Expert Systems With Applications 189 (2022) 116087

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Structural similarity index (SSIM) revisited: A data-driven approach

Illya Bakurov [a], Marco Buzzelli [b], Raimondo Schettini [b], Mauro Castelli [a,*], Leonardo Vanneschi [a]

[a] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312, Lisboa, Portugal
[b] Department of Informatics Systems and Communication, University of Milano – Bicocca, Viale Sarca, 336, 20126 Milano, Italy

ARTICLE INFO

Keywords:
Image quality assessment measures
Structural similarity
Evolutionary computation
Scale selection
Image processing

ABSTRACT

Several contemporaneous image processing and computer vision systems rely upon the full-reference image quality assessment (IQA) measures. The single-scale structural similarity index (SS-SSIM) is one of the most popular measures, and it owes its success to the mathematical simplicity, low computational complexity, and implicit incorporation of Human Visual System's (HVS) characteristics. In this paper, we revise the original parameters of SSIM and its multi-scale counterpart (MS-SSIM) to increase their correlation with subjective evaluation. More specifically, we exploit the evolutionary computation and the swarm intelligence methods on five popular IQA databases, two of which are dedicated distance-changed databases, to determine the best combination of parameters efficiently. Simultaneously, we explore the effect of different scale selection approaches in the context of SS-SSIM. The experimental results show that with a proper fine-tuning (1) the performance of SS-SSIM and MS-SSIM can be improved, in average terms, by 8% and by 3%, respectively, (2) the SS-SSIM after the so-called standard scale selection achieves similar performance as if applying computationally more expensive state-of-the-art scale selection methods or MS-SSIM; moreover, (3) there is evidence that the parameters learned on a given database can be successfully *transferred* to other (previously unseen) databases; finally, (4) we propose a new set of reference parameters for SSIM's variants and provide their interpretation.

1. Introduction

An English adage says "a picture is worth a thousand words", and the fact that people upload about 350 million new photos each day on Facebook alone (Smith, 2013) is a strong argument in its favor. In fact, digital imagery became a new way of communication and, due to technological progress, taking a photo or recording a video and then sharing it to a virtual community of millions is now a matter of few clicks. Nevertheless, various aspects play a significant role in the degradation of the visual quality of a digital image or video; these can be raindrops adhered to a window or camera lens, a low-light environment, an accidental shake of the capturing device (or the target object) during the acquisition process, an improper definition of the camera's parameters, etc. As if these factors were not enough, in the background, there is a whole set of technological steps that occur before the final users can make use of the digital image acquisition process' outcome; typically, these are organized into a pipeline of sequential steps, like digitization, compression, storage, transmission and reproduction, and may result in a noticeable visual degradation of the final rendering. From the perspective of media quality perception, this might result in an annoying viewing experience; whereas, for instance, from the perspective of a radiology practice, this might result in visual artifacts that can make it harder for a radiologist or an automated system to detect a disease on an MRI. The scientific community has developed numerous computational systems to remove the undesirable visual artifacts that degrade images' visual quality and reduce their usefulness for the underlying tasks. Despite the heterogeneity of contexts that motivated their creation and the discrepancy of methodological approaches, these systems have at least one element in common — their assessment is mainly based on the structural similarity index (SSIM), a popular full-reference image quality assessment (FR-IQA) measure. Before presenting the reader with the core topic of our research, we provide a smooth immersion into the wold of IQA. We start by introducing IQA measures and their classification, then we make a short overview of the research track, highlight some of the advancements and problems relevant to the context of our study; finally, we provide the reader with some examples which, in our opinion, illustrate the importance of our research in both.

The IQA measures are commonly divided in two categories: subjective and objective. For the applications in which humans ultimately view the images, the most appropriate method for quantifying image

* Corresponding author.
E-mail addresses: ibakurov@novaims.unl.pt (I. Bakurov), marco.buzzelli@unimib.it (M. Buzzelli), raimondo.schettini@unimib.it (R. Schettini), mcastelli@novaims.unl.pt (M. Castelli), lvanneschi@novaims.unl.pt (L. Vanneschi).

https://doi.org/10.1016/j.eswa.2021.116087
Received 27 August 2020; Received in revised form 2 August 2021; Accepted 12 October 2021
Available online 27 October 2021

quality is through the subjective evaluation of the human visual system (HVS), i.e., by involving people to assess image quality (Wang et al., 2004). However, in practice, subjective evaluation becomes complex, time-consuming, expensive and highly sensible to the experimental design (Streijl et al., 2016). Considering the aforementioned limitations, several researchers have proposed objective measures that can automatically, i.e., without human involvement, estimate the perceived visual quality. The objective IQA measures can be classified based on the availability of a pristine reference image. When a measure compares a reference image and its potentially corrupted variant, it is classified as *full-reference* IQA (FR-IQA). When the reference image is not available, the measure is classified as *no-reference* IQA (NR-IQA). When the reference image is not entirely available, i.e., only some partial information is provided, like a given set of the extracted features, the measure is classified as *reduced-reference* IQA (RR-IQA). In this paper, we focus our attention on the FR-IQA. The following paragraphs are intended to provide the reader with an introduction to the FR-IQA measures and highlight some of the advancements and problems relevant to the context of our study.

Until (approximately) the last fifteen years, the assessment of signal quality and fidelity, which also includes digital images, was mainly performed through the mean squared error (MSE) and its derivations. In the context of FR-IQA, to determine the degree of distortion, this class of measures relies upon the global amount of pixel errors obtained from reference–distortion pairs. The major advantages of this class of measures are the mathematical simplicity, the differentiability, i.e., they can be exploited as a guiding loss for gradient-based learning of Artificial Neural Networks (ANNs), and the low computational complexity. However, this class of measures does not happen to correlate well with humans' perception of visual quality; several experiments show the low sensitivity of MSE-based approaches when considering several levels of different distortion types, even though the images present significantly different visual quality (Wang & Bovik, 2009). Several researchers have proposed other FR-IQA measures that explicitly incorporate characteristics of the HVS to overcome these limitations.

The structural similarity index (SSIM) (Wang et al., 2004) is an FR-IQA measure inspired by the theory that HVS is highly adapted for extracting structural information from the scenes. The incorporation of this characteristic as the intrinsic component of an IQA measure allowed authors to outperform not only the MSE-based measures but also the existing state-of-the-art perceptual image quality measures, showing a better correlation with the subjective evaluation provided by the human observers, such as the mean opinion score (MOS) and differential MOS (DMOS), on different IQA databases. The increased performance, simple mathematical formulation, differentiability, and high degree of computational parallelization allowed SSIM to become one of the most popular FR-IQA measures in the scientific community, and it has been used as a proxy evaluation for human assessment in different image processing (IP) and computer vision (CV) applications. The following paragraph presents several illustrative examples of SSIM's utilization in a varied set of application fields.

In Qian et al. (2018), SSIM was used to quantitatively compare the state-of-the-art methods with a novel attentive generative adversarial network (GAN) to remove raindrops from a single image. The work of Toizumi et al. (2019) utilized SSIM in the context of satellite imagery to assess a novel framework to train an artifact-free thin cloud removal model using GAN with thick cloud masks, and compare it against conventional methods. The research of Rundo et al. (2019) employed SSIM in the context of medical imaging applied to different clinical scenarios, like uterine fibroid segmentation in MR-guided focused ultrasound surgery and brain metastatic cancer segmentation in neuroradiosurgery. Specifically, SSIM is used to compare the conventional state-of-the-art image enhancement techniques with a novel framework for image enhancement, based on genetic algorithms, to improve the appearance and the visual quality of images characterized by a bimodal gray-level intensity histogram. The work of Zhao et al. (2019) relied on SSIM to compare state-of-the-art techniques for image denoising with a novel convolutional neural network (CNN), called subband denoising CNN (SD-CNN), that incorporates frequency information with spatial context and, therefore, recovers image details more effectively. The research of Zini et al. (2020) employed SSIM to compare state state-of-the-art-models for JPEG restoration with a novel deep residual autoencoder, which effectively restores images with any level of compression that leverages both the learning capacity of deep residual networks and prior knowledge of the JPEG compression pipeline. In Bianco et al. (2021), SSIM was applied to assess a novel approach to analyze the capability of deep visual representations to characterize different types of image distortions intrinsically. To demonstrate the usefulness of their approach, authors experiment on different image quality assessment tasks. In Bianco et al. (2020), the authors used SSIM to assess a novel method for single image dehazing that exploits a physical model to recover the haze-free image by estimating the atmospheric scattering parameters. Similarly, in Wang et al. (2020), the authors used SSIM to assess a novel weakly supervised network based on the multi-level multi-scale block and quantitatively compared the proposed approach with state-of-the-art methods for single image dehazing. Furthermore, in Shao et al. (2020), the authors proposed a domain adaptation framework for single image dehazing, which includes two parts: an image translation module and two domain-related dehazing modules. To quantitatively assess and compare their method with state-of-the-art approaches for image dehazing, the authors used SSIM.

Since its introduction, SSIM was mainly used at a single spatial-scale (SS-SSIM). However, in practice, subjective evaluation is highly dependent on the numerous viewing conditions. These include, among many others, the conditions of the displaying device, such as display resolution and response time, viewing distances, network bandwidth and latency, etc. In this context, the single-scale approach may only be appropriate for specific settings. Notably, it was shown that SSIM, as with several other IQA measures, is highly sensitive to spatial scale selection (Gu et al., 2015; Venkataramanan et al., 2021). The problem of determining a proper spatial scale originated two conceptually different research tracks and, consequently, solutions. The first consisted of estimating the overall similarity by aggregating the inner similarity indexes obtained from a range of reasonable spatial scales; this approach has originated the multi-scale SSIM (MS-SSIM), which has inspired several other metrics like Visual Information Fidelity (VIF) (Sheikh & Bovik, 2006; Wang et al., 2003). Instead of considering a range of scales, the second research track focused on determining the most appropriate spatial scale before computing the single-scale similarity between reference–distortion pairs (Gu et al., 2015, 2013a, 2013b; Lin & Kuo, 2011). Both proved to be effective approaches in accommodating the diversity of viewing conditions.

In this work, we take a data-driven approach to revise the default parameters of a popular FR-IQA measure – SSIM – under the light of the above-mentioned discussion regarding scale selection. Specifically, we apply the genetic algorithm (GA) and particle swarm optimization (PSO), in a comparative fashion, to find the best combination of parameters for the single-scale SSIM (SS-SSIM), considering different scale selection approaches, and its multi-scale extension (MS-SSIM). Although, both GA and PSO can be used interchangeably in many applications, these are conceptually different population-based stochastic metaheuristics (i.e., with different inspirations and mechanics). The paragraph below briefly presents both modi operandi and discusses their main advantages and limits.

GA and PSO belong to a larger class of evolutionary computation and swarm intelligence metaheuristics. They are generally applied to solve optimization problems that are computationally hard (as we show in this paper, the problem of optimizing SSIM's variants can be classified as NP-hard). They are known for their capability to produce fit solutions, which are either optimal or sub-optimal, in a reasonable amount of time. GA is a meta-heuristic introduced by Holland (1992),

which was strongly inspired by Darwin's theory of evolution. The algorithm starts with a random-like population of candidate-solutions (aka chromosomes). Then, by mimicking the natural selection and genetically-inspired variation operators, such as the crossover and the mutation, the algorithm breeds an offsprings' population that replaces the previous population (a.k.a. the parent population). This procedure is iterated until reaching some stopping criteria. PSO is another form of population-based stochastic metaheuristics introduced by Kennedy and Eberhart (1995), and, contrarily to GAs, it was inspired by the social behavior of living organisms, such as the birds and fishes when looking for some food source. Following PSO's nomenclature, a candidate-solution is called a particle, and a population is called a swarm. Each particle's position in the search space is updated based on a procedure that takes into account the cognitive component of the particle (i.e., particle's personal memory) and the social component of the particle (i.e., its cooperation with the swarm). Since its introduction in 1995, PSO rapidly became popular in the scientific community. Compared to GAs, PSO is known to be more efficient in terms of time-complexity, requires fewer parameters, and is simpler to implement. Several studies point to PSO's superiority over GAs in terms of performance across several domains (Jatana & Suri, 2020; Wihartiko et al., 2018). However, there is also evidence for the opposite (Hamzaoui & Arellano, 2018; Rhodes, 2019). The work of Bakurov et al. (2021) presents a benchmark of a varied set of functions evaluated across several dimensions and shows no clear advantage of one metaheuristic over another. In this context, we decided to employ both GA and PSO comparatively to leverage algorithms' potential in this particular application. Moreover, our motivation relies on top of one of the fundamental theorems in the field of optimization – the no free lunch theorem – which roughly states that the average performance of any pair of algorithms across all possible problems is exactly identical (Wolpert & Macready, 1997). Finally, given that the principal interest of this study is to revise IQA measures' parameters in a data-driven manner, by comparatively applying the GA and PSO, we can bypass an exhaustive hyper-parameters' exploration for each metaheuristic. Instead, their parameters are chosen according to our best understanding and findings from the literature.

