in general be symmetric (i.e. $f(X, Y) \neq f(Y, X)$). Therefore, the resulting behavior is more fit for a FR-IQAM, where either of the two inputs is defined as the reference, as opposed to a more generic image similarity function. In this sense, we decided to label this set of terminals *asymmetric mid-level* (T_{M_A}).

The set of functions that allow to define and combine the luminance, contrast, and structural similarities, include multiplication, addition and division, as shown in Equations 1 to 4. We extend this function set by also including the subtraction, and the power of two, in line with the T_L terminal set.

C. Middle-Level Terminals (Symmetric)

The third set, from now on called *symmetric mid-level* (T_{M_S}), increments the abstraction level as it covers the same pre-computed μ and σ statistics from T_{M_A}, although paired so that each individual terminal is symmetric and, by extension, any resulting similarity expression will be symmetric as well:

- $T_{M_S 1} := \mu_X \cdot \mu_Y$
- $T_{M_S 2} := \mu_X^2 \cdot \mu_Y^2$
- $T_{M_S 3} := \mu_X + \mu_Y$
- $T_{M_S 4} := \mu_X^2 + \mu_Y^2$
- $T_{M_S 5} := \sigma_X \cdot \sigma_Y$
- $T_{M_S 6} := \sigma_X^2 \cdot \sigma_Y^2$
- $T_{M_S 7} := \sigma_X + \sigma_Y$
- $T_{M_S 8} := \sigma_X^2 + \sigma_Y^2$
- $T_{M_S 9} := \sigma_{XY}$

A solution computed from this terminal set can be applied for image similarity instead of the more specific FR-IQAM, because it does not require the explicit indication of a reference and distorted image.

The set of functions that allow to define and combine the original SSIM components only include division and multiplication, according to Equations 1 to 4. We extend this function set by also including the addition and subtraction, to once again enlarge the space of explored solutions.

D. High-Level Terminals

The fourth and the last terminal set, from now on called *high-level* (T_H), increases the level of abstraction by including precomputed SSIM's major components: the luminance, the contrast and the structural comparison measures, defined by Equations 1, 2 and 3, respectively. Following the rationale behind MS-SSIM [10], we considered it necessary to estimate the luminance, contrast and structure components from a range of different spatial-scales, to account for the varying viewing conditions (such as the display resolution and the distance from the display to the observer):

- $T_{H1} := l_{scale=1}(X, Y)$
- $T_{H2} := c_{scale=1}(X, Y)$
- $T_{H3} := s_{scale=1}(X, Y)$
- $\ldots$
- $T_{H(N-2)} := l_{scale=M}(X, Y)$
- $T_{H(N-1)} := c_{scale=M}(X, Y)$
- $T_{H(N)} := s_{scale=M}(X, Y)$

Following the work of Wang et al. [10], we set the number of scales $M = 5$. Therefore, the total number of multiscale high-level terminals ($T_{H_{MS}}$) is 15. We will also experiment with a single-scale variant of the same high-level terminals ($T_{H_{SS}}$).

The function set necessary to reach the same complexity as MS-SSIM includes the multiplication, and the exponentiation operator. We select a discrete set of exponents to be explored through GP: [0.05, 0.15, 0.30, 0.50, 0.80, 1/0.80, 1/0.50, 1/0.30, 1/0.15, 1/0.05]. These, when combined through the power rule for exponents, allow the GP to potentially reach exponentiations that are outside the initial set. We also integrate the above function set with the inclusion of addition, subtraction, and division operators.

IV. EXPERIMENTS

A. IQA Datasets

We explore the feasibility of the proposed approach using five well-known datasets for assessing image quality aspects. In this section, the reader can find a detailed description of each dataset. For a summarized description, the reader is referred to Table I.

1) KADID10K: The Konstanz Artificially Distorted Image quality Database (KADID-10k) is a large dataset for full-reference image quality assessment, designed with the specific goal of building a collection of annotated image pairs that is suitable for the training of data-hungry machine learning models [39]. The large scale annotation objective was achieved with the use of crowdsourcing platform figure-eight.com. The remote workers were asked to rate an image pair on a five-point scale, ranging from very annoying to imperceptible.

Due to the inherent variety of devices and experimental environments, no information about viewing distance is available. The KADID-10k dataset contains 81 pristine images selected from Pixabay.com and rescaled to 512 × 384 pixel resolution, each degraded by 25 distortions in 5 levels, for a total of 10125 distorted images with 30 quality scores each.

2) TID2008: The Tampere image dataset 2008 (TID2008) is a well-known and publicly available dataset specifically designed for the evaluation of full-reference image visual quality assessment metrics [40], [41]. The central aspect of this dataset is that it was created upon reference images that account for a wide variety of visual scenes and contains several different types of distortion that relate to various peculiarities of the human visual system. More specifically, TID2008 was built from 25 512 × 384px reference images taken from the Kodak lossless true-color image suite [49], apart from one artificially synthesized image. For each reference image, authors have applied 17 types of distortions with four different levels for each type of distortion, resulting in 1700 reference-distortion pairs. To subjectively evaluate the visual quality of distorted images, more than 800 volunteers were involved. Moreover, to remove possible judgemental bias, volunteers with different cultural levels (researchers, tutors, and students) from three different countries (Finland, Italy, and Ukraine) were involved. The subjective test was carried out at the viewing distance of three times the image height. In total, about 256’000 individual human quality judgments were performed, and, as a result, MOS values were obtained. Further details about the dataset, namely a complete enumeration of distortion types and levels, can be found in [40] and [41].

3) TID2013: The Tampere image dataset 2013 (TID2013) is an extension of the aforementioned TID2008 and contains more distortion types and levels [42]. This dataset is publicly available and rapidly became popular in the scientific community. The authors motivated the creation of TID2013 mainly by the new types of distortions and improved methodologies of quantitative subjective tests. More specifically, they re-utilized the reference images used for TID2008 and applied 24 types of distortions, with five different levels each, resulting in a dataset containing 3000 reference-distortion pairs. The visual quality of the reference-distortion pairs was assessed through 985 subjective experiments with volunteers from five different countries (Finland, France, Italy, Ukraine, and the USA). Similarly to TID2008, the subjective test was carried out at the viewing distance of 3 times the image height. In total, about 524’340 individual human quality judgments were performed, and, as a result, MOS values were obtained. Further details about the dataset, namely a complete enumeration of distortion types and levels, can be found in [42].

4) LIVE2: The Laboratory for Image & Video Engineering (LIVE) database (conceived in the university of Texas at Austin) is one of the most renowned datasets in the research community [43]; in our research, we rely on its second release (made available in 2006). The dataset contains 779 reference-distortion pairs, generated from artificially corrupting 29 reference images using 5 distortion types: JPEG compression, JPEG2000 compression, Gaussian blur, white noise, and bit errors in JPEG200 bit stream (169, 175, 145, 145, and 145 reference-distortion pairs, respectively). The subjective test was carried out at a viewing distance of 3 to 3.75 times the image height, and the scores are reported in the form of DMOS.

5) LIVEM: The LIVE Multiply Distorted Image Quality Database (LIVEM) was collected with the specific goal of monitoring the quality of visual content that may be corrupted by multiple distortions [45], a line of research correlated with the increasing efforts to improve bandwidth usage in more realistic scenarios [50]. The dataset collects opinion scores from 37 subjects, for a total of 8880 judgments on 15 pristine reference images and 405 multiply-distorted images of two types: blur followed by JPEG, and blur followed by noise. The subjects, mostly male between 23 and 30 years old, were presented each stimulus for 8 seconds from a distance approximately equal to 4 times the screen height, and were subsequently asked to provide a quality value on a continuous scale from 0 to 100.