The main contributions of this paper can be summarized as follows:

- We present a methodology to enhance the SS-SSIM and MS-SSIM measures by exploring their parameters search space efficiently, exploiting evolutionary computation and Swarm Intelligence.
- We obtain statistically significant improvements on both measures: in average terms, SS-SSIM is improved by 8% apropos the Spearman correlation with human MOS, and MS-SSIM by 3%.
- We provide a set of recommended parameters, which can be used by the scientific community to both train and evaluate various computer vision solutions using a similarity function that is highly correlated with the human response.
- We show that, with proper parameters, SS-SSIM can be as effective as if applying the more computationally-demanding state-of-the-art optimal scale selection (OSS) or using its multi-scale counterpart.
- We conduct an extensive cross-dataset analysis to highlight the actual efficacy of our parametrizations in a novel scenario, and provide distortion-type details to identify their most effective fields of application.
- We experimentally corroborate the hypothesis that the HVS is highly adapted for extracting structural information from the scenes, which is prioritized over contrast and luminance information.

The document is organized as follows. Section 2 provides a definition of SSIM, its multi-scale extension, and the spatial scale selection approaches used in this study. Section 3 enumerates and describes the research contributions that relate the most to ours. In Section 4 we formalize the research objectives and propose a scientific method to achieve them. The Section 5 presents the benchmark environment and the hyper-parameters that were used in our experiments. The Section 6 exhibits the experimental results and provides a detailed discussion of the main findings. Finally, Section 7 summarizes the main contributions of the paper and suggests possible directions for future research.

2. Background

2.1. Single-scale structural similarity

Almost every FR-IQA measure developed upon principles of HVS can be characterized by two steps (from now on, we will refer to IQA measures as *algorithms*). First, algorithms gather local information of the reference–distortion pair of digital signals to obtain local metrics. Second, algorithms combine these metrics into an overall quality assessment (Kuo et al., 2016). Under this perspective, SSIM is not an exception as it separately measures the local brightness (a.k.a. *luminance*), contrast, and structure of both images, and then aggregates all the local assessments to obtain the overall measure. Unlike MSE-based measures, which compare pair's differences pixel-by-pixel of the whole range, SSIM operates on patches obtained from a sliding-window. This technique better resembles the functioning of HVS because our eyes can easily perceive local information differences in a specific area of the two images, instead of the individual differences in pixel value in the whole area. Formally, SSIM performs a comparison between a pristine reference image x and a potentially corrupted version of the same image y based on three independent components extracted at a single spatial scale (resolution): luminance, contrast, and structure. Each image's patch average μ represents the luminance information. Thus the luminance comparison 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 is a small quantity introduced for numerical stability, as are C_2 and C_3 in the following equations for the other components. The three quantities are given as functions of the dynamic range of the pixel values L ($L = 255$ for 8 bits/pixel gray-scale images) and two scalar constants $K_1 \ll 1$ and $K_2 \ll 1$ (traditionally set to 0.01 and 0.03, respectively): $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$, $C_3 = C_2/2$. Contrast is represented through the use of standard deviation σ. Consequently, the contrast-based comparison is:

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

The structure element is represented through the standardization of each image with the corresponding mean and standard deviation. Comparison of the structure can be obtained through the inner product of these signals:

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

where:

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \mu_x\right)\left(y_i - \mu_x\right). \tag{4}$$

Finally, the three components are combined into a unique expression that is weighted with exponents α, β, and γ:

$$SSIM(x,y) = [l(x,y)]^\alpha \cdot [c(x,y)]^\beta \cdot [s(x,y)]^\gamma. \tag{5}$$

Since it was first introduced to the scientific community, SSIM rapidly grew in popularity and propelled a *significant* amount of research in different sub-fields, including the aforementioned recent contributions presented in Section 1.

2.2. Multi-scale structural similarity

It is known that subjective evaluation is highly dependent on the viewing conditions such as the display resolution, the distance from the display to the observer, the environment illumination, the intrinsic capability of the observer's visual system, etc. For example, as the viewing distance to a fixed-sized monitor increases, the viewing angle shrinks gradually and fewer image details can be noticed. By leveraging such a biological phenomena, one could extract positive outcomes. For instance, applying a higher level of image compression to save the bandwidth without deteriorating the perceived visual quality (assuming a fixed monitor size) (Gu et al., 2015). However, in the context of IQA, overlooking viewing conditions when estimating the perceived visual quality might result in significant underperformance of the measures.

In this context, Wang et al. (2003) pointed out that SSIM considered at a single scale (SS-SSIM) would be appropriate only for specific viewing conditions, as such, unable to incorporate their vast diversity. To remedy this drawback, they have proposed a MS-SSIM by estimating the perceived visual quality by aggregating of inner similarity indexes calculated from a range of different spatial-scales (resolutions). Formally MS-SSIM is defined as:

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

Taking the reference–distortion pair as the input (x and y, respectively), the measure computes, at each scale j, the contrast and structural similarities (c_j and s_j, respectively). When changing 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. The exponents α_M, β_j, and γ_j are used to adjust the relative importance of different components. The overall cross-scale evaluation is then given by a weighted product of the aforementioned components extracted at different scales.

2.3. Selection of spatial scale

After being introduced to the scientific community, MS-SSIM has become one of the most *precise* FR-IQA measures. Nevertheless, several researchers investigated how to consider diverse viewing conditions in the context of IQA through a different perspective. Instead of considering a range of admissible scales, researchers have focused on determining the *right* scale before computing the similarity between reference–distortion pairs in a single-scale approach (Gu et al., 2015, 2013a, 2013b; Lin & Kuo, 2011). We have considered two models for scale selection in this work, both described in the following sub-sub-sections.

2.3.1. Standard scale selection

The first is a simple although popular scale selection method that assumes a typical viewing distance (3∼5 times of the image height or width) and estimates the down-scaling factor as a function of the image's height:

$$Z_d = max(1,\ round(H_i/256)), \tag{7}$$

where H_i indicates the image's height. Note that, from now, we will refer to this method as standard scale selection (SSS).

2.3.2. Optimal scale selection

The second is the so-called optimal scale selection (OSS) method (Gu et al., 2015), which makes use of a cascade of adaptive high-frequency clipping (AHC) in the discrete wavelet transform domain and adaptive resolution scaling. More specifically, it is a method designed upon the AHC model (Gu et al., 2013a) and so-called Self-Adaptive Scale Transform (SAST) model (Gu et al., 2013b).

The AHC model, instead of using the spatial domain, focuses on discarding part of image details by adaptive high-frequency clipping in the discrete wavelet transform (DWT) domain, and then synthesizing the AHC model filtered sub-band coefficients back to an image at its original resolution to be used by IQA metrics. Formally the AHC model is characterized by a weighting function applied to all the LH, HL and HH sub-bands of a wavelet transformed image (Gu et al., 2013a):

$$w(i,d,l) = \frac{b \cdot k^{t(L-l)}}{\alpha^{(\frac{d}{d_o})}}, \tag{8}$$

where L indicates the decomposition layer (set to 4) and α, k, t, d_o correspond to the model parameters (set to 10, 10, 2, and 512 respectively)- Notice that the parameter setting follows the original definition in Gu et al. (2013a).

The SAST model was designed to simulate the spatial filtering mechanism of the HVS (Gu et al., 2013b). Its fundamental idea is to estimate the suitable scaling parameter from the original image resolution and the given viewing distance before resizing input images to boost the performance. The OSS model uses a modified SAST model, formally defined as (Gu et al., 2015):

$$Z'_{sast} = Z_{sast}^{(1-\frac{|\gamma-\gamma_o|^{\beta}}{\alpha})}, \tag{9}$$

where α and β control the speed of modification process caused by different aspect ratios (both are selected as 2), γ stands for the aspect ratio of reference–distortion pair whereas γ_o for the aspect ratio human eyes are well suited to (selected as 9:16), which is also called optimal because it the international standard format for digital television; Z_{sast} represents the SAST model, formally given by Gu et al. (2013b):

$$\sqrt{\frac{H_i \cdot W_i}{H_v \cdot W_v}} = \sqrt{\frac{1}{4tan(\frac{\theta_H}{2}) \cdot tan(\frac{\theta_W}{2})} \cdot (\frac{H_i}{D})^2 \cdot \frac{1}{\gamma}}, \tag{10}$$

where H_v and W_v are the visual height and width, θ_H and θ_W separately indicate horizontal and vertical visual angles and, following Gu et al. (2015), assumed to be 40° and 50° respectively. Notice that parameters' selection follows the original definition in Gu et al. (2015, 2013b).

In such a way, the OSS model is given by the modified SAST model, after applying the AHC model. Combining both AHC and SAST, the OSS model better removes indiscernible details caused by the varying viewing conditions in different but complementary domains.

3. Related works

3.1. Optimization of SSIM parameters

Researchers' interest in optimizing different aspects of the SSIM is not novel. To estimate the relative importance of different scales in MS-SSIM, Wang et al. (2003) applied an image synthesis approach. In their experiments, they considered five scales ($M = 5$) and, assuming $\alpha_j = \beta_j = \gamma_j$ to simplify parameters' selection, conducted a quantitative subjective test with ten original 64 × 64 images with different types of content and 12 distortion levels. The test involved eight subjects, and each was shown the ten sets of test images, one set at a time. The viewing distance was fixed to 32 pixels per degree of visual angle. The subject was asked to compare the quality of the images across scales and detect one image from each of the five scales. At the end, the authors obtained the following set of parameters: $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2364$, $\beta_5 = \gamma_5 = 0.1333$ respectively (Wang et al., 2003). In our opinion, the quantitative subjective test that was conducted to estimate the relative importance of scales is only appropriate for the previously described viewing conditions. That is to say, we speculate that the potential of MS-SSIM might be unexplored due to a limited experimental setup that was used to deduce the relative importance of different scales. Moreover, the authors did

not provide a clear justification for why they considered using five scales.

Silvestre-Blanes (2011) optimized the value of regularization constants K_1 and K_2 through parameters enumeration. They observed that changing these values from the default setting would introduce variations in the output similarity measure up to 0.5 points and affect the ranking of distorted images similarity, thus proving the measure's sensitivity with respect to such parameters. However, the optimization was conducted in a limited search space, evaluating a total of only six parameter combinations, which prompted us to explore this idea more in-depth by resorting to continuous optimization and extending the evaluation to other parameters as well. In addition to the parameters' optimization, the authors derived a function that maps image complexity to optimal SSIM window size. As such, they were able to produce a dynamic value for the window size parameter. In our work, we search for a unique window size that is optimal for a given dataset, and we dynamically rescale each image based on viewing distance through Optimal Scale Selection (OSS), thus implicitly impacting the relationship between image size and window size. Finally, the analyzed work is evaluated on the LIVEv2 dataset only. Although this provides a reasonable starting point to evaluate parameters optimization, we consider it fundamental to assess the generalizability of the proposed solution and therefore extend our evaluation to other FR-IQA datasets.

Charrier et al. (2012) addressed the problem of optimizing MS-SSIM exponents α, β, and γ, dedicated respectively to the luminance, contrast, and structural components. They used maximum likelihood difference scaling, a psychophysical method that allows estimating a perceptual interval scale on image quality, to assess and tune the performance of MS-SSIM. The authors explored a 15-dimensional search space, given by three parameters at five different scales, through GA. Similarly, we exploit GA as well as PSO to explore different search spaces. We also extend our analysis to SS-SSIM optimization and show how proper parameters optimization can, in fact, reduce the gap in performance between single-scale and multiple-scale analysis. Finally, the authors specifically addressed the problem of image compression and how the corresponding artifacts affect perceived image quality. For this reason, they focused their optimization on JPEG2000-related distortions on a custom dataset, and evaluated the proposed solution both on the LIVEv2 and TID2008 datasets. Aware of the importance of resorting to a commonly-adopted benchmark, we structure our work by optimizing for five popular image quality datasets independently and performing a cross-dataset evaluation to assess the robustness and reliability of the optimized parameters.

Skurowski and Janiak (2014) devised a log–log regression-based optimization procedure to select the optimal values for MS-SSIM α, β, and γ exponents and constrained each component exponent to be the same at each scale of the processing. In our work, we experiment with different constraints and levels of optimization and provide a solution that takes advantage of such constraints in reducing the computational complexity of MS-SSIM at little-to-no cost in terms of performance drop. The authors performed a regression by minimizing least squares (L2) and least absolute deviation (L1), which required reformulating MS-SSIM to an approximated version, thus not directly addressing the official similarity measure. Their optimization procedure was run on the TID2008 dataset and cross-evaluated on the LIVE and CSIQ datasets, producing mixed results.

Recent advancements show the benefit of applying meta heuristics to optimize the SS-SSIM. Specifically, in Bakurov et al. (2020), the authors explored a varied set of meta-heuristics to optimize the α, β, and γ exponents, as well as the sliding window size, used to compute the similarity values. Experimental results point that better correlations with human-expressed MOS can be obtained. The study's outcomes motivated us to continue this research track by augmenting the number of optimizable parameters in SS-SSIM and including a wide range of MS-SSIM parameters.

In the field of FR-IQA, several SSIM improvements were proposed based on the exploitation of visual features at different spatial scales (resolutions) (Wang et al., 2003). Nevertheless, relatively few works explore how the set of weights used to adjust the relative importance of SS-SSIM's components – α, β, and γ – influences its correlation with human perception of image fidelity and quality. We speculate that this oversight has to do with the fact that Wang et al. in their manuscript, which presents a structural similarity paradigm for IQA (Wang et al., 2004), set $\alpha = \beta = \gamma = 1$ to simplify the aforementioned expression of SSIM into:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{x,y} + C2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{11}$$

Consequently, numerous libraries which implement SS-SSIM are, by default, parametrized as $\alpha = \beta = \gamma = 1$.

4. Proposed method

In this work, we apply and compare two popular and conceptually different population-based stochastic metaheuristics – Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) – to optimize the parameters of SS-SSIM, under the light of different spatial scale selection approaches and MS-SSIM. Our work's objectives go beyond mere optimization and can be summarized in by the following points:

- excel measures' correlation with subjective evaluation;
- compare different metaheuristics;
- compare single-scale SSIM against its multi-scale counterpart in the context of parameter optimization;
- study the appropriateness of different scale selection methods, in the context of SS-SSIM optimization;
- understand the usefulness of different optimization parameters;
- provide an interpretation for the learned parameters;
- verify to which extent the set of parameters learned on a given IQA database can be transferred to other, previously unseen, databases;

We will rely upon formal definition of optimization problem (OP) and its components – the search space S and the fitness function f – to describe the proposed method.

4.1. The search space

The work presented in this paper is conceptually divided into six parts, each corresponding to different types of OP characterized by a unique search space (S). In the following, we provide a detailed characterization for each search space and formally define chromosomes' representation.

4.1.1. $SS-SSIM_{(\alpha,\ \beta,\ \gamma)}$

To begin with, we consider the problem of optimizing the relative importance of SS-SSIM's components: luminance (α), contrast (β), and structure (γ). Following the original settings (Wang et al., 2004), we use an 11×11 circular-symmetric Gaussian filter that from now on, will be mentioned as *sliding-window* and represented by w, $w = \{w_i | i = 1, 2, ..., N\}$, having standard deviation of 1.5 and normalized to unit sum ($\sum_{i=1}^{N} w_i = 1$); the parameters K_1 and K_2 were set to 0.01 and 0.03.

From now on, the search space for optimization of $\{\alpha,\ \beta,\ \gamma\}$ for SS-SSIM will be denoted as $SS-SSIM_{(\alpha,\ \beta,\ \gamma)}$. Adopting the nomenclature from evolutionary computation, a given candidate-solution for an instance of continuous OP can be seen as a fixed-length chromosome of real-values whose length is directly proportional to the search space's hyper-cube dimensionality. Considering $S = SS-SSIM_{(\alpha,\ \beta,\ \gamma)}$, the chromosome c at iteration i is defined as a 3D real-valued vector of the form $\vec{X}_{c,i} = [X_{\alpha,c,i},\ X_{\beta,c,i},\ X_{\gamma,c,i}]$. Since the three real-values represent the exponents associated with SS-SSIM's components, we

considered it reasonable to bound chromosomes' values in $(0,\ 3]$ real-valued interval meaning that we allow the exponents to vary up to the cubic order. Consequently, $SS-SSIM_{(\alpha,\ \beta,\ \gamma)}$ is defined in a 3D real-valued hyper-cube bounded in $(0,\ 3]$ intervals at each dimension.

Given the fact that we explore SS-SSIM in the context of different spatial scale selection approaches, to distinguish between benchmarks that involve optimization of $SS-SSIM_{(\alpha,\ \beta,\ \gamma)}$ after applying SSS and OSS methods, we extend the above-defined nomenclature by including the scale selection approach in the superscript. In this sense, $SS-SSIM_{(\alpha,\ \beta,\ \gamma)}$ after applying SSS will be denoted by $SS-SSIM^{sss}_{(\alpha,\ \beta,\ \gamma)}$, whereas after OSS, it will be denoted by $SS-SSIM^{oss}_{(\alpha,\ \beta,\ \gamma)}$.

At this point, it might seem that, since the number of fitting parameters can be said to be relatively *small*, one could be tempted to simply apply an exhaustive search of the parameters' space. However, our rationale is the most appropriate solution for the following reasons. Although we bound the search space in $(0,\ 3]$ hype-cube, the space is continuous, which means that the set of candidate solutions is, in theory, infinite. Nevertheless, even if one admits discretization $SS-SSIM_{(\alpha,\ \beta,\ \gamma)}$ (in this paper, we consider 3 decimal points), there will be 3000^3 candidate-solutions to evaluate. Considering that one candidate-solution takes, in average terms, 5 s to be evaluated on TID2008 (made of 1700 reference–distortion pairs), divided into batches of size 100, using an MSI GS65 Stealth Thin 8RF computer and GPU capabilities, then 3000^3 candidate-solutions will take 135 000 000 000 s or 1 562 500 days. Whereas a single execution (run) of an optimization heuristic like GA, if parametrized as in Table 2, takes about 720 s. Moreover, as one will be presented further, in this study, we consider significantly more complex search spaces comprising a higher amount of optimizable parameters (up to 45); this makes the exhaustive search of the parameters' space even less appropriate for this kind of experiments.

4.1.2. $SS-SSIM_{full}$

In continuation, we extend the aforementioned problem 4.1.1 by including six additional parameters in the joint optimization process. Following the original definition of SS-SSIM (Wang et al., 2004), Wang et al. introduced two scalar constants K_1 and K_2, to scale C_1, C_2 and (indirectly) C_3; the latter were included as functions of the dynamic range of the pixel values (L) to avoid instability during the calculations of SS-SSIM. Wang et al. defined $K_1 = 0.01$ and $K_2 = 0.03$, and mentioned that these values were *"somewhat arbitrary"* and that SS-SSIM was *"fairly insensitive to variations of these values"*. However (Silvestre-Blanes, 2011) proved the opposite. For this reason, we have considered it important to include both K_1 and K_2 in the joint optimization of SS-SSIM's parameters. Since both normalization constants are desired to be small, they were bounded in $(0,\ 0.3]$ in the chromosome.

Additionally, we have considered the size of the sliding-window and the standard deviation used to initialize its values (denoted by w and σ, respectively). It happens that the default $w = 11$ and $\sigma = 1.5$ might be correlated to the viewing conditions used in the subjective tests of the underlying visual data. By allowing to optimize w and σ, we optimize the SS-SSIM with respect to the viewing conditions of the subjective tests, more specifically, the visual resolution (number of pixels per degree of the visual field).

Finally, since SS-SSIM works upon a systematic application of a circular-symmetric Gaussian filter, which is equivalent to a convolution, we have considered optimizing the stride of the aforementioned filter (denoted by s) and its dilation rate (denoted by d). In the original settings (Wang et al., 2004), the sliding-window's stride and dilation were set to $(1,\ 1)$ and 1 respectively, but no clear justification was provided for this setup, neither were experiments done to verify the suitability of these parameters. The stride can be characterized as the amount of movement between applications of the filter to an input image and, in the majority of cases, it is symmetrical in height and width; by default, most convolutions (including SS-SSIM's) use $s = (1,\ 1)$. However, s can be enlarged, which will affect how the filter is applied and, consequently, the size of the resulting feature map. More specifically, the larger the stride, the fewer operations are performed to apply convolution. We speculate that, by using a larger stride, we could reduce the computational effort of SS-SSIM without the loss of performance. A dilated convolution can be formally defined as $\left(k *_d f\right)_t = \sum_{\tau=-\infty}^{\infty} k_\tau \cdot f_{t-d\tau}$, where f is the input signal, k is the filter (a.k.a. kernel), and d the dilation factor (Yu & Koltun, 2016). Dilated convolutions, unlike standard convolutions, allow the filter to use the spatial information only at each dth pixel. Consequently, this allows for enlarging the receptive field without loss of resolution or coverage. Motivated by the potential advantages of the aforementioned facts, we decided to include the dilation factor d in the optimization of SS-SSIM.

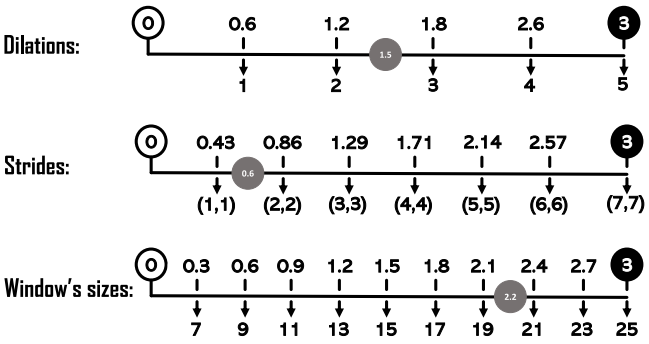

Fig. 1. Illustration of how dilation rate, stride, and sliding-window size are mapped to the interval $]0,\ 3]$, the set of chromosome's admissible values. Assuming that d, s, and w are defined in the chromosome as 1.5, 0.6, and 2.2, respectively (these are represented by the gray circles in the corresponding real lines), the SS-SSIM will have a 21×21 circular-symmetric Gaussian filter, applied with a stride $(2,\ 2)$ and a dilation rate of 3.

From now on, the search space for SS-SSIM's optimization made of $\{\alpha,\ \beta,\ \gamma,\ K_1,\ K_2,\ d,\ s,\ w,\ \sigma\}$ will be denoted as $SS-SSIM_{full}$. Considering $S = SS-SSIM_{full}$, the chromosome c at iteration i is then formally defined as a 9D real-valued vector of the form $\vec{X}_{c,i} = [X_{\alpha,c,i},\ X_{\beta,c,i},\ X_{\gamma,c,i},\ X_{K_1,c,i},\ X_{K_2,c,i},\ X_{d,c,i},\ X_{s,c,i},\ X_{w,c,i},\ X_{\sigma,c,i}]$ which values are allowed to vary in $(0,\ 3]$.

To deal with the search spaces which include integer-based parameters (which is the case of stride, dilation factor, and window size), we have considered a mapping between integer values and the continuous space, accounting for the fact that we are searching for the best combination of the aforementioned parameters from the perspective of continuous optimization problem-solving. To include parameters d, s, and w in the chromosome, we have decided to create a special mapping from hyper-cube's dimensions which regard dilation factor, stride, and window's size to the set of *admissible* integer-based quantities. More specifically, we have divided the $(0,\ 3]$ real-valued interval in even sub-intervals, each representing an admissible value in the mapped space. In such a way, we easily convert continuous values into integer-based quantities. Our inspection of the experimental results (see Section 6) proves that this approach is effective in widely exploring the search space. Fig. 1 illustrates such continuous→integer mapping. Each horizontal line corresponds to $(0,\ 3]$ real-valued intervals respective to a given dimension in the hyper-cube. The vertical arrows delimit the aforementioned even sub-intervals such that the upper values are the sub-intervals' upper-bounds, whereas the lower values are the mapped integer-based values for the respective parameters. Assuming that, in the chromosome, $X_{d,c,i} = 1.5$, $X_{s,c,i} = 0.6$ and $X_{w,c,i} = 2.2$ (these values are represented by the gray circles in the corresponding horizontal lines), the SS-SSIM will have a 21×21 circular-symmetric Gaussian filter, applied with a stride $(2,\ 2)$ and a dilation rate of 3. As illustrated in the figure, in our experiment, we define $d \in \{1,\ 2,\ 3,\ 4,\ 5\}$, $s \in \{(1,\ 1),\ (2,\ 2),\ (3,\ 3),\ (4,\ 4),\ (5,\ 5),\ (6,\ 6),\ (7,\ 7)\}$ and $w \in \{7,\ 9,\ 11,\ 13,\ 15,\ 17,\ 19,\ 21,\ 23,\ 25\}$.

Similarly to the nomenclature adopted in 4.1.2, $SS-SSIM_{full}$ after applying SSS will be denoted by $SS-SSIM^{sss}_{full}$, whereas after OSS by $SS-SSIM^{oss}_{full}$.

4.1.3. $MS-SSIM_{(\alpha_j=\beta_j=\gamma_j)}$

Besides optimizing SS-SSIM, in this work, we also considered its multi-scale version (MS-SSIM). To start with, we considered the problem of finding the optimal weights for MS-SSIM's spatial scales, assuming $\alpha_j = \beta_j = \gamma_j$ as in Wang et al. (2003), where j represents the jth scale and $j = \{j_i | i = 1, 2, 3, 4, 5\}$. In this sense we wanted to revise, in a data-driven approach, the appropriateness of the set of weights proposed through the experimental setup described in Wang et al. (2003).