6) VDID2014: The VDID2014 is a viewing distance-changed IQA dataset, first published in 2015 to deploy the impact of viewing distances and image resolutions on IQA [46]. VDID2014 was built from eight reference images with resolutions of 768 × 512px and 512 × 512px. Note that the largest four are original from the Kodak lossless true-color image suite [49]. For each reference image, the authors have applied four types of distortions with five different levels for each type, resulting in a dataset containing 160 reference-distortion pairs. Twenty different volunteers subjectively evaluated the visual quality of distorted images. The subjective test was carried out at a viewing distance of four and six times the image height. The authors reported their results in the form of DMOS. Further details about the dataset can be found in [46].

7) CSIQ: The computational and subjective image quality (CSIQ) dataset is another popular dataset for the evaluation of image quality aspects. The main reason for the inclusion of this dataset in our benchmark was the fact that it was built upon a completely different set of reference images than those in TID2008, TID2013, and VDID2014. The CSIQ dataset was conceived from 30 512 × 512px reference images taken from public-domain sources, predominantly from the United States national park service. For each reference image, the authors applied six types of distortions, with five different levels for each type, resulting in a dataset containing 866 reference-distortion pairs. Thirty-five different volunteers subjectively evaluated the visual quality of the reference-distortion pairs. In total, 5000 individual human quality judgments were performed. The subjective test was carried out at a viewing distance of five times the image height. Unlike for TID2008, the authors reported their results in the form of differential MOS (DMOS), where larger values stand for greater visual distortion when compared to the reference. For this reason, a high negative correlation is expected between FR-IQA measures and DMOS. Further details about the dataset, namely a complete enumeration of distortion levels, can be found in [47].

8) PieAPP: The PieAPP dataset was collected in conjunction with the development of the corresponding FR-IQAM [25]. It separates the training and test image pairs both in terms of reference images and types of distortions, with 160 references for training affected by 75 distortions, and 59 references for test affected by 31 distortions. Such distortions include traditional FR-IQA artifacts as well as more complex artifacts related to the application of various computer vision and image processing algorithms. An important characteristic of this dataset is that annotations are provided in terms of the probability that humans will prefer one distorted image over another, in a comparison with the reference. These are collected via Amazon Mechanical Turk on a statistically significant, but relatively small, sample of all possibile combination pairs, and then transferred to a larger set of pairs. In our experiments, we considered the test set of PieAPP, composed of 600 reference-distortion pairs.

These datasets differ not only in terms of reference images, distortion types and magnitudes, viewing distance and image resolution, but also in terms of the quality-labeling schemes. For example, to collect the subjective quality scores, the authors of TID2008, TID2013 and PieAPP rely upon tristimulus methodology (where the latter database focuses exclusively on the probability of pairwise preference to avoid the set-dependence of Swiss tournaments characteristic of the formed two datasets). On the other hand, the authors of KADID10K and LIVE2 rely upon the double stimulus impairment scale method where the test sequences are presented in pairs. Contrariwise, a single stimulus method was used in LIVEM and VDID2014; here, the test sequences are presented one at a time and are rated independently. In our experiments, one of the datasets (KADID10K) is used for training and the remaining six to assess the robustness of the proposed approach towards different experimental conditions, including different quality-labeling schemes. The experimental results in Section V show that our method is robust to previously unseen experimental conditions, including the quality-labeling. Specifically, both Pearson's and Spearman's correlation coefficients show that our method significantly improves upon the respective baseline measures. Additionally, we consider the possibility to train upon pairwise probability of preference as ground-truth labels as an interesting direction for future extensions of our proposed method.

B. Data Usage

A common issue with supervised machine learning is overfitting: a situation when the algorithm learns overly-specific correlations to the point of memorizing samples and fails to generalize on unseen data. We exploit GP to formulate novel FR-IQAMs that effectively resemble human subjective evaluation on a wide variety of visual scenes and viewing conditions. Thus, it is of our primary interest that the obtained similarity expressions achieve superior performance on several different FR-IQA datasets and not just the ones "seen" during the training phase. To ensure that, we performed both inter and intra-dataset cross-validation. Specifically, we chose one IQA dataset for training and validating the proposed GP approach, whereas the remaining datasets were used to assess the generalization ability of the solutions learned. Taking into account the size and the diversity of datasets considered in our study, we chose TID2013 for training and validation, leaving the remaining datasets for an unbiased assessment of the generalization ability. Specifically, during the training at the beginning of each run, 70% of TID2013's reference images (along with the respective distortions) were randomly selected for optimization and the remaining 30% for assessment (referred to as "validation", in the continuation of the paper).

TABLE II

ENUMERATION OF GP HYPER-PARAMETERS. NOTE THAT $P(C)$ AND $P(M)$ INDICATE THE CROSSOVER'S AND THE MUTATION'S PROBABILITIES

Parameters	Values
№runs	30
№generations	20
Population's size	$\{100_{EDDA}$, 1000$\}$
Terminals (T)	$\{T_L,\ T_{M_A}, T_{M_S},\ T_H\}$
Functions (F)	{+, -, x, /, MEAN, MAX, MIN, x^i}, $i \in$ [0.05, 0.15, 0.3, 0.5, 0.8, 1/0.8, 1/0.5, 1/0.3, 1/0.15, 1/0.05]}
Initialization	$EDDA_5$ with 100% subtree mutation
Selection	tournament with selection pressure of 8%
Crossover	swap crossover
Mutation	{subtree mutation, hoist mutation}
$P(C)$	{0.7, 0.05}
$P(M)$	{0.3, 0.95}
Stopping criteria	№generations

C. Experimental Setup

This section presents the experimental settings. Table II provides a complete enumeration of the experimental parameters for the proposed GP-driven approach.

Taking into account the stochastic nature of Genetic Programming, we repeated the experiments 30 times (runs), each with a different seed for the pseudo-random numbers generator used to partition the data and to initialize and execute the algorithms. Throughout our experiments, we guaranteed, for each experiment, an equal computational effort (measured as the total number of fitness evaluations per run). Specifically, given that we used EDDA initialization technique with 100 individuals per deme, each left to evolve for 5 generations ($EDDA_5$), the effort to generate 1000 initial solutions for the GP algorithm comprises $1000 \times 100 \times 5 = 500000$ fitness evaluations. The main evolutionary process requires $1000 \times 20 = 20000$ fitness evaluations; therefore, the total computational effort comprises 520000 fitness evaluations per run.

The selection was tournament-based, with pressure 8%. The pressure was slightly increased from the traditionally used 5% to stimulate faster convergence given the relatively small amount of generations [29]. During EDDA's initialization, subtree mutation was the only variation operator being used in order to foster the search-space exploration. Once the population was initialized, the first half of GP's main evolutionary process (i.e., the first 10 generations) was conducted using the swap crossover and the subtree mutation, with probabilities 0.7 and 0.3, respectively. In the second half of the process (i.e., the remaining 10 generations), the hoist mutation was instead used, and the mutation's probability was increased to

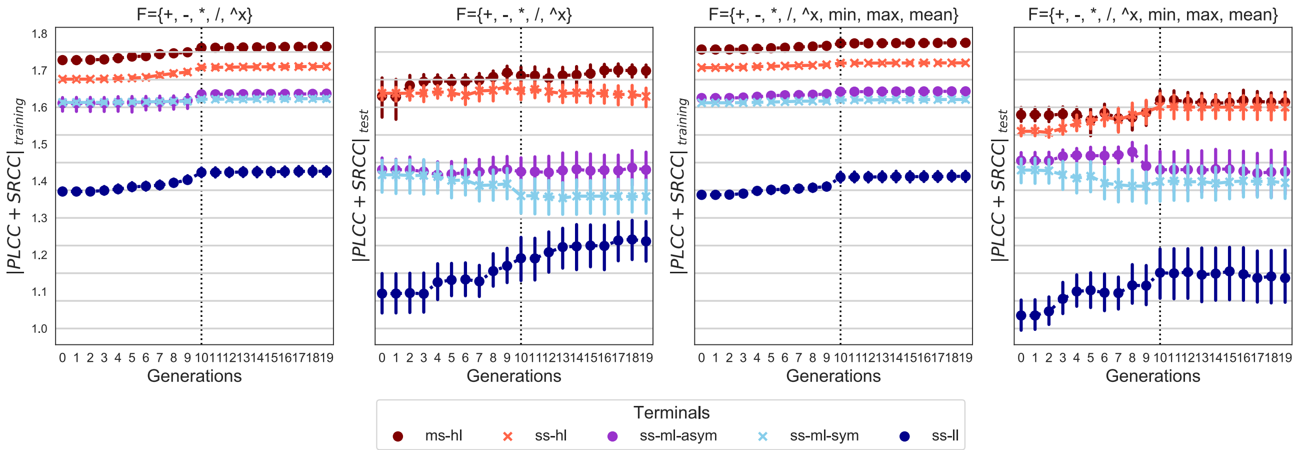


Fig. 3. Visualization of the learning curves computed on the training and test sets of KADID10K for different configurations of terminals and functions: single scale (ss) vs. multi-scale (ms); low-level (ll) vs. middle-level asymmetric (ml-asym), vs. middle-level symmetric (ml-sym), vs. high-level (hl) terminals.