From now on, we denote the search space for MS-SSIM made of $\{\iota_1, \iota_2, \iota_3, \iota_4, \iota_5\}$ as $MS-SSIM_{\alpha_j=\beta_j=\gamma_j}$, where $\iota_j = \beta_j = \gamma_j$ for $j < 5$ and $\iota_j = \alpha_j = \beta_j = \gamma_j$ for $j = 5$. Under this perspective, the chromosome c at iteration i is defined as a 5D real-valued vector of the form $\vec{X}_{c,i} = [X_{\iota_1,c,i}, X_{\iota_2,c,i}, X_{\iota_3,c,i}, X_{\iota_4,c,i}, X_{\iota_5,c,i}]$. Similarly to 4.1.1 and 4.1.5, we allow the exponents to vary up to the cubic order.

4.1.4. $MS-SSIM_{(\alpha_M,\ \beta_j\neq\gamma_j)}$

We consider that assuming equal importance for different SSIM components calculated at each scale (i.e., $\alpha_j = \beta_j = \gamma_j$), proposed by Wang et al. to simplify parameter selection (Wang et al., 2003), might deteriorate MS-SSIM's potential to estimate the perceived visual quality. For this reason, we have decided to relax the aforementioned assumption and optimize the relative importance of each component at each scale. That is, when combining measurements of contrast and structure at different scales, we allow the search-algorithms to optimize the exponents β_j and γ_j, $\forall\ j \in \{1, 2, 3, 4, 5\}$ and α_j at scale $j = 5$.

From now on, we denote the search space for MS-SSIM made of $\{\beta_1, \gamma_1, (...), \beta_5, \gamma_5, \alpha_5\}$ as $MS-SSIM_{(\alpha_M,\ \beta_j\neq\gamma_j)}$. Under this perspective, the chromosome c at iteration i is defined as a 11D real-valued vector of the form $\vec{X}_{c,i} = [X_{\beta_1,c,i}, X_{\gamma_1,c,i}, X_{\beta_2,c,i}, X_{\gamma_2,c,i}, (...), X_{\beta_5,c,i}, X_{\gamma_5,c,i}, X_{\alpha_5,c,i}]$, which values are allowed to vary up to the cubic order.

4.1.5. $MS-SSIM^{(0,\ 1)}_{(\alpha_j=\beta_j=\gamma_j)}$

If one applies a periphrasis to the MS-SSIM, one could get something like an aggregation of SS-SSIM components, calculated on a range of plausible scales. In fact, the similarity indexes that are calculated at different MS-SSIM scales use exactly the same 11×11 circular-symmetric Gaussian filter, generated with $\sigma = 1.5$, and are applied with $s = (1, 1)$ and $d = 1$; besides that, all the measures use equal scaling constants ($K_1 = 0.01$ and $K_2 = 0.03$). We speculate that, by assuming equality of s, d, w σ, K_1, and K_2 during the calculation of similarity indexes at every spatial scale, the performance of MS-SSIM might be under-explored. For this reason, we decided to (jointly) optimize not only the scales' relative importance but also the parameters of SSIMs at different spatial scales. At this point, we define the following search space: $\{\iota_1, K_{(1,\ 1)}, K_{(2,\ 1)}, s_1, d_1, w_1, \sigma_1, (...), \iota_5, K_{(1,\ 5)}, K_{(2,\ 5)}, s_5, d_5, w_5, \sigma_5\}$, where $\iota_j = \beta_j = \gamma_j$ for $j < 5$ and $\iota_j = \alpha_j = \beta_j = \gamma_j$ for $j = 5$.

Furthermore, we consider it necessary to revise the suitability of the range of scales proposed by Wang et al. (2003) when designing MS-SSIM. In fact, the authors did not provide a clear justification for the set of considered spatial scales when computing MS-SSIM. We speculate that an unaware inclusion of inner similarity indexes, calculated across all the five scales, may introduce a fruitless computational burden to the measure and deteriorate its performance. To shed light on this issue, we decided to extend the aforementioned search space by adding a combinatorial sub-space defined as $\{b_1, b_2, b_3, b_4, b_5\}$, where $b_j \in \{0, 1\}\ \forall\ j \in \{0, 1, 2, 3, 4, 5\}$ and works as an interrupter for scale j. In other words, if at a given scale j, a candidate-solution assumes value 1, then the jth scale is included in MS-SSIM's calculation; otherwise, it is excluded.

Given the combinatorial nature of the proposed extension, we decided to process candidate-solutions in a parallel fashion alternating between continuous and combinatorial sub-spaces. Consequently, we adapted GA to act upon the respective sub-spaces in parallel, manipulating each with a distinct type of operators (consult Section 5.3 for more details). Under this perspective, chromosome c at iteration i is defined as a tuple of two vectors. The first is a 35D vector of the form $\vec{X}_{(c,\ i)} = [X_{(\iota_1,\ c,\ i)}, X_{(K_{(1,\ 1)},\ c,\ i)}, X_{(K_{(2,\ 1)},\ c,\ i)}, X_{(s_1,\ c,\ i)}, X_{(d_1,\ c,\ i)}, X_{(w_1,\ c,\ i)}, X_{(\sigma_1,\ c,\ i)}, (...), X_{(\iota_5,\ c,\ i)}, X_{(K_{(1,\ 5)},\ c,\ i)}, X_{(K_{(2,\ 5)},\ c,\ i)}, X_{(s_5,\ c,\ i)}, X_{(d_5,\ c,\ i)}, X_{(w_5,\ c,\ i)}, X_{(\sigma_5,\ c,\ i)}]$, which values are allowed to vary in $(0,\ 3]$ real-valued intervals. The second is a 5D vector of the form $\vec{X}_{(c,\ i)} = [X_{(1,\ c,\ i)}, X_{(2,\ c,\ i)}, X_{(3,\ c,\ i)}, X_{(4,\ c,\ i)}, X_{(5,\ c,\ i)}]$, which values can be either 1 or 0. From now on, we will define this search space as $MS-SSIM^{(0,\ 1)}_{(\alpha_j=\beta_j=\gamma_j)}$.

4.1.6. $MS-SSIM^{(0,\ 1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$

Following the rationale exposed in Section 4.1.4, we relax the assumption $\alpha_j = \beta_j = \gamma_j$ in the search space $MS-SSIM^{(0,\ 1)}_{(\alpha_j=\beta_j=\gamma_j)}$, which generates another search space that we denote by $MS-SSIM^{(0,\ 1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$. In fact, this is precisely the same search space that differs only by the first representation of the chromosome, which now is a 40D vector: $\vec{X}_{(c,\ i)} = [X_{(\beta_1,\ c,\ i)}, X_{(\gamma_1,\ c,\ i)}, X_{(K_{(1,\ 1)},\ c,\ i)}, X_{(K_{(2,\ 1)},\ c,\ i)}, X_{(s_1,\ c,\ i)}, X_{(d_1,\ c,\ i)}, X_{(w_1,\ c,\ i)}, X_{(\sigma_1,\ c,\ i)}, (...), X_{(\beta_5,\ c,\ i)}, X_{(\gamma_5,\ c,\ i)}, X_{(K_{(1,\ 5)},\ c,\ i)}, X_{(K_{(2,\ 5)},\ c,\ i)}, X_{(s_5,\ c,\ i)}, X_{(d_5,\ c,\ i)}, X_{(w_5,\ c,\ i)}, X_{(\sigma_5,\ c,\ i)}]$, which values are allowed to vary in $(0,\ 3]$ real-valued interval. We find it important to highlight that the second representation is kept the same as in $MS-SSIM^{(0,\ 1)}_{(\alpha_j=\beta_j=\gamma_j)}$.

4.2. The fitness function

Since the goal of our OPs is to find a set of parameters for an SSIM-based IQA measure that maximizes its similarity with the subjective evaluation provided by human observers (the target), we formalized the fitness function f as Spearman's rank correlation coefficient (SRCC) between both measures — the subjective evaluation and the respective outcome of SSIM; in fact, the SRCC is a widely accepted and used evaluation measure for IQA metrics in the community (Wang & Bovik, 2006; Wang et al., 2004). In such a way, $f : S \rightarrow [-1, 1]$, with higher values representing higher similarity when the target is MOS; the opposite when the target is DMOS. Fig. 2 provides a detailed description of the fitness calculation procedure for a given candidate-solution s considering:

5. Experimental setup

The objective of this section is to describe the IQA databases and the experimental parameters that were used in our benchmark.

5.1. IQA databases

The experimental results of this document are reported on five well-known databases for assessing image quality aspects. In this subsection, the reader can find a detailed description of these databases. It is important to highlight that two of them, VDID2014 and CID:IQ, are dedicated viewing distance-changed image databases, i.e., the underlying perceptual experiments were conducted at (two) different viewing distances. For a summarized description, the reader is referred to Table 1.

5.1.1. TID2008

The Tampere image database 2008 (TID2008) is a well-known and publicly available database that satisfies the main requirements for evaluating full-reference metrics (Ponomarenko et al., 2009a, 2009b). The crux is that the database was created upon reference images that comprise a wide variety of scenes and contains several different types of distortion that relate to various peculiarities of the HVS. More specifically, TID2008 was built from 25 512×384 reference images

Let s be a candidate-solution, S the underlying search space related to a given IQA measure IQA_m, (X, Y) a tuple of reference-distortion pairs and MOS the respective target:

1. if $s \in S$:
   (a) estimate the objective quality scores by parametrizing IQA_m with s;
   (b) compute SRCC between values obtained in 1.a) and the target; assign it as the fitness value of s;
2. else:
   (a) assign s a *very bad* fitness value (like -1.0);

Fig. 2. Pseudo-code for fitness-value calculation-

Table 1
Summary characteristics of IQA databases considered in our experiments. Notice that the columns W_ixH_i and D/H_i stand for image resolution and viewing distance in terms of image height, respectively, while *References, Distortions, and Pairs* refer to the number of reference images, distortion types, and resulting reference–distortion pairs, respectively.

Name	W_ixH_i	D/H_i	#References	#Distortions	#Pairs
TID2008 (Ponomarenko et al., 2009a, 2009b)	512 × 384	3	25	17	1700
TID2013 (Ponomarenko et al., 2015)	512 × 384	3	25	24	3000
CSIQ (Larson & Chandler, 2010)	515 × 515	5	30	6	866
VDID2014 (Gu et al., 2015)	768×512, 512×512	4, 6	8	4	160
CID:IQ (Liu et al., 2014)	800 × 800	2.45, 4.75	23	6	690

taken from the Kodak lossless true color image suite (Franzen, 1999), except for 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 a database containing 1700 reference–distortion pairs. More than 800 volunteers with different cultural levels (researchers, tutors, and students) from three different countries (Finland, Italy, and Ukraine), subjectively evaluated the visual quality of distorted images. 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 database, namely a complete enumeration of distortion types and levels, can be found in Ponomarenko et al. (2009a, 2009b).

5.1.2. TID2013

The Tampere image database 2013 (TID2013) is an extension of the aforementioned TID2008, which contains more distortion types and levels (Ponomarenko et al., 2015). Similarly to TID2008, the database is publicly available and rapidly became popular in the scientific community. Ponomarenko et al. motivated the creation of TID2013 mainly by the new types of distortions and improved methodologies of quantitative subjective tests. More specifically, the authors of TID2013 re-utilized the reference images used for TID2008. For each reference image, the authors applied 24 types of distortions for each reference image with five different levels each, resulting in a database containing 3000 reference–distortion pairs. The visual quality of distorted images was gathered by performing 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 database, namely a complete enumeration of distortion types and levels, can be found in Ponomarenko et al. (2015).

5.1.3. CSIQ

The computational and subjective image quality (CSIQ) database is another popular database for IQA of measures and other aspects of image quality. The main reason for the inclusion of this database in our benchmark was the fact that it was built upon completely different reference images than those in TID2008, TID2013, and VDID2014. The CSIQ database was built from 30 512 × 512 reference images taken from public-domain sources, predominantly from the U.S. national park service. For each reference image, the authors applied six types of distortions with five different levels for each type of distortion, resulting in a database containing 866 reference–distortion pairs. The distortion types comprise commonly encountered acquisition, registration, and compression artifacts: global contrast decrements, additive pink Gaussian noise, Gaussian blurring, JPEG compression and JPEG2000 compression. Thirty-five different volunteers subjectively evaluated the visual quality of the distorted images. 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 database, namely a complete enumeration of distortion levels, can be found in Larson and Chandler (2010).

5.1.4. CID:IQ

Unlike the previously presented TID2008, TID2013, and CSIQ sources, CID:IQ is a viewing distance-changed IQA database, i.e., the underlying perceptual experiments were conducted at (two) different viewing distances (Liu et al., 2014). Moreover, it is uniquely characterized by following *objective* design principles when selecting the reference images, i.e., the reference images were selected based on objective analytical procedures following the most recent achievements in the research field. CID:IQ was built from 23 800 × 800 reference images. For each reference image, the authors have applied six types of distortions with five different levels of degradation each, resulting in a database containing 690 reference–distortion pairs. The distortion types comprise commonly encountered acquisition, registration, and compression artifacts: Poisson noise, Gaussian blur, constant hue minimum ΔE gammut mapping, SGCK gammut mapping, JPEG compression and JPEG2000 compression. Seventeen different volunteers subjectively evaluated the visual quality of distorted images. The subjective test was carried out at a viewing distance of 2.45 and 4.75 times the image height. The authors reported their results in the form of MOS. Further details about the database can be found in Liu et al. (2014).

5.1.5. VDID2014

The VDID2014 is another viewing distance-changed IQA database, first published in 2015 with the objective of deploying the impact of viewing distances and image resolutions on IQA (Gu et al., 2015). VDID2014 was built from eight reference images with resolutions of 768 × 512 and 512 × 512. It is worth noting that the largest four are original from the Kodak lossless true color image suite (Franzen, 1999). For each reference image, the authors have applied four types of distortions with five different levels for each type of distortion, resulting

in a database containing 160 reference–distortion pairs. The distortion types comprise commonly encountered acquisition, registration, and compression artifacts: white noise in the RGB components, Gaussian blur, JPEG compression, and JPEG2000 compression. 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 database can be found in Gu et al. (2015).

5.2. Data usage

It is worth highlighting that we approach the problem of optimizing SSIM-based measures from the perspective of machine learning (ML) in this work. Since ML refers to the task of inducing a general pattern when provided a set of training (a.k.a. learning) examples, the ML algorithms, in this work GA and S-PSO, are expected to achieve a fair generalization on unseen examples of the same pattern. A common issue faced in almost every ML application is overfitting — a situation when the algorithms simply memorize the set of training examples instead of learning the underlying pattern. To ensure that our ML algorithms are suggesting parameters for SSIM that are as good on previously unseen reference–distortion pairs as on those used for learning, although being of the same kind, we performed internal cross-validation and computed both *training* and *unseen* fitness. More specifically, during the optimization of an IQA measure on a given database, we have left out 30% of the reference–distortion pairs to estimate the algorithms' generalization ability and then create the possibility to, further, compare this estimate with the real fitness observed on previously unseen data. The IQA databases were partitioned randomly, using a different seed for a pseudo-random number generator at each run, following the Monte Carlo cross-validation scheme.

Given the fact we worked with real-world images, i.e., data-structures of considerable size, to accelerate our algorithmic procedures, we decided to use only 50% of reference–distortion pairs of the training partition at each iteration; these were selected at random and without replacement. In other words, we used batch-training where the batch-size equals 50% of the training partition. Such a training scheme allowed us to significantly reduce the training times while still ensuring an equal positive probability of using any training data instance. Moreover, we can confidently say that such a training scheme did not introduce a *significant* information loss from the obtained results.

5.3. Parameters

In this sub-section, the reader can find a detailed description of the parameters used in our experiments. It is paramount to highlight that the objective of this paper is not to perform an exhaustive hyper-parameter exploration of the optimization algorithms. Instead, our goal is to prove the suitability of the proposed method for IQA measures' optimization. Therefore, we experimented with two semantically diverse algorithms whose parameters were chosen according to our best understanding and findings from the literature. In the first paragraph, the reader is exposed to the set of parameters that are shared across the optimization metaheuristics, along with the computational cost. The second and the third paragraphs describe, in detail, the parameters of GA along with the aforementioned adaptation of GA towards simultaneous processing of combinatorial and continuous sub-spaces in the case of $MS-SSIM^{(0,\ 1)}_{(\alpha_j=\beta_j=\gamma_j)}$ and $MS-SSIM^{(0,\ 1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$. The third paragraph chronicles the parameters of PSO. Finally, the last paragraph introduces details regarding the search spaces' hyper-cube. For a summarized description, the reader is referred to Table 2.

Considering the stochastic nature of the search-algorithms employed and results' volatility upon data partitions (i.e., to provide a robust and statistically-consistent analysis of experimental results), we have repeated experiments ten times (runs), each with a different pseudo-random number generator (a.k.a. *seed*), used for partitioning the data, algorithms' initialization and their subsequent execution. We have fixed an equal population-size and number of generations at values 40 and 50, respectively, throughout our experiments; the exception for this setting was the proposed adaption of GA, which from now on, will be referenced as GA_2, where the population-size and number of generations were fixed at 25 and 80, respectively. Thus, the resulting computational effort for each experiment was defined at 2000 fitness evaluations per algorithm at one run. As it was described in 4.1, the search space consists of a multidimensional hyper-cube whose dimensions are defined in $(0,\ 3]$ real-valued interval (recall that the number of dimensions varies according to the problem-instance). For this reason, the initial candidate-solutions were generated under continuous uniform distribution $\sim U(0,\ 3]$. The stopping criterion for the search-algorithms was defined as the number of generations.

In our experiments, GA was used with tournament selection with a selection pressure of 10%. Such a high selection pressure was adopted to foster the convergence given a reduced amount of generations that GA was provided. The survival was elitist, i.e., the parent that presents the best training fitness, if better than the best offspring, was automatically copied to the next generation. In this sense, it becomes possible to maintain the traits of the most fitting individuals and flow their genetic material to the next generation, improving, therefore, the convergence (Du et al., 2018). It is worth noting that, in this work, the GAs are divided across the two kinds of the search space: purely continuous and mixed (continuous and combinatorial); to distinguish between both, the variant of GA that is applied in a mixed search space will be referenced as GA_2. For the purely continuous search spaces, which is the case of $SS-SSIM_{(\alpha,\ \beta,\ \gamma)}$, $SS-SSIM_{full}$, $MS-SSIM_{(\alpha_j=\beta_j=\gamma_j)}$ and $MS-SSIM_{(\alpha_M,\ \beta_j=\gamma_j)}$, GA was used with geometric crossover and ball-mutation. Both operators are representation independent search-operators, defined in precise geometric terms using the notions of line segment and ball, that generalize search operators for the major representations used in GAs, such as binary strings, real vectors, permutations, and syntactic trees (Moraglio & Poli, 2004). In the case of geometric crossover, the offspring always stands on the segment joining the points representing the parents in the D-dimensional hyper-cube. The box-mutation consists of a random perturbation of chromosome values in a given range; it was applied at every position of the chromosome with a probability of 0.3 and a perturbation magnitude generated from $N(\mu = X_{j,c,i},\ \sigma = 0.1)$, where j represents the jth position in the chromosome, and c and i uniquely identify the chromosome c at iteration i. For the search spaces consisting of a mix between continuous and combinatorial sub-spaces, which is the case of $MS-SSIM^{(0,1)}_{full_{(\alpha_j=\beta_j=\gamma_j)}}$ and $MS-SSIM^{(0,1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, we used an adaption of GA where variation operators act upon each sub-space appropriately in a parallel fashion. To operate upon continuous sub-spaces, we decided to reuse the aforementioned operators and parameters; whereas to manipulate candidate-solutions co-represented by 5D binary vectors, we considered using point crossover and bit-flip mutation as defined in Mitchell (1998); the latter was applied at every position of the chromosome with a probability of 0.3 (similarly to the aforementioned ball mutation). The probabilities of applying crossover ($P(C)$) and mutation ($P(M)$) were set at 0.7 and 0.3, respectively, following the recommendation provided in the literature (Mitchell, 1998), independently of the type of search space.

As we explained in 4.1.5, GA_2 was designed to operate upon chromosomes represented by a tuple of vectors: the first represents SS-SSIMs' parameters at different spatial scales, the second indicates which spatial scales to include in the MS-SSIM's calculation. In this sense, chromosome processing alternates between continuous and combinatorial sub-spaces (as such, variation operators). While the optimization of continuous sub-space was conducted traditionally, iteration after iteration, optimization of the combinatorial sub-space was paused and only manipulated at each 8th iteration (binary update frequency). We

Table 2
Enumeration of hyper-parameters for GA and PSO.

Algorithm(s)	Parameter	Value
{GA, S-PSO, GA_2}	#runs	10
	#fitness evaluations	2000
	Initialization	$U_{(0,\ 3]}$
	Stopping criteria	#generations
{GA, PSO}	#generations	40
	Population size	50
GA	Selection type	Tournament
	Selection pressure	0.1
	Elitism	True
	Crossover	Geometric
	Mutation	Ball
	$P(C)$	0.7
	$P(M)$	0.3
GA_2	#generations	80
	binary update frequency	8
	Population size	25
	Crossover	(Geometric, Point)
	Mutation	(Ball, Flip)
PSO	Synchronization of the swarm	True
	Neighborhood model	*gbest*
	Social factor	1.0
	Cognitive Factor	1.0
	Inertia weight	0.79

introduced such a *delay* to allow GA_2 to adjust continuous parameters after, in our opinion, disruptive changes that would result in the combinatorial sub-space.

Following the work of Bartashevich et al. (2020), PSO was used with equal weights for acceleration coefficients $C_1 = C_2 = 1.0$, and the inertia weight w was set to 0.79. At a given iteration, particles in the swarm are updated with the same *gbest*, obtained from examining the whole swarm. This operation means that the neighborhood model we used in our experiments is *gbest*, and the swarm's position updates are synchronous. In such a way, the update can be performed in a computationally more efficient way, therefore accelerating the experiments' execution.

Whenever a given candidate-solution leaves the search space's hyper-cube, i.e., when some of the coordinates at a given position get out of the $(0,\ 3]$ real-valued interval, the coordinates were randomly reinitialized following $U_{(0,\ 3]}$. This action is based on the scientific community's common practice regarding hard-constrained continuous problem-solving (Bakurov et al., 2021; Bartashevich et al., 2017).

6. Experimental results

This section presents and discusses the experimental results; due to the vast amount of experimental findings, we have divided this section into five parts: the overall performance, the statistical assessment, the cross-database analysis, the performance by distortion groups, and the parameters' presentation and discussion. We highlight the fact that the analysis is mostly based on SRCC, which is a widely accepted evaluation measure for IQA metrics (Wang & Bovik, 2006; Wang et al., 2004).

6.1. Overall performance of optimization framework

Fig. 3 exhibits a series of five box-plots, one per IQA database. The box-plots show the distribution of the SRCC, calculated on the unseen data, between the subjective evaluation provided by the human observers and the objective evaluation of the proposed improvements of the SSIM. More specifically, the y-axis reports the SRCC, whereas the x-axis identifies the target search spaces (following the nomenclature defined in 4.1). The three stylized lines in each sub-figure represent the baseline SSIMs. The baseline SS-SSIM is present in two versions: after the SSS, illustrated in black dashed lines, and after the OSS, shown in blue dotted lines; also, the baseline MS-SSIM, calculated after gray-scaling reference–distortion pairs, is provided in red dot–dashed lines. The boxes' colors stand for different optimization algorithms: blue for GA, golden for S-PSO and green for GA_2. Notice that GA_2 was only used for $MS-SSIM^{(0,1)}_{(\alpha_j=\beta_j=\gamma_j)}$ and $MS-SSIM^{(0,1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, the search spaces whose candidate-solutions are co-represented by 5D binary vectors.

From Fig. 3, it becomes clear that optimizing SS-SSIM on the set of parameters $\{\alpha,\ \beta,\ \gamma,\ K_1,\ K_2,\ d,\ s,\ w,\ \sigma\}$ generally yields better SRCC than on $\{\alpha,\ \beta,\ \gamma\}$ alone, independently of the scale selection method. It is worth detailing that $SS{-}SSIM_{full}$ generally outperforms the baseline variants of SSIM, including the multi-scale extension (MS-SSIM). By comparing two scale selection methods, one can conclude that the advantage of applying a more complex OSS is not granted; first, for some IQA databases (like CSIQ and CID:IQ), the baseline SS-SSIM after SSS seems to outperform SS-SSIM after OSS, which results in a more productive optimization of SS-SSIM after SSS when compared to OSS; second, for the databases where OSS seems more advantageous for SS-SSIM, the benefit of applying OSS, although visible, seems to be minimal when compared to SSS; the only IQA database where OSS seems to provide a clear visual advantage is VDID2014.

By looking at MS-SSIM's optimization, it becomes clear that optimizing the relative importance of each component at each scale, instead of assuming their equality, allows the measure to achieve better SRCC. However, to our surprise, a deeper fine-tuning of MS-SSIM as in $MS-SSIM^{(0,1)}_{(\alpha_j=\beta_j=\gamma_j)}$ and $MS-SSIM^{(0,1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$ (where candidate-solutions are co-represented by 5D binary vectors), did not provide a clear visible advantage when compared to $MS-SSIM_{(\alpha_j=\beta_j=\gamma_j)}$ and $MS{-}SSIM(\alpha_M,\ \beta_j\neq\gamma_j)$, respectively. When comparing SS-SSIM with its multi-scale counterparts, it becomes clear that the fine-tuned SS-SSIM can be as good or even better; the only two IQA databases where this does not hold, although the difference seems to be relatively small, are TID2008 and TID2013. Considering that SS-SSIM can be said five times computationally less complex, we consider this finding as a strong argument to reinforce its relevance as a low-cost and high-quality IQA measure. Finally, we will comment on the differences between different optimization algorithms. To ensure a fair comparison, the comparison will be carried between GA and S-PSO alone since GA_2 is a specifically-designed adaption of GA for searching in two distinct search spaces simultaneously. By analyzing the box-plots, one can argue that, in general terms, the two algorithms behave identically; such a level of *agreement* between these two conceptually distinct metaheuristics suggests a fair level of convergence on the underlying optimization problems.