0.95 to foster the pruning of the trees (this leaves 0.05 to crossover's probability).

As described in Section III, the set of functions F depends on the choice of terminals, and it includes the traditional arithmetic operators ($\{+, -, \text{x}, /\}$), as well as an exponentiation operator that raises an input terminal to the power of a given exponent (x^i}, $i \in$ [0.05, 0.15, 0.3, 0.5, 0.8, 1/0.8, 1/0.5, 1/0.3, 1/0.15, 1/0.05]); all the elements of F admit two operands, except for the exponentiation operator which takes only one. In addition to the operators that are tightly related to the existing SSIM and MS-SSIM formulations, we considered the inclusion of the minimum, the maximum, and the mean operators (i.e., two values are taken at each index and the operator is applied) as they were successfully used for the fusion of saliency estimation maps through genetic programming [51]. Therefore, we divided all our experiments into two groups: with and without {MEAN, MAX, MIN}.

It is necessary to point out that, for each experiment, the initial population of the GP was seeded with the default version of SSIM (regarding the underlying set of terminals). The rationale behind this was to foster convergence and avoid the evolution to take an undesirable and, consequently, counter-productive path by providing a high-quality starting genotype.

V. Experimental Results

A. Results of GP Optimization

In this section, we show and discuss the experimental findings. We start with GP's learning curves because their analysis is essential to demonstrate the utility of the proposed approach and identify the potential overfitting. Figure 3 is made of four sub-plots, alternating training and test statistics on the KADID10K dataset, respectively without and with {MEAN, MAX, MIN} operators. Every sub-plot presents five colored curves, each color representing a different terminal set: dark blue, purple, sky blue, red and dark red stand for T_L, T_{M_A}, T_{M_S}, T_H (single-scale) and T_H (multi-scale), respectively. The curves were obtained by aggregating the elites' validation fitness across 30 runs. The vertical line in the middle of the plots (at generation 10), represents the transition point between the two phases of GP (characterized by different mutation strategies). By observing the figure, one can notice that whenever overfitting patterns emerge, manifested by decreasing test fit curves, the introduction of hoist mutation prevents further degradation in most cases. Additionally, one can see that using 20 generations is sufficient to achieve a fair level of convergence: aggregated objective function values tend to stabilize in the last generations, suggesting that further optimization would not bring a significant improvement. When comparing the curves between the two sets of functions, we can observe that there is a slight advantage from using the {MEAN, MAX, MIN} operators in the function set. Moreover, by adding the {MEAN, MAX, MIN} operators to the function set, it is possible to evolve the single-scale similarity expressions based on T_H to a performance level comparable with those evolved at multiple scales.

Table III shows summarized SRCC and PLCC values obtained by the elite individuals trained on KADID10K and extracted at the end of the evolutionary process. For each combination of terminals and functions that characterize our experiments, the table shows the maximum, the average, and the standard deviation of the fitness values in the columns "MAX", "AVG" and "STD", respectively, aggregated for both operator sets (with and without {MEAN, MAX, MIN}). Moreover, the column "SSIM"' provides the baseline estimate of SRCC and PLCC for each terminal set calculated using the spatially-aggregated features according to Equation 9. For this reason, the rank-based correlation with the subjective evaluation might not correspond to that observed in the literature. For example, the value of SSIM for CSIQ (0.867) indicates the SRCC of SS-SSIM calculated from $T_{H_{SS}}$. From the table, we can see that the elite individuals obtained by the proposed approach exhibit a larger or comparable SRCC and PLCC with the subjective judgment in absolute (MAX) and average terms (AVG) when compared to the baseline.

B. Analysis of Selected Individuals

In this section, we analyze some of the solutions obtained with the proposed optimization strategy across the various terminal levels. We remind the reader that all solutions are

TABLE III

CROSS-DATASET ASSESSMENT OF ELITE INDIVIDUALS AT THE LAST GENERATION OPTIMIZED ON KADID10K. FOR EACH COMBINATION OF TERMINAL SET AND DATASET, WE REPORT SRCC STATISTICS (MAXIMUM, AVERAGE AND STANDARD DEVIATION), AS WELL AS THE SSIM-EQUIVALENT THAT CAN BE OBTAINED FROM THE CORRESPONDING TERMINAL SET. PLCC STATISTICS ARE PROVIDED IN THE SUPPLEMENTARY MATERIALS

SRCC

	TID2008				TID2013				LIVE2				LIVEM			
	SSIM'	MAX	AVG	STD	SSIM'	MAX	AVG	STD	SSIM'	MAX	AVG	STD	SSIM'	MAX	AVG	STD
T_L	0.635	0.752	0.683	0.025	0.666	0.729	0.699	0.011	0.874	0.882	0.875	0.007	0.749	0.767	0.731	0.029
T_L (MMM)	0.635	0.768	0.686	0.034	0.666	0.754	0.700	0.017	0.874	0.889	0.873	0.008	0.749	0.774	0.714	0.034
T_{M_A}	0.448	0.828	0.799	0.015	0.436	0.761	0.728	0.018	0.670	0.914	0.894	0.022	0.396	0.859	0.853	0.008
T_{M_A} (MMM)	0.448	0.835	0.799	0.013	0.436	0.760	0.726	0.016	0.670	0.916	0.899	0.016	0.396	0.861	0.855	0.005
T_{M_S}	0.771	0.843	0.804	0.015	0.729	0.764	0.738	0.011	0.899	0.901	0.885	0.017	0.828	0.856	0.852	0.005
T_{M_S} (MMM)	0.771	0.824	0.800	0.012	0.729	0.763	0.736	0.010	0.899	0.900	0.878	0.045	0.828	0.856	0.853	0.004
$T_{H_{SS}}$	0.791	0.863	0.848	0.006	0.753	0.791	0.780	0.005	0.900	0.892	0.884	0.006	0.850	0.846	0.837	0.006
$T_{H_{SS}}$ (MMM)	0.791	0.865	0.847	0.008	0.753	0.796	0.780	0.007	0.900	0.900	0.892	0.003	0.850	0.851	0.843	0.007
$T_{H_{MS}}$	0.857	0.874	0.856	0.007	0.788	0.794	0.777	0.008	0.903	0.918	0.903	0.012	0.831	0.861	0.844	0.014
$T_{H_{MS}}$ (MMM)	0.857	0.892	0.867	0.016	0.788	0.808	0.788	0.014	0.903	0.913	0.902	0.004	0.831	0.861	0.845	0.013