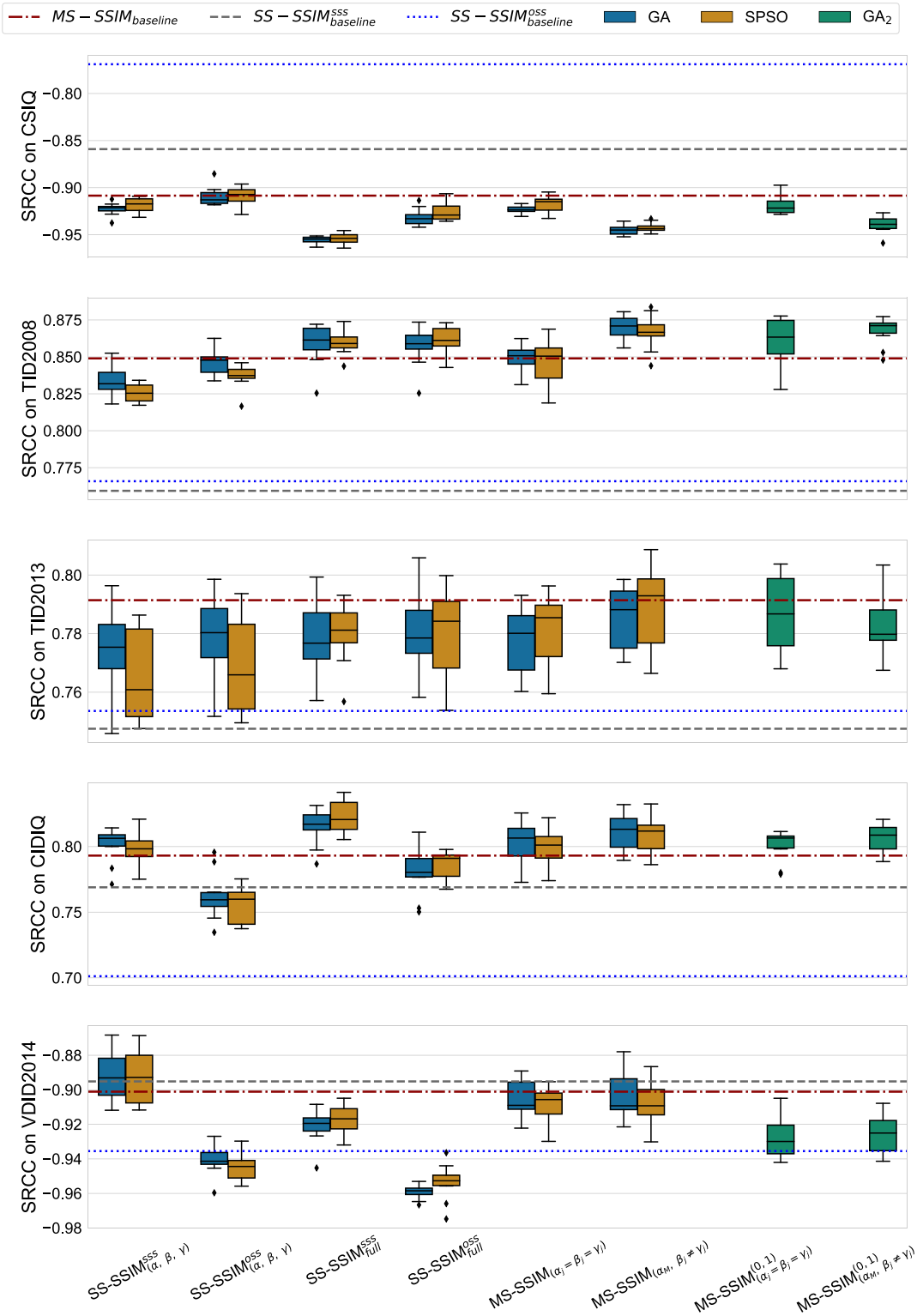

Fig. 3. Box-plot of SRCC calculated on unseen data, at a given IQA database and for each search space.

To better understand which variant of SSIM's optimization (from now on, referenced as the experiment), is the most appropriate at each IQA database, we analyzed both SRCC and the performance ranks on unseen data partitions. More specifically, on a given IQA database, we have sorted the experiments according to the SRCC and assigned a value (the rank) for each run and optimization algorithm. Then we aggregated these ranks across all the runs and selected the five most performing experiments for each IQA database (notice that lower values stand for better ranking). Additionally we computed the median SRCC on unseen data partitions across all the runs for each experiment. This information can be found in Table 3, and the reader can find the table's description in the following paragraphs.

From Table 3, we can see that for the two distance-changed IQA databases (CID:IQ and VDID2014) and CSIQ, the experiments which achieve the highest generalization ability comprise optimization of SS-SSIM$_{full}$; notice that all the remaining ranks, except for

Table 3
Rank of different SSIM's optimization forms based on their generalization ability, for each IQA database.

IQA database	Search space's ID	Algorithm	$\widetilde{SRCC}_{unseen}$	Rank
CSIQ	$\text{SS-SSIM}^{sss}_{full}$	GA	−0.951	1
	$\text{SS-SSIM}^{sss}_{full}$	S-PSO	−0.952	2
	$\text{MS-SSIM}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA	−0.936	3
	$\text{MS-SSIM}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	S-PSO	−0.934	4
	$\text{MS-SSIM}^{(0,1)}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA_2	−0.935	5
TID2008	$\text{MS-SSIM}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA	0.865	1
	$\text{MS-SSIM}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	S-PSO	0.858	2
	$\text{MS-SSIM}^{(0,1)}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA_2	0.859	3
	$\text{SS-SSIM}^{oss}_{full}$	S-PSO	0.855	4
	$\text{SS-SSIM}^{sss}_{full}$	GA	0.844	5
TID2013	$\text{MS-SSIM}^{(0,1)}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA_2	0.793	1
	$\text{MS-SSIM}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA	0.787	2
	$\text{MS-SSIM}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	S-PSO	0.783	3
	$\text{MS-SSIM}^{(0,1)}_{full_{(\alpha_j=\beta_j=\gamma_j)}}$	GA_2	0.788	4
	$\text{SS-SSIM}^{oss}_{full}$	GA	0.781	5
CID:IQ	$\text{SS-SSIM}^{sss}_{full}$	S-PSO	0.817	1
	$\text{SS-SSIM}^{sss}_{full}$	GA	0.810	2
	$\text{MS-SSIM}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA	0.806	3
	$\text{MS-SSIM}^{(0,1)}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA_2	0.806	4
	$\text{MS-SSIM}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	S-PSO	0.806	5
VDID2014	$\text{SS-SSIM}^{oss}_{full}$	GA	−0.960	1
	$\text{SS-SSIM}^{oss}_{full}$	S-PSO	−0.958	2
	$\text{SS-SSIM}^{oss}_{(\alpha,\ \beta,\ \gamma)}$	GA	−0.952	3
	$\text{SS-SSIM}^{oss}_{(\alpha,\ \beta,\ \gamma)}$	S-PSO	−0.951	4
	$\text{MS-SSIM}^{(0,1)}_{(\alpha_j=\beta_j=\gamma_j)}$	GA_2	−0.930	5

VDID2014, comprise optimization of MS-SSIM. Regarding the TID2008 and TID2013 IQA databases, which exhibit the most variety of distortion types and levels, it becomes clear that the best correlation with MOS is achieved when optimizing the multi-scale variant of SSIM after the relaxation of the $\alpha_j = \beta_j = \gamma_j$ assumption; the second-best group of experiments comprises the search space $\text{SS-SSIM}^{oss}_{full}$. By comparing $\widetilde{SRCC}_{test}$ between the most well-ranked MS-SSIM and SS-SSIM, one can notice that the difference happens to range from 0.01 (for TID2008) to 0.02 (for VDID2014); bearing in mind the increased complexity of MS-SSIM, we consider this difference minor yet benevolent for SS-SSIM's revised importance. When assessing the impact of scale selection, one can conclude that OSS allows the SS-SSIM to achieve the highest generalization for VDID2014 (where it occupies the top 4 in the ranking), TID2008, and TID2013; whereas for CSIQ and CID:IQ, it is the SSS that allows the measure to achieve the highest generalization (such experiments occupy the top 2 in the ranking). Finally, we will comment on the differences between the metaheuristics. Following the rationale exposed during the analysis of Fig. 3, the comparison will be carried out between GA and S-PSO only. From the table one can confirm the aforementioned *agreement* between GA and S-PSO — both achieve a very similar $\widetilde{SRCC}_{unseen}$; although, the difference is minimal, it is clear that GA tends to achieve slightly better ranks.

To prove the effectiveness of the optimization approach and identify potential overfitting, we analyzed the learning curves. Fig. 4 provides an illustrative example comprising two forms of SSIM's optimization: $\text{SS-SSIM}^{sss}_{full}$ and $\text{MS-SSIM}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$. Our choice was based on two factors: first, we wanted to include one single and one multi-scale variant of SSIM; second, we decided to include the experiments which, according to the analysis of Fig. 3 and Table 3, were shown to be among those which generalize better. It is worth remembering that, to reduce the training times, we conducted the optimization on batches of training data (consult 5.3 for more details); for this reason, the training curves are not monotonically decreasing functions. The following paragraph describes the figure.

From Fig. 4, one can notice that, in general terms, the optimization process does not exhibit clear overfitting patterns. Additionally, one can see that after 10–15 generations, the level of generalization tends to stabilize, meaning that no further optimization brings a significant improvement. It is worth noticing that TID2013 is the IQA database where the optimization seems to bring little-to-no benefits. A possible explanation can be found in the richness and variably of distortion types, including a wide range of color-specific distortions, making the optimization of SSIM naturally harder.

We finish this subsection by comparing the complexity of each optimization algorithm in terms of the processing time. Table 4 presents the average processing times of a single run of each experiment identified by the tuple (**Search space's ID**, **Algorithm**). It happens that the optimization algorithms' time-complexity depends on (i) the imaging database, because images usually have different size across different DBs, (ii) the IQA measure, as different measures have different time-complexity, and (iii) the experiment's type, because different experiments imply a different search-space, therefore processing time. Therefore, we averaged the experiments' processing time across databases. From the table, we can see that all the experiments involving SS-SSIM are approximately 10x faster than the MS-SSIM's. This fact is directly linked to the time-complexity of each measure: SS-SSIM computes similarity at a single spatial scale, whereas MS-SSIM computes (and then aggregates) the similarity across five different spatial scales; moreover, the number of optimized parameters for the MS-SSIM is tendentiously larger than for its single scale counterpart. It is pertinent to highlight that S-PSO takes notably less time than GA on the same experiments. This is mostly related to the fact particles' updates rely upon vectorized operations; specifically, given that all the particles in the swarm use the same *gbest*, one can make use of more optimal and pre-compiled functions and mathematical operations on array objects to update the whole swarm at once. For a more detailed comparison between GA and PSO, including an empirical comparison on a varied set of optimization problems, the reader is referred to Bakurov et al. (2021).