	VDID2014				CSIQ				PieAPP			
	SSIM'	MAX	AVG	STD	SSIM'	MAX	AVG	STD	SSIM'	MAX	AVG	STD
T_L	0.909	0.919	0.909	0.007	0.810	0.854	0.818	0.016	0.199	0.231	0.205	0.011
T_L (MMM)	0.909	0.917	0.903	0.014	0.810	0.857	0.827	0.016	0.199	0.257	0.209	0.015
T_{M_A}	0.612	0.918	0.902	0.009	0.345	0.937	0.856	0.046	0.159	0.389	0.237	0.062
T_{M_A} (MMM)	0.612	0.916	0.902	0.009	0.345	0.927	0.859	0.047	0.159	0.267	0.205	0.026
T_{M_S}	0.909	0.906	0.895	0.010	0.885	0.924	0.882	0.017	0.234	0.342	0.248	0.039
T_{M_S} (MMM)	0.909	0.905	0.895	0.013	0.885	0.904	0.872	0.034	0.234	0.341	0.242	0.035
$T_{H_{SS}}$	0.895	0.920	0.911	0.009	0.867	0.919	0.886	0.020	0.293	0.303	0.286	0.008
$T_{H_{SS}}$ (MMM)	0.895	0.918	0.911	0.004	0.867	0.926	0.859	0.033	0.293	0.330	0.305	0.016
$T_{H_{MS}}$	0.901	0.928	0.920	0.004	0.915	0.929	0.860	0.044	0.285	0.370	0.292	0.024
$T_{H_{MS}}$ (MMM)	0.901	0.927	0.916	0.006	0.915	0.953	0.883	0.066	0.285	0.313	0.294	0.013

TABLE IV

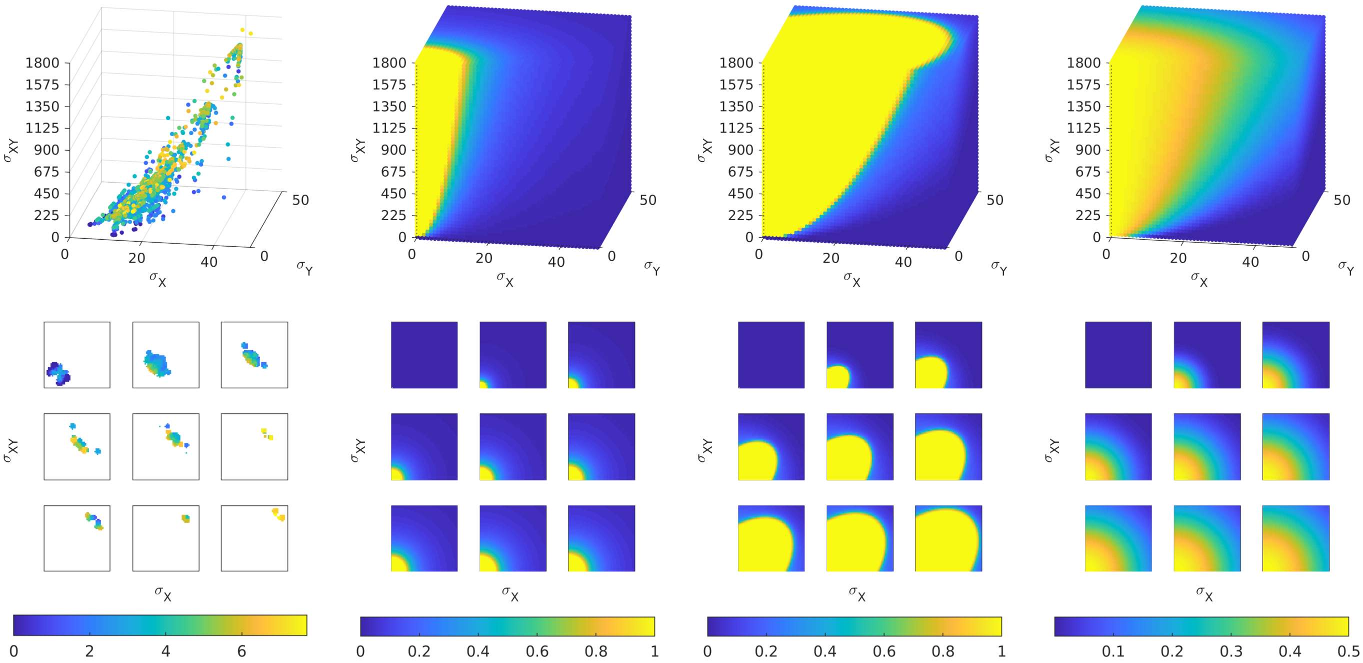
SRCC AND PLCC STATISTICS OF SELECTED INDIVIDUALS OBTAINED FROM THE CROSS-DATASET GENETIC PROGRAMMING OPTIMIZATION WITH DIFFERENT TERMINAL SETS. WE REPORT DIFFERENT VERSIONS OF STRUCTURAL SIMILARITY (SSIM) FOR DIRECT COMPARISON

IQAM	SRCC TID2008	TID2013	LIVE2	LIVEM	VDID2014	CSIQ	PieAPP
SSIM	0.775	0.741	0.901	0.850	0.842	0.876	**0.316**
OP-SSIM [4]	0.858	**0.792**	0.901	0.845	**0.949**	0.864	0.276
ESSIM [11]	0.762	0.762	**0.938**	**0.861**	0.912	0.866	0.244
GP-SSIM T_L	0.752	0.729	0.881	0.738	0.913	0.850	0.194
GP-SSIM T_{M_A} (MMM)	0.811	0.732	0.901	**0.861**	0.896	0.918	0.203
GP-SSIM T_{M_S}-A	0.787	0.734	0.889	0.853	0.897	0.870	0.250
GP-SSIM T_{M_S}-B	0.788	0.725	0.897	0.854	0.900	0.886	0.237
GP-SSIM $T_{H_{SS}}$-A	0.854	0.784	0.888	0.846	0.911	**0.919**	0.286
GP-SSIM $T_{H_{SS}}$-B (MMM)	**0.860**	0.790	0.893	0.851	0.912	0.914	0.295
MS-SSIM	0.853	0.787	**0.945**	0.845	0.900	0.913	**0.321**
OP-MS-SSIM [4]	0.871	**0.805**	0.909	**0.879**	0.899	0.927	0.286
GP-SSIM $T_{H_{MS}}$-A	0.874	0.794	0.904	0.851	0.917	0.921	0.286
GP-SSIM $T_{H_{MS}}$-B (MMM)	**0.878**	0.794	0.913	0.860	**0.927**	**0.953**	0.263
IQAM	**PLCC TID2008**	**TID2013**	**LIVE2**	**LIVEM**	**VDID2014**	**CSIQ**	**PieAPP**
SSIM	0.773	0.790	**0.945**	0.800	0.826	0.861	0.245
OP-SSIM [4]	0.783	0.771	0.814	0.827	0.810	0.866	0.250
ESSIM [11]	0.703	0.713	0.612	0.781	0.741	0.664	0.024
GP-SSIM T_L	0.718	0.754	0.862	0.760	0.887	0.826	**0.310**
GP-SSIM T_{M_A} (MMM)	0.798	0.764	0.905	0.849	0.881	**0.909**	0.241
GP-SSIM T_{M_S}-A	0.785	0.781	0.884	0.867	0.868	0.843	0.174
GP-SSIM T_{M_S}-B	0.737	0.728	0.785	0.806	0.729	0.733	0.102
GP-SSIM $T_{H_{SS}}$-A	0.843	0.839	0.851	0.857	0.865	0.889	0.286
GP-SSIM $T_{H_{SS}}$-B (MMM)	**0.852**	**0.844**	0.854	**0.876**	**0.892**	0.896	0.240
MS-SSIM	0.839	0.833	**0.933**	**0.883**	0.891	0.897	0.051
OP-MS-SSIM [4]	0.785	0.773	0.746	0.785	0.829	0.803	0.184
GP-SSIM $T_{H_{MS}}$-A	**0.846**	**0.837**	0.859	0.866	0.876	0.877	**0.307**
GP-SSIM $T_{H_{MS}}$-B (MMM)	0.824	0.814	0.899	0.740	**0.895**	**0.918**	0.295

For each row group: best results in bold, and second best underlined.

to be intended as further processed by a non-linarity in the form of a sigmoid function. Table IV shows the performance of selected individuals, evaluated in the cross-dataset scenario from Table III, directly compared with the original SSIM and MS-SSIM formulation, as well as an optimized version of both metrics (OP-SSIM and OP-MS-SSIM) from the work of Bakurov et al. [4], and ESSIM from the work of Venkataramanan et al. [11]. Our GP-based individuals show that, with a proper combination of high-level terminals, it is possible to reach performance comparable or even superior to MS-SSIM while resorting to only single-scale information. OP-SSIM [4] (both in single and multi scale versions) shows competitive performance in terms of rank correlation, however, when tested in terms of linear correlation, it is always outperformed by the proposed expression optimization. A similar observation can be derived from the analysis of ESSIM [11] results: by resorting to the official code implementation [52], we were able to run the method on all yuv420p-encoded datasets and to extract both SRCC and PLCC statistics. The resulting performance shows superior results for ESSIM on the LIVE2 and LIVEM datasets in terms of rank correlation, but significantly lower performance for other datasets and for all linear correlations. Additional comparisons with methods from the state of the art are provided later in Section V-C.