6.2. Statistical assessment

In this sub-section, the reader can find the statistical assessment and the analysis of the experimental results. The statistical assessment

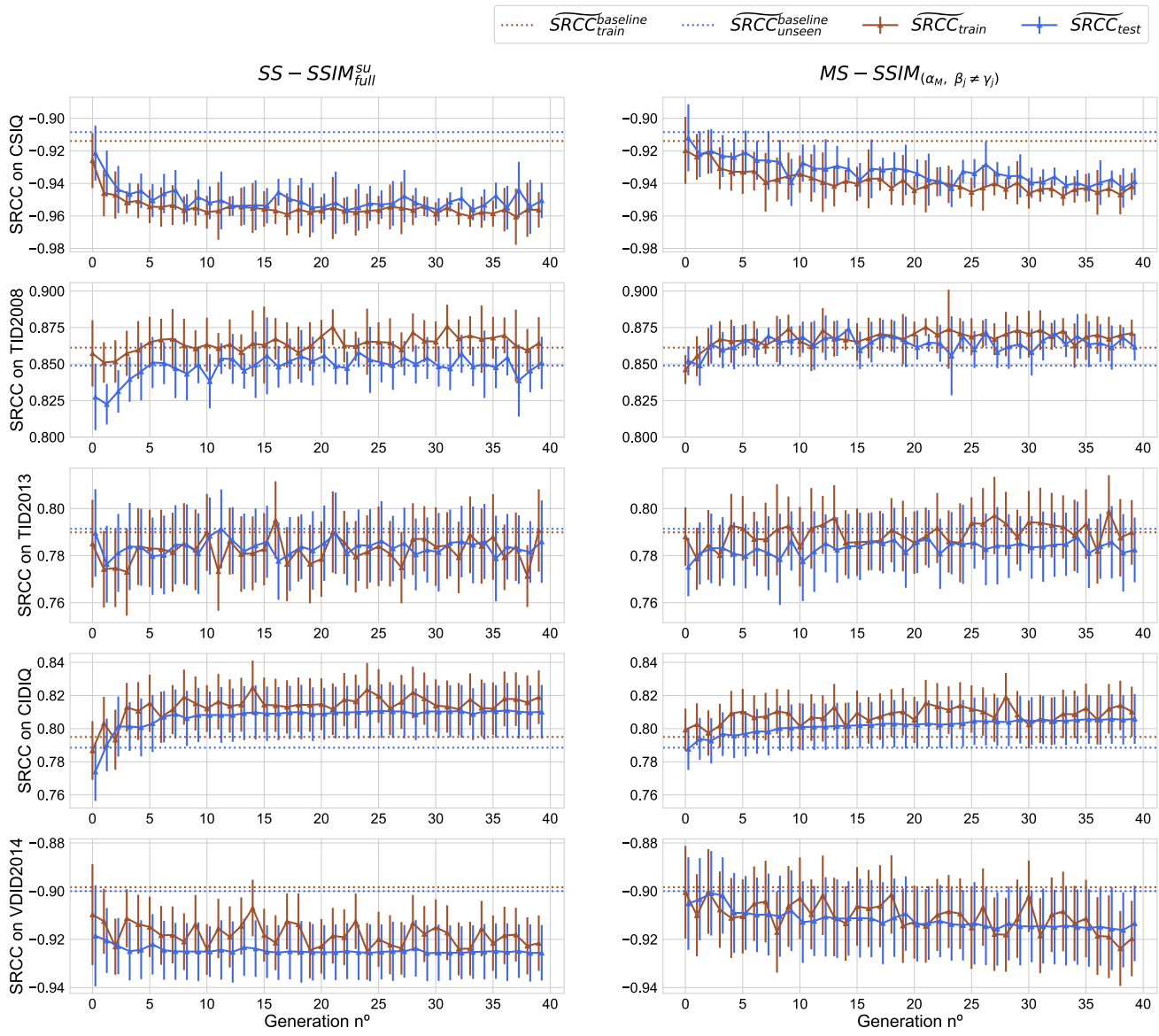

Fig. 4. Learning curves from optimizing SS-SSIM$^{sss}_{full}$ and MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$ with GA.

Table 4
Algorithms' time complexity across different search spaces.

Search space's ID	Algorithm	Time (s)
MS-SSIM$^{(0,1)}_{full_{(\alpha_j=\beta_j=\gamma_j)}}$	GA_2	12391.36
MS-SSIM$^{(0,1)}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA_2	11975.21
MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$	GA	13125.79
	SPSO	13009.36
MS-SSIM$_{(\alpha_j=\beta_j=\gamma_j)}$	GA	13212.84
	SPSO	13048.65
SS-SSIM$^{sss}_{(\alpha,\ \beta,\ \gamma)}$	GA	1422.79
	SPSO	1200.05
SS-SSIM$^{sss}_{(\alpha,\ \beta,\ \gamma)}$	GA	1520.93
	SPSO	1342.95
SS-SSIM$^{sss}_{full}$	GA	1610.06
	SPSO	909.08
SS-SSIM$^{sss}_{full}$	GA	1173.96
	SPSO	629.70

was performed through the Wilcoxon rank-sum test for pairwise data comparison (from now on, referenced as Wilcoxon's test), under the null hypothesis that the differences between two related paired samples are symmetric about zero. More specifically, we compared the SRCC between two samples, obtained on unseen data partitions and at the end of the evolutionary process (the records were taken at the last generation). It is worth pointing out that we reject the null hypothesis when the p-value of the test is smaller or equal to 5% (i.e., we assume a significance level, formally represented by α, of 5%).

Table 5 provides a statistically sustained comparison between the proposed variants of SSIM's optimization and the respective baseline. Through Wilcoxon's test, we compared the SRCC achieved by each type of SSIM's optimization against the respective baseline; recall that the comparison was performed across unseen data partitions. For those experiments where two different metaheuristics were engaged, GA and S-PSO, Wilcoxon's test was conducted on the subset involving GAs because it was shown to generalize (slightly) better (see Table 6 and its description for more details). The column *Search Space'a ID* represents the different variants of SSIM's optimization, following the nomenclature adopted in 4.1, whereas the column *Baseline* represents the respective baseline parameter set for the SSIM, as defined in Wang et al.

(2003); the column *Preprocessing*, as the name suggests, represents the preprocessing that the IQA databases were subject to, before applying the measures; the columns *Statistic* and *p-value* represent the test's statistic and p-value, respectively; finally, the column *Sign* holds value + when the proposed optimization of SSIM correlates with subjective evaluation better than the baseline, in median terms, − otherwise. See the following paragraph for the main findings.

From Table 5, one can observe a clear superiority of the proposed optimization approaches: the majority are statistically better than the respective baseline, assuming a significance level of 5%. The only optimizations for which the difference is not statistically significant are SS-SSIM$_{(\alpha,\ \beta,\ \gamma)}$ and MS-SSIM$^{(0,1)}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$; the latter is the only optimization variant that does not outperform the baseline numerically.

Table 6 exhibits the results of a series of Wilcoxon's tests conducted on semantically different subsets of experiments; with each test, we tried to achieve, in a statistically rigorous way, one of the research objectives defined in 4. The columns *Sample*$_A$ and *Sample*$_B$ label the two related paired samples — the objects of statistical assessment; the columns *Statistic* and *p-value* represent test's statistic and p-value, respectively; finally, the column *Sign* holds value + when *Sample*$_B$ achieves a higher generalization ability, in median terms, than *Sample*$_A$. See the following two paragraphs for the main findings.

The first test compares the subset of experiments conducted using GA against a related subset but involving S-PSO; from the results, one can say that GA is statistically better than S-PSO. The second test compares the two scale selection methods for the experiments involving the two kinds of SS-SSIM optimization: SS-SSIM$_{(\alpha,\ \beta,\ \gamma)}$ and SS-SSIM$_{full}$; from the test's results, one can say there is no statistically-sustained difference between SSS and OSS, although, in median terms, OSS achieves slightly better generalization ability. The third and fourth tests compare two most-performing optimization approaches for each kind of SSIM — SS-SSIM$_{full}$ against MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$; the only difference between these two tests consists of the scale selection approach used for SS-SSIM$_{full}$: the third test compares SS-SSIM$^{sss}_{full}$ with MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$, whereas the fourth test compares SS-SSIM$^{oss}_{full}$ with MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$. From the test results, one can say that MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$ is not statistically different from SS-SSIM$_{full}$, despite achieving a better overall generalization (in median terms) regardless of the scale selection method. Nevertheless, the p-value associated with the comparison of the third test is about three times smaller than that of the fourth test; this result suggests that SS-SSIMoss_{full} better approximates to MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$ than SS-SSIMsss_{full}.

In general terms, we can conclude the following. First, GA is better than S-SPO when considering the optimization tasks formalized in 4. Second, there is no statistically sustained difference across scale selection methods; when relating tests outcomes with discussions originated from Fig. 3 and Table 3, it becomes clear this topic deserves better investigation by the scientific community as the so-called Optimal Scale Selection (OSS) does not happen to be *optimal* for all the IQA databases considered in our benchmarks. Third, a *proper* parameter setting for SS-SSIM results in a performance as good as its fine-tuned multi-scale counterpart.

6.3. Cross-database analysis

To verify to which extent the set of parameters learned on a given IQA database can be transferred to other (previously unseen) databases, we created the so-called Spearman's rank cross-database correlation table (Table 7). In this table, the column *Trained on* represents the IQA databases that were used to estimate SSIM parameters (that from now on, will be referenced as training databases), whereas *Tested on* represents the databases on which those were assessed (that from now on, will be referenced as testing databases). Note that, differently from the previously shown and discussed experimental results, where the assessment was made on a given partition of data, a cross-database assessment was performed for all the reference–distortion pairs. The column *Benchmark* uniquely identifies the set of parameters learned at each search space for each training database; each identifier is provided as a tuple consisting of the SSIM's search space (labeled following the nomenclature defined in 4) and the metaheuristic that was employed. For those experiments involving SS-SSIM, the label includes the best scale selection method in the superscript. The three bottom rows regard the baseline set of parameters. The values presented in the table regard SRCC between measure's proposed optimization and databases' subjective evaluation. The red–green color range allows the reader to better understand the benchmarks with the highest SRCC: the greener the values at a given column are, the better the SRCC is. See the following two paragraphs for the main findings.

The analysis of Table 7 suggests that, as was expected, the highest correlation at a given IQA database can be achieved if the optimization system is performed on the database itself. Nevertheless, the table also suggests a high potential for transferring the learned parameters to other IQA databases. When considering the optimization of MS-SSIM, the most cross-database generalization can be observed on the tuple (MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$, GA). When trained on CID:IQ, CSIQ, TID2008 or TID2013, this benchmark is shown to outperform the baseline SSIM almost for all the databases; the only exception is VDID2014 where the baseline SS-SSIM^{oss} achieves the best results. When considering the optimization of SS-SSIM, the most cross-database generalization can be observed on the tuple (SS-SSIM$^{sss}_{full}$, S-PSO), if trained on CSIQ or TID2008, and (SS-SSIM$^{oss}_{full}$, S-PSO), if trained on TID2013.

It is important to highlight the IQA databases which allow the optimization system to achieve the highest cross-database generalization, regardless of the measure. From the table's analysis, we consider that CSIQ, TID2008 and TID2013 are the most befitting in the context of *knowledge transfer*; the reason for such adequacy can be found in databases' variability: these IQA databases happen to have the highest amount of reference images and distortion types. Under this light, it turns out to be clear why the parameters learned on VDID2014 did not exhibit a good cross-database generalization — VDID is composed of just eight reference images and four distortion types; nevertheless, the tuples (SS-SSIM$^{oss}_{(\alpha,\ \beta,\ \gamma)}$, GA) and (SS-SSIM$^{oss}_{full}$, S-PSO) trained on TID2013 happen to outperform the baseline SSIMs on VDID204.

6.4. Performance by distortion groups

Our experiments also examined the performance of the proposed parameters set on different image distortion types. Table 8 reports SRCC calculated for each distortion type. To illustrate the performance of the SSIM optimization framework by distortion groups, we relied on TID2008 as it is among the largest, most heterogeneous, and most popular IQA databases; moreover, TID2008 is primarily composed of structural distortion types which are more appropriate for the assessment of SSIM-based techniques because the latter were designed to exploit the theory about HVS's high adaptability for the extraction of structural information from the scenes.

The column *Distortion Type* uniquely identifies the databases' distortion types. The row *Training IQA DB* identifies the so-called training IQA database — the one that was used to optimize a given type of SSIM (these are reported as table's columns). Note that this table also includes the baseline SS-SSIM, reported for each scale selection method, and MS-SSIM. The table is divided into two halves: the first regards SS-SSIM, the second is MS-SSIM. The red–green color range allows the reader to better understand the benchmarks with the highest SRCC: the greener the values at a given row are, the better the SRCC is. See the following two paragraphs for the main findings. Note that the reported SSIMs were optimized using GAs.

From Table 8, one can clearly see that the proposed optimization approaches outperform the respective baselines in the majority of distortion groups. The most significant improvement in terms of SRCC can be observed from SS-SSIM$^{oss}_{full}$ trained on TID2008. Surprisingly, SS-SSIM$^{sss}_{full}$ trained on entirely different IQA database achieves slightly worse levels of correlation; this fact is further proof of high for cross-database generalization ability of the proposed optimization approach.

Table 5
Statistical assessment against the baseline. The table provides both statistic value and the respective *p*-value after Wilcoxon's paired signed rank test, under the null hypothesis that the median difference between pairs of observations is zero.

Baseline	Preprocessing	Search space's ID	Statistic	p-value	Sign
SS-SSIM	SSS	SS-SSIM$_{(\alpha,\ \beta,\ \gamma)}$	40	8.03E−09	+
		SS-SSIM$_{full}$	0	7.56E−10	+
	OSS	SS-SSIM$_{(\alpha,\ \beta,\ \gamma)}$	0	7.56E−10	+
		SS-SSIM$_{full}$	0	7.56E−10	+
MS-SSIM	grayscale	MS-SSIM$_{(\alpha_j=\beta_j=\gamma_j)}$	534	0.3178	+
		MS-SSIM$_{(\alpha_M,\ \beta_j\neq\gamma_j)}$	99	2.01E−07	+
		MS-SSIM$^{(0,1)}_{(\alpha_j=\beta_j=\gamma_j)}$	561	0.4602	–
		MS-SSIM$^{(0,1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$	193	1.78E−05	+

Table 6
Results of Wilcoxon's paired signed rank test, under the null hypothesis that the median difference between two paired samples is zero. The table provides both statistic value and the respective p-value.

Sample$_A$	Sample$_B$	Statistic	p-value	Sign
GA	S-PSO	19358.0	0.032	–
SSS	OSS	9579.0	0.565	+
SS-SSIM$^{sss}_{full}$	MS-SSIM$_{(\alpha_M,\ \beta_j\neq\gamma_j)}$	2021.0	0.083	+
SS-SSIM$^{oss}_{full}$	MS-SSIM$_{(\alpha_M,\ \beta_j\neq\gamma_j)}$	2203.0	0.268	+

Table 7
Spearman's rank cross-database correlation table. Color format is normalized per-column, with green indicating best correlation, and red least correlation.