The expressions for some of these individuals are particularly compact, as shown in the following, while we refer the reader to the supplementary material for additional individuals. The middle-level symmetric solutions (T_{M_S}) exploit a set of terminals that includes both local average μ and local variance σ. However, some of the output solutions completely exclude μ components. Considering that μ_X and μ_Y are used in the original SSIM formulation to define the luminance component, this result is coherent with the findings of Bakurov et al. [3], [4], according to which the luminance similarity component is assigned an extremely low weight in the evaluation of overall similarity. Additionally, the emergent removal of local average terminals makes it possible to visually analyze the behavior of the solutions since their formulation is now dependent only on three variables: σ_X, σ_Y,



(a) MOS on TID2008 (b) SSIM-like (only σ) (c) GP-SSIM T_{M_S}-A (d) GP-SSIM T_{M_S}-B

Fig. 4. Visualization of MOS data distribution for middle-level variables, and corresponding responses by middle-level symmetric FR-IQAMs.

and σ_{XY}. Figure 4(a) presents the actual data distribution and scores of the TID2008 dataset, relying solely on the variance variables. Each point identifies an image-distortion pair, for which variance statistics are extracted and then aggregated, following Equation 9. Color is used to encode the associated MOS value, from low blue to high yellow. The bottom part of the plot shows horizontal slices of the distribution, extracted at nine different values of σ_{XY} as indicated in the top part of the plot. Adopting the same visualization syntax, Figure 4(b) shows the application of an SSIM-like expression that excludes the luminance component for similarity assessment, and it is provided as a comparison reference.

Figure 4(c) shows GP-SSIM T_{M_S}-A: a compact yet high-performing individual based on middle-level symmetric terminals:

$$\text{GP-SSIM } T_{M_S} - \text{A} = \frac{{\sigma_y}^2+{\sigma_x}^2}{\sigma_{xy}} - \left(\frac{\left({\sigma_y}^2 - \sigma_x\sigma_y + {\sigma_x}^2\right)^{0.05} + \sigma_{xy}}{{\sigma_y}^2 - \sigma_x\sigma_y + {\sigma_x}^2}\right)^{11.11} \quad (10)$$

Figure 4(d) presents GP-SSIM T_{M_S}-B, another solution based on middle-level symmetric terminals that can be visualized thanks to its reliance on only three components:

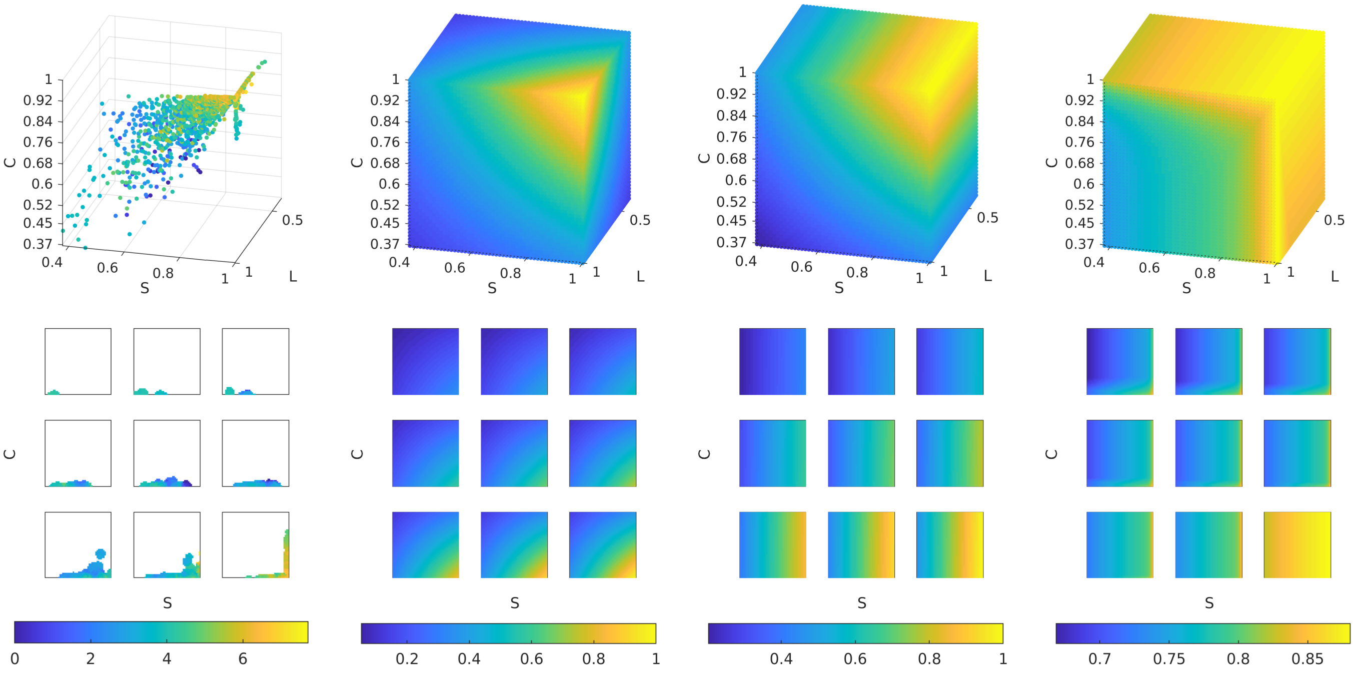
$$\text{GP-SSIM } T_{M_S} - \text{B} = \frac{\left({\sigma_y}^2 + {\sigma_x}^2\right)^{1.25}}{\left|\sigma_{xy}\right|^{2.5}} - \frac{{\sigma_y}^2 + {\sigma_x}^2}{\sigma_{xy}} \quad (11)$$

In general, one can observe that the original data distribution is extremely localized, within the feature space defined by σ_X, σ_Y, and σ_{XY}: large portions of the space are unaccounted for, and the gradient from low similarity to high similarity (as expressed with MOS) takes place in a very compact range. Despite the sparsity of the data, each slice clearly shows a higher MOS when both σ_X and σ_Y are smaller. All three visualized solutions for FR-IQA present a gradient that follows the same direction as the MOS data, although differently distributed. The gradient intensity has little effect on the rank correlation between metric and MOS, since the ordinal relationship among various reference-distortion pairs is preserved in either case. Linear correlation, however, can also be a desirable property, as it provides a consistent perception of "similarity differences": to this extent, the individual visualized in Figure 4(c) provides a better fit with the MOS distribution, which is reflected in generally higher PLCC.

Similarly, it is possible to visually inspect the behavior of high-level terminal solutions when constrained to a single scale, as is the case for the original SSIM formulation. Figure 5(a) shows the training distribution for luminance (L) contrast (C) and structure (S) similarity on the TID2008 dataset, highlighting once again the sparse nature of the training data when considering these SSIM-related features. Figure 5(b) presents the behavior of the original SSIM formulation, which combines the three individual similarities through a multiplication. Figure 5(c) shows the effects of reducing the impact of the luminance component, as suggested in the optimization performed by Bakurov et al. [3], [4].

Figure 5 (d) shows the behavior of GP-SSIM $T_{H_{SS}}$-B, the best high-level solution at single scale, which exploits the {MEAN, MIN, MAX} function set:

$$\begin{aligned}\text{GP-SSIM } T_{H_{SS}} - \text{B} \\ &= s^{0.6} + 0.5\left(\max\left(c^{1.25}, l^{39.0625}, \min\left(l^{6.67}, 0.5\,(s+l)\right)\right)\right) \\ &\quad + \max\left(c, 0.0098\left(l^{40.0} + c^{0.05}\right)^{6.67}, s^{0.8}, s,\right. \\ &\quad \left.\left(0.5\left(0.5\,(ls+l) + l^{6.67}\right)\right)^{66.6}\right) \quad (12)\end{aligned}$$



(a) MOS on TID2008 (b) SSIM-like (c) OP-SSIM [4] (d) GP-SSIM $T_{H_{SS}}$-B

Fig. 5. Visualization of MOS data distribution for high-level variables, and corresponding responses by high-level FR-IQAMs. The color coding goes from less-similar (blue) to most-similar (yellow). (a) MOS data distribution from TID2008. (b) SSIM-like behavior. (c) Behavior of the solution from Bakurov et al. [4] that was found to mainly ignore luminance. (d) Best cross-dataset high-level single-scale solution behavior with {MEAN, MIN, MAX} functions (GP-SSIM $T_{H_{SS}}$-B).

Here, l, c and s respectively refer to the luminance, contrast and structure components between the reference and distortion image, as defined in Equations 1, 2, 3. The general behavior follows the concept of reducing the influence of luminance in the computation of the overall similarity. This is manifest in the tendency of having high similarity values (depicted in yellow) regardless of the value of the luminance component. In addition, the genetic programming approach allows the generation of a complex distribution, that better matches the training data distribution, while still generalizing on a wide range of datasets, as shown numerically in Table III and Table IV.

Figure 6 presents a visual tree-based representation of our best overall individual GP-SSIM $T_{H_{MS}}$-B (based on extended function set with {MEAN, MAX, MIN}). In this case, a component-based visualization such as those presented for other individuals is not possible, as the solution depends on more than three variables. We refer the reader to the supplementary material for the full equation of this individual (as well as all individuals selected for Table IV) and for a detailed view of its performance per distortion type. Figure 7 illustrates visually the different behavior of MS-SSIM and GP-SSIM $T_{H_{MS}}$-B. We compute the similarity according to the two analyzed functions for all reference-distortion image pairs in the TID2008 dataset. We obtain three possible rankings of all pairs, respectively based on users MOS, MS-SSIM, and our GP-SSIM $T_{H_{MS}}$-B. This procedure allows us to visually compare similarity metrics that produce outputs in different ranges, focusing only on the relative ordering that they produce and mirroring the use of the Spearman Rank correlation coefficient in standard quantitative evaluation. For selected image pairs of different distortion types and levels, we report the three similarity ranks. The first three columns represent the case where the expected similarity is very low, as shown by MOS rank, which is well-matched by our rank, and less well-captured by standard MS-SSIM. The fourth and fifth columns show examples of our rank failing to capture the extremely low image similarity expressed by MOS, with the Gaussian

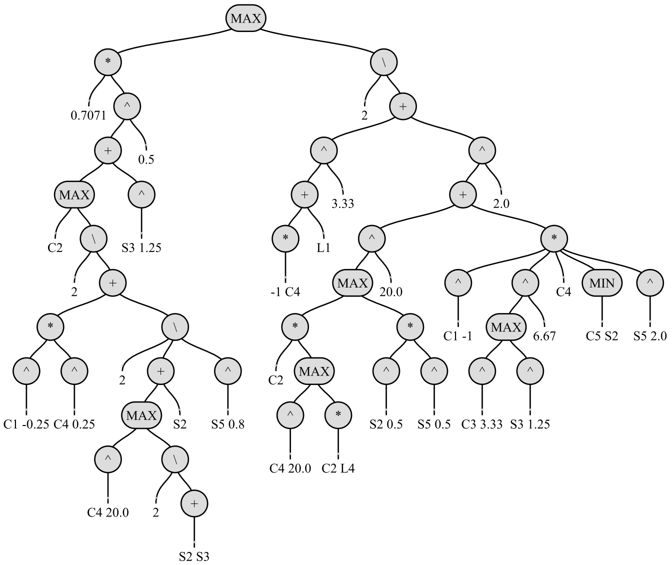


Fig. 6. Tree-based representation of our best overall individual GP-SSIM $T_{H_{MS}}$-B based on extended function set with {MEAN, MAX, MIN}.

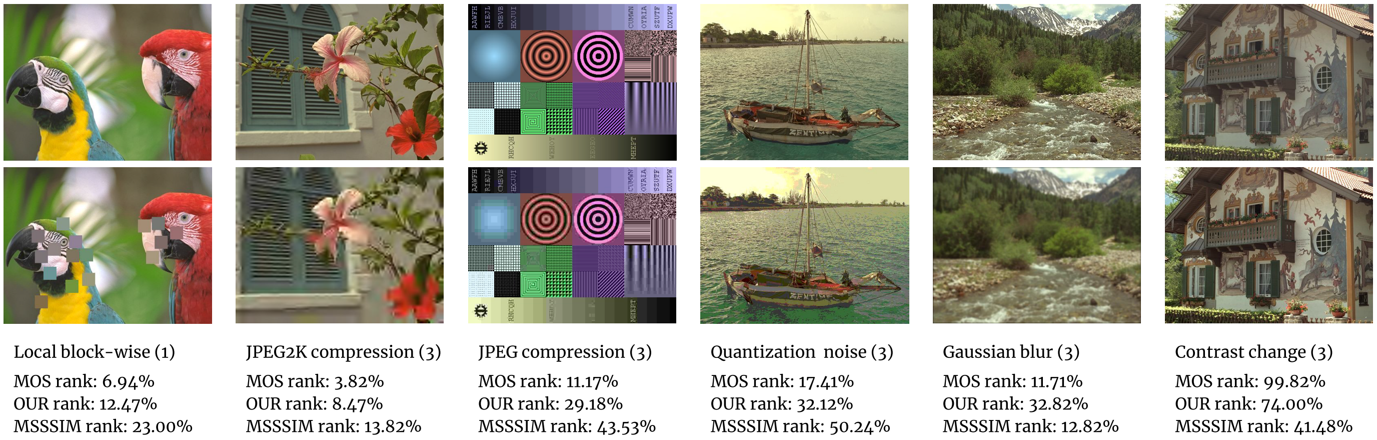


Fig. 7. Visual examples of similarity rank between reference and distorted images. Higher similarity ranks correspond to higher predicted quailty. Reported values include human mean opinion scores (MOS), our multiscale similarity (OUR), and multiscale structural similarity (MSSSIM).

blur, in particular, being better handled by MS-SSIM. Finally, the last column shows a case of strong contrast change that is disregarded by most users (high MOS) and similarly ignored by our solution.

As noted by Bakurov et al. [36], the original SSIM can be subject to a probabilistic interpretation, where the "probability" of two images being similar is computed as the probability that they are simultaneously similar according to luminance, contrast, and structure (their values being in the range between 0 and 1), combined via the multiplication rule. In this paper we obtain high-level individuals (T_H) that exploit both multiplications and additions, which can be interpreted as a complex rule set for the integration of both mutually independent and dependent sub-similarity probabilities. The underlying components of these probabilities, i.e. the high-level terminals, are manipulated through the use of exponents in order to calibrate their impact on the overall similarity: a high exponent will lead to a values closer to 0, which in turn will make the term highly-impactful when combined with others via multiplication, or negligible when combined via addition. To this extent, the high-valued exponents shown in some individuals might lead to such small values that they might lead to a practical simplification of the expressions, equating some terms to zero: we reserve this study, as well as the potential for numerical representation issues, for future works.

C. Comparison With State of the Art

We extend the evaluation of FR-IQAMs generated from our Genetic Programming approach by including a comparison with existing solutions from the state-of-the-art, at varying levels of complexity, introduced in Sections II-B and II-C. The results are presented in Table V for the same datasets.

The first rows in the table represent two different views of our approach. The very first one, "GP-SSIM (ours)", corresponds to GP-SSIM $T_{H_{MS}}$-B: our best cross-dataset individual, optimized on the KADID10K dataset. The second row, "(GP-SSIM per dataset)", reports the best performance on each dataset, extracted from different individuals also optimized on KADID10K. This is provided as a reference, as it is not intended for direct comparison, and it suggests that the training set is adequately large and heterogeneous so as to guarantee per-dataset performance similar to what can be obtained with a selected individual.

TABLE V

SRCC and PLCC Statistics of Our Best Overall Individual GP-SSIM $T_{H_{MS}}$-B, Compared to State of the Art Metrics for FR-IQA

IQAM	TID2008	TID2013	LIVE2	SRCC LIVEM	VDID2014	CSIQ	PieAPP
PSNR	0.525	0.687	0.873	0.626	0.861	0.810	0.268
RMSE	0.525	0.687	0.873	0.626	0.861	0.810	0.268
VIF (2006) [13]	0.655	0.677	0.964	0.836	0.908	0.911	0.181
MAD (2010) [17]	0.834	0.781	0.967	0.865	0.925	0.947	0.266
IW-SSIM (2010) [16]	0.856	0.778	0.957	**0.884**	0.918	0.921	0.298
RFSIM (2010) [15]	0.863	0.774	0.900	0.833	0.914	0.929	0.190
FSIMc (2011) [18]	0.884	0.851	0.965	0.867	0.926	0.931	0.378
GSM (2011) [19]	0.855	0.795	0.955	0.845	0.919	0.913	0.357
SFF (2013) [21]	0.877	0.851	0.965	0.870	**0.931**	**0.963**	0.273
DOG-SSIM (2015) [22]	**0.926**	**0.907**	0.961	-	-	0.952	0.464
SC-QI (2016) [24]	0.905	0.905	0.948	-	-	0.943	0.360
NLPD (2016) [23]	-	0.800	0.937	-	-	0.932	-
GMSD (2013) [20]	0.891	0.804	0.960	0.845	0.927	0.950	0.297
DeepSim (2017) [5]	0.887	0.846	**0.974**	0.877	0.921	0.919	0.500
DeepIQA (2017) [6]	0.908	0.831	0.947	0.794	0.920	0.909	0.537
PieAPP (2018) [25]	0.788	0.876	0.919	0.769	0.895	0.892	**0.831**
LPIPS (2018) [26]	0.731	0.670	0.932	0.849	0.886	0.876	0.492
DISTS (2020) [27]	0.773	0.830	0.954	0.866	0.917	0.929	0.693
GP-SSIM (ours)	0.878	0.794	0.913	0.860	0.927	0.953	0.263
(GP-SSIM per dataset)	(0.892)	(0.808)	(0.918)	(0.861)	(0.928)	(0.953)	(0.389)
IQAM	TID2008	TID2013	LIVE2	PLCC LIVEM	VDID2014	CSIQ	PieAPP
PSNR	0.489	0.677	0.865	0.684	0.655	0.819	0.288
RMSE	0.406	0.597	0.538	0.660	0.704	0.697	0.170
VIF (2006) [13]	0.638	0.771	0.960	0.868	0.872	0.913	0.192
MAD (2010) [17]	0.829	0.827	**0.968**	**0.894**	0.925	0.950	0.198
IW-SSIM (2010) [16]	0.858	0.764	0.952	0.847	0.828	0.914	0.023
RFSIM (2010) [15]	0.862	0.812	0.895	0.866	0.907	0.913	0.255
FSIMc (2011) [18]	0.834	0.877	0.961	0.822	0.847	0.919	0.481
GSM (2011) [19]	0.846	0.797	0.944	0.750	0.808	0.898	0.336
SFF (2013) [21]	0.882	0.809	0.963	0.874	0.921	**0.964**	0.036
DOG-SSIM (2015) [22]	**0.928**	**0.919**	0.963	-	-	0.942	0.417
SC-QI (2016) [24]	0.890	0.907	0.937	-	-	0.927	0.267
NLPD (2016) [23]	-	0.839	0.932	-	-	0.923	-
GMSD (2013) [20]	0.872	0.855	0.957	0.863	0.887	0.945	0.242
DeepSim (2017) [5]	0.876	0.872	**0.968**	0.885	0.885	0.919	0.516
DeepIQA (2017) [6]	0.917	0.834	0.940	0.834	**0.928**	0.901	0.568
PieAPP (2018) [25]	0.610	0.859	0.908	0.803	0.816	0.877	**0.842**
LPIPS (2018) [26]	0.772	0.749	0.934	0.844	0.890	0.896	0.503
DISTS (2020) [27]	0.800	0.855	0.954	0.879	0.918	0.928	0.716
GP-SSIM (ours)	0.824	0.814	0.899	0.740	0.895	0.918	0.295
(GP-SSIM per dataset)	(0.859)	(0.846)	(0.908)	(0.888)	(0.921)	(0.928)	(0.419)

Best results in bold, and second best underlined.

The original SSIM owes its success in the scientific community to the mathematical simplicity, low computational complexity, and implicit incorporation of characteristics from the human visual system; it motivated several related measures which, however, tend to add further levels of complexity in order to better estimate the quality scores. For example, the

DOG-SSIM decomposes a given reference-distortion pair of images into five levels of frequency bands to further estimate their similarity score. Similarly, to estimate a quality score using FSIM, two complementary feature maps have to be constructed first: the phase congruency and the gradient magnitude. Alike, FSIMc transforms the RGB color space to the YIQ color space and then defines three chromatic feature maps. Contrarily to other SSIM-inspired extensions, our approach relies upon the same building blocks as the original SSIM (in both single and multi-scale variants), with the difference that they are averaged on the spatial dimensions before computing the similarity measure, and are combined into a different mathematical formulation. From Table IV we can conclude that our approach improves upon the standard SS-SSIM and MS-SSIM for most of the image quality assessment databases. Although the same does not happen when comparing our method with the state of the art in Table V, the solutions we propose are mathematically less complex; this is particularly relevant when comparing with state of the art IQA measures which rely upon even more complex feature extraction methods such as the deep convolutional neural networks (DeepSim, DeepIQA, PieAPP, LPIPS and DISTS). Excellent performance levels are also reported by lower-level methods, such as SFF and DOG-SSIM. Where SFF is trained on a limited set of nine custom images, DOG-SSIM is instead evaluated in a cross-validation setup within each dataset, without information about the type of split (cross-distortion or cross-reference). To this extent, an important point of discussion emerges when dealing with data-driven methods including, but not limited to, those related to image quality assessment. In this context, the training dataset used for optimization can have a significant impact on the resulting metric, which will inherit the biases related to the selection of degradations, cultural extraction of the human judges, experimental setup, and other factors. This suggests as a potential direction for future works, an investigation of the variation in performance when optimizing for different sources of data.

VI. Conclusion

We have presented an optimization process to define full-reference image quality measures based on genetic programming. Our choice of terminal sets, used to construct the evolving programs, has been guided by the objective of formulating SSIM-like measures due to their efficacy and effectiveness. For our optimized solutions, we have analyzed the general behavior in terms of elite statistics as well as the specific behavior of selected individuals. In particular, we compared them against different variations of SSIM on a challenging cross-dataset optimization setup. Additionally, dataset-specific optimization has proven the possibility to reach a correlation with human mean opinion scores that is competitive against existing complex solutions from image quality assessment.

With this work we have proven that it is possible to "re-learn" a formulation for image quality assessment which can be considered roughly equivalent, in terms of complexity and performance, to a metric that was developed by experts and refined through years of iterations in research. Furthermore, we have shown that, given sufficient resources and complexity budget, the same optimization framework can lead to formulations that are superior in performance, although at the cost of reducing their interpretability. Nonetheless, the computational steps of our best-performing solution are significantly fewer than what is required by state of the art deep learning methods, opening the possibility of a dedicated study on their explainability and interpretation.

Thanks to our choice of terminal and function sets, our optimized image quality measures are by construction differentiable. This characteristic opens the doors to potentially employ such measures as loss functions for gradient-based backpropagation in neural networks training. For this reason, future research will be focused on further analyzing our solutions in terms of their ability to facilitate gradient propagation. As an additional direction for future work, we will consider exploring different terminal sets, possibly inspired by other existing full-reference image quality measures, to investigate the possibility of merging different existing measures. Finally, it is important to acknowledge the relevance of the video domain in nowadays applications of image quality. Effective algorithms for video compression, for example, apply a controlled reduction in frame quality based on an analysis of its temporal neighborhood. As such, a method for quality assessment in video sequences, such as Video Multi-Method Assessment Fusion (VMAF) [53], should also be specifically designed for the analysis of multiple frames at once, in order to properly address cross-frame artifacts and to correlate with the mechanisms of human perception of moving images. We consider this a fundamental direction for future explorations of our research.

References

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[2] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Syst.*, vol. 22, no. 2, pp. 213–227, Mar. 2016.

[3] I. Bakurov, M. Buzzelli, M. Castelli, R. Schettini, and L. Vanneschi, "Parameters optimization of the structural similarity index," in *London Imaging Meeting*, vol. 2020, no. 1. Springfield, VA, USA: Society for Imaging Science and Technology, 2020, pp. 19–23.

[4] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (SSIM) revisited: A data-driven approach," *Expert Syst. Appl.*, vol. 189, Mar. 2022, Art. no. 116087. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421014238

[5] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "DeepSim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, Sep. 2017.

[6] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[7] *Research Anthology on Multi-Industry Uses of Genetic Programming and Algorithms*, Information Resources Management Association, Engineering Science Reference, IGI Global, Hershey, PA, USA, 2021.

[8] (2003). *The SSIM Index for Image Quality Assessment*. Accessed: Sep. 16, 2022. [Online]. Available: https://www.cns.nyu.edu/~lcv/ssim/

[9] E. Peli, "Contrast in complex images," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 7, no. 10, pp. 2032–2040, 1990.

[10] Z. Wang, E. P. Simoncelli, and A. C. Bovil, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Conf. Signals Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.

[11] A. K. Venkataramanan, C. Wu, A. C. Bovik, I. Katsavounidis, and Z. Shahid, "A hitchhiker's guide to structural similarity," *IEEE Access*, vol. 9, pp. 28872–28896, 2021.

[12] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[13] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

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

[15] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 321–324.

[16] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[17] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.

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

[19] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[20] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[21] H.-W. Chang, H. Yang, Y. Gan, and M.-H. Wang, "Sparse feature fidelity for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4007–4018, Oct. 2013.

[22] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual DOG model fused with random forest," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3282–3292, Nov. 2015.

[23] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized Laplacian pyramid," *Electron. Imag.*, vol. 28, no. 16, pp. 1–6, Feb. 2016.

[24] S.-H. Bae and M. Kim, "A novel image quality assessment with globally and locally consilient visual quality perception," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2392–2406, May 2016.

[25] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1808–1817.

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[27] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, pp. 2567–2581, 2022.

[28] V. V. Lukin, N. N. Ponomarenko, O. I. Ieremeiev, K. O. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," *Proc. SPIE*, vol. 9394, pp. 172–183, Mar. 2015.

[29] J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.

[30] A. Khan, A. S. Qureshi, N. Wahab, M. Hussain, and M. Y. Hamza, "A recent survey on the applications of genetic programming in image processing," *Comput. Intell.*, vol. 37, no. 4, pp. 1745–1778, Nov. 2021.

[31] R. Joshi and S. Tapaswi, "Image similarity: A genetic algorithm based approach," *World Acad. Sci., Eng. Technol.*, vol. 27, pp. 327–331, Mar. 2007.

[32] B. Syam and Y. Rao, "An effective similarity measure via genetic algorithm for content based image retrieval with extensive features," *Int. Arab J. Inf. Technol.*, vol. 10, no. 2, pp. 143–151, 2013.

[33] R. da S. Torres et al., "A genetic programming framework for content-based image retrieval," *Pattern Recognit.*, vol. 42, no. 2, pp. 283–292, Feb. 2009.

[34] C. D. Ferreira, J. A. Santos, R. da S. Torres, M. A. Gonçalves, R. C. Rezende, and W. Fan, "Relevance feedback based on genetic programming for image retrieval," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 27–37, Jan. 2011.

[35] R. T. Calumby, R. da Silva Torres, and M. A. Gonçalves, "Multimodal retrieval with relevance feedback based on genetic programming," *Multimedia Tools Appl.*, vol. 69, no. 3, pp. 991–1019, Apr. 2014.

[36] I. Bakurov, M. Buzzelli, M. Castelli, R. Schettini, and L. Vanneschi, "Genetic programming for structural similarity design at multiple spatial scales," in *Proc. Genetic Evol. Comput. Conf.*, Jul. 2022, pp. 911–919.

[37] L. Vanneschi, I. Bakurov, and M. Castelli, "An initialization technique for geometric semantic GP based on demes evolution and despeciation," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2017, pp. 113–120.

[38] I. Bakurov, L. Vanneschi, M. Castelli, and F. Fontanella, "EDDA-V2—An improvement of the evolutionary demes despeciation algorithm," in *Parallel Problem Solving From Nature PPSN XV*, A. Auger, C. M. Fonseca, N. Lourenço, P. Machado, L. Paquete, and D. Whitley, Eds. Cham, Switzerland: Springer, 2018, pp. 185–196.

[39] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.

[40] N. Ponomarenko et al., "TID2008-A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectronics*, vol. 10, pp. 30–45, Jan. 2009.

[41] N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, and V. Lukin, "Metrics performance comparison for color image database," in *Proc. 4th Int. Workshop Video Process. Quality Metrics*, 2009, pp. 1–6.

[42] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[43] 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, no. 11, pp. 3440–3451, Nov. 2006.

[44] (2006). *Live Image Quality Assessment Database Release 2*. Accessed: Oct. 4, 2019. [Online]. Available: http://live.ece.utexas.edu/research/quality

[45] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Conf. Rec. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012, pp. 1693–1697.

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

[47] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.

[48] *CSIQ Image Quality Database*. Accessed: Oct. 6, 2019. [Online]. Available: http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=23

[49] R. Franzen. (1999). *Kodak Lossless True Color Image Suite*. Accessed: Oct. 4, 2019. [Online]. Available: http://r0k.us/graphics/kodak/

[50] C. Rota and M. Buzzelli, "MdVRNet: Deep video restoration under multiple distortions," in *Proc. 17th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2022, pp. 419–426.

[51] S. Bianco, M. Buzzelli, G. Ciocca, and R. Schettini, "Neural architecture search for image saliency fusion," *Inf. Fusion*, vol. 57, pp. 89–101, May 2020.

[52] A. K. Venkataramanan. (2020). *GitHub Utlive/Enhanced_SSIM: A C Implementation of Enhanced SSIM Based on the Recommendations Made in 'A Hitchhiker's Guide to SSIM'*. Accessed: Dec. 19, 2022. [Online]. Available: https://github.com/utlive/enhanced_ssim

[53] Netflix. (2016). *GitHub Netflix/VMAF: Perceptual Video Quality Assessment Based on Multi-Method Fusion*. Accessed: Oct. 4, 2022. [Online]. Available: https://github.com/Netflix/vmaf