		SRCC				
	Tested on	**CID:IQ**	**CSIQ**	**TID2008**	**TID2013**	**VDID2014**
Trained on	**Benchmark**					
VDID2014	(MS-SSIM$^{(0.1)}_{(\alpha_j=\beta_j=\gamma_j)}$, GA$_2$)	0.646	-0.808	0.630	0.668	-0.930
	(MS-SSIM$^{(0.1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA$_2$)	0.675	-0.851	0.702	0.705	-0.932
	(MS-SSIM$_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA)	0.759	-0.883	0.814	0.769	-0.916
	(MS-SSIM$_{(\alpha_j=\beta_j=\gamma_j)}$, S-PSO)	0.755	-0.877	0.811	0.766	-0.915
	(SS-SSIM$^{sss}_{(\alpha,\ \beta,\ \gamma)}$, GA)	0.639	-0.679	0.64	0.684	-0.949
	(SS-SSIM$^{sss}_{full}$, GA)	0.566	-0.737	0.592	0.649	-0.959
CID:IQ	(MS-SSIM$^{(0.1)}_{(\alpha_j=\beta_j=\gamma_j)}$, GA$_2$)	0.8	-0.904	0.848	0.782	-0.899
	(MS-SSIM$^{(0.1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA$_2$)	0.802	-0.910	0.852	0.785	-0.907
	(MS-SSIM$_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA)	0.809	-0.939	0.863	0.792	-0.904
	(MS-SSIM$_{(\alpha_j=\beta_j=\gamma_j)}$, S-PSO)	0.799	-0.912	0.856	0.789	-0.900
	(SS-SSIM$^{sss}_{(\alpha,\ \beta,\ \gamma)}$, GA)	0.8	-0.903	0.774	0.726	-0.894
	(SS-SSIM$^{sss}_{full}$, S-PSO)	0.827	-0.918	0.755	0.713	-0.880
CSIQ	(MS-SSIM$^{(0.1)}_{(\alpha_j=\beta_j=\gamma_j)}$, GA$_2$)	0.787	-0.922	0.862	0.789	-0.908
	(MS-SSIM$^{(0.1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA$_2$)	0.73	-0.936	0.85	0.779	-0.888
	(MS-SSIM$_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA)	0.776	-0.952	0.874	0.791	-0.904
	(MS-SSIM$_{(\alpha_j=\beta_j=\gamma_j)}$, GA)	0.754	-0.929	0.845	0.779	-0.899
	(SS-SSIM$^{sss}_{(\alpha,\ \beta,\ \gamma)}$, GA)	0.766	-0.932	0.802	0.742	-0.895
	(SS-SSIM$^{sss}_{full}$, S-PSO)	0.737	-0.965	0.853	0.771	-0.916
TID2008	(MS-SSIM$^{(0.1)}_{(\alpha_j=\beta_j=\gamma_j)}$, GA$_2$)	0.788	-0.915	0.863	0.792	-0.911
	(MS-SSIM$^{(0.1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA$_2$)	0.784	-0.928	0.87	0.8	-0.900
	(MS-SSIM$_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA)	0.793	-0.948	0.880	0.796	-0.905
	(MS-SSIM$_{(\alpha_j=\beta_j=\gamma_j)}$, GA)	0.792	-0.917	0.859	0.789	-0.901
	(SS-SSIM$^{sss}_{(\alpha,\ \beta,\ \gamma)}$, GA)	0.745	-0.861	0.854	0.784	-0.928
	(SS-SSIM$^{sss}_{full}$, S-PSO)	0.746	-0.953	0.872	0.786	-0.912
TID2013	(MS-SSIM$^{(0.1)}_{(\alpha_j=\beta_j=\gamma_j)}$, GA$_2$)	0.765	-0.912	0.85	0.789	-0.893
	(MS-SSIM$^{(0.1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA$_2$)	0.791	-0.927	0.871	0.805	-0.899
	(MS-SSIM$_{(\alpha_M,\ \beta_j\neq\gamma_j)}$, GA)	0.797	-0.937	0.872	0.801	-0.901
	(MS-SSIM$_{(\alpha_j=\beta_j=\gamma_j)}$, GA)	0.796	-0.910	0.857	0.791	-0.901
	(SS-SSIM$^{oss}_{(\alpha,\ \beta,\ \gamma)}$, GA)	0.735	-0.840	0.851	0.786	-0.935
	(SS-SSIM$^{oss}_{full}$, S-PSO)	0.706	-0.864	0.858	0.792	-0.949
Baseline	$SS-SSIM^{sss}$	0.773	-0.864	0.773	0.741	-0.895
	$SS-SSIM^{oss}$	0.695	-0.772	0.777	0.748	-0.931
	$MS-SSIM^{grayscale}$	0.793	-0.913	0.858	0.790	-0.901

6.5. Parameters distribution

In this sub-section, the reader will be presented with the suggested parameters, along with their discussion. The results will be provided for SS-SSIM$_{full}$, MS-SSIM$_{(\alpha_M,\ \beta_j\neq\gamma_j)}$ and MS-SSIM$^{(0,\ 1)}_{(\alpha_M,\ \beta_j\neq\gamma_j)}$ as these types of search spaces were shown to exhibit the highest correlation with subjective evaluation.

Table 9 shows the set of parameters learned for SS-SSIM$_{full}$ on each IQA database. More specifically, we averaged the parameters at the end of the evolutionary process across all the runs. The results are divided across the two scale selection methods: SSS and OSS. It is worth highlighting that we show the results obtained by GA, as it was found to be slightly better than S-PSO (consult Section 6.2 for more details). See the following two paragraphs for the main findings.

Table 8
SRCC values of SSIM-based IQA metrics for each distortion type in TID2008.

Training IQA DB	TID2008			CSIQ	TID2008		CSIQ
Distortion Type	**SS-SSIM^{sss}**	**SS-SSIM^{oss}**	**SS-SSIM$^{oss}_{full}$**	**SS-SSIM$^{sss}_{full}$**	**MS-SSIM**	**MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$**	**MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$**
Additive Gaussian noise	0,810	0,826	0,864	0,847	0,815	0,811	0,810
Noise in color components	0,804	0,806	0,855	0,860	0,806	0,805	0,806
Spatially correlated noise	0,814	0,841	0,889	0,870	0,823	0,830	0,830
High frequency noise	0,873	0,890	0,906	0,910	0,871	0,878	0,875
Impulse noise	0,674	0,697	0,788	0,805	0,691	0,692	0,704
Quantization noise	0,859	0,864	0,898	0,891	0,866	0,867	0,867
Gaussian blur	0,955	0,956	0,966	0,958	0,957	0,956	0,958
JPEG compression	0,924	0,924	0,936	0,942	0,932	0,930	0,930
JPEG2000 compression	0,963	0,973	0,971	0,962	0,970	0,970	0,970
JPEG transmission	0,867	0,849	0,849	0,857	0,872	0,869	0,866
JPEG2000 transmission	0,858	0,859	0,878	0,887	0,861	0,861	0,859
Mean shift	0,723	0,723	0,735	0,782	0,734	0,737	0,737
Contrast change	0,525	0,525	0,637	0,559	0,638	0,634	0,634
Non eccentricity pattern	0,711	0,722	0,706	0,682	0,740	0,746	0,737
Masked noise	0,780	0,767	0,824	0,861	0,811	0,813	0,809
Image denoising	0,953	0,962	0,967	0,958	0,959	0,957	0,955
Local block-wise distortions	0,845	0,841	0,867	0,832	0,769	0,693	0,719

Table 9
Table of aggregated (by average) parameters learned for SS-SSIM$_{full}$ at each IQA database.

SSIM	IQA DB	α	β	γ	K_1	K_2	ws	stride	dilation
SS-SSIM$^{sss}_{full}$	CSIQ	0.211	0.208	2.506	0.249	0.114	15 × 15	(4, 4)	4
	TID2008	0.123	0.533	2.164	0.234	0.096	19 × 19	(4, 4)	3
	TID2013	0.107	1.237	1.943	0.205	0.093	19 × 19	(4, 4)	2
	CID:IQ	1.136	0.557	1.827	0.207	0.020	15 × 15	(4, 4)	1
	VDID2014	0.703	0.082	2.236	0.116	0.211	15 × 15	(4, 4)	4
SS-SSIM$^{oss}_{full}$	CSIQ	0.020	0.054	2.565	0.211	0.094	19 × 19	(5, 5)	2
	TID2008	0.077	0.457	2.478	0.206	0.075	21 × 21	(4, 4)	2
	TID2013	0.092	0.844	1.531	0.217	0.065	17 × 17	(4, 4)	1
	CID:IQ	0.685	0.220	2.450	0.205	0.023	11 × 11	(4, 4)	1
	VDID2014	1.919	1.837	1.270	0.138	0.228	13 × 13	(4, 4)	3

By analyzing Table 9, one can clearly see that SSIM's optimization suggests unequal importance of its inner components: in general, the structure (γ) is prioritized over contrast (β), whereas contrast is prioritized over the luminance (α). In such a way, we empirically, in a data-driven manner, reinforce the hypothesis that the HVS is highly adapted for extracting structural information from the scenes and prove the inappropriateness of assigning them equal weights in SS-SSIM. A further indication of the relative importance of different components is given by optimizing normalization constants K_1 and K_2. In general, the SSIM optimization suggests setting them respectively at least 20 and 3 times higher than the default values (0.01 and 0.03) (Wang et al., 2004). Specifically, K_1 directly affects constant C_1, which is used to protect luminance Eq. (1) from a division by zero. A higher value in such an equation forces the ratio to be closer to 1: in the final SSIM computation, where all components are combined via multiplication, this behavior makes the luminance component have a lesser impact on the overall similarity score. The most impressive revelation, in our opinion, was related to the sliding-window's size, stride and dilation — all of them are suggested to be significantly larger than the ones proposed in the literature. It happens that, by optimizing the sliding-window size, we allow the system to optimize the SS-SSIM with respect to the visual resolution used during databases' subjective evaluation of the underlying visual data (number of pixels per degree of the visual field). From the experimental results, we can conclude that the default 11 × 11 window size does not happen to fit with IQA databases' experimental settings; the only exception to this observation can be found at CID:IQ, when optimizing SS-SSIM$^{oss}_{full}$. When considering the proposed sliding-window's stride, it becomes clear that SS-SSIM can be used with significantly fewer applications of the filter to an input reference–distortion pair of images, without loss of performance; from the table, the number of filter's movements between convolutional applications, on both the horizontal and vertical axis, can be enlarged about four times. The usage of spatial information and the receptive field was also substantially revised through the dilation rate. The optimization system generally suggests a larger dilation rate, at least two times higher than the default parameter (which is 1), and the value is highly dependent on the IQA database. One of the effects of applying a dilation term larger than 1 is that image distortions are sampled instead of densely analyzed. A preference for smaller dilation suggests the existence of fine-level distortions that cannot be overlooked and should be considered in the overall similarity score. Conversely, a larger dilation might suggest to either actively ignore unimportant fine-level distortions, or that such distortions are not present at all. In practice, the biggest difference can be observed between datasets CID:IQ (optimal dilation term 1), and VDID2014 (optimal dilation term 3/4). Since the two datasets are built with a very similar set of distortions, the different behavior is found in the characteristics of the reference images, which are affected differently by the same distortions.

Finally, we consider it is necessary to highlight the noticeable differences between SSS and OSS methods. First, the dilation rate for the latter happens to be smaller when compared to SSS; in our opinion, this is a natural consequence of the adjustments brought by a more *aware* spatial-scale selection approach in the usage of spatial information and the receptive field. Regarding other parameters, the general distribution can be considered to be comparable between the two scale selection methods.

Table 11 shows the set of parameters learned for MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$ on each IQA database. Similarly to Table 9, we have averaged the parameters obtained by the GA, at the end of the evolutionary process and across all the runs. See the following two paragraphs for the main findings.

Similar to what we observed in Table 9, the measure's parameters associated with the structural component (γ_j) are, in general terms, suggested to be larger than those associated with contrast β_j and luminance α_5. This fact serves as another supportive argument to favor the theory that HVS is highly adapted for extracting structural information from

Table 10
Table of aggregated (by average) parameters learned for MS-SSIM$_{(\alpha_M,\ \beta_j \neq \gamma_j)}$ at each IQA database.

IQA DB	β_1	β_2	β_3	β_4	β_5	γ_1	γ_2	γ_3	γ_4	γ_5	α_5
CSIQ	0.248	0.479	0.315	0.228	0.097	0.028	1.769	1.959	1.857	1.727	0.309
TID2008	0.031	0.729	0.349	0.105	0.098	0.091	0.891	1.314	1.542	1.477	0.343
TID2013	0.323	1.187	0.398	0.058	0.085	0.110	0.539	1.337	1.485	1.496	0.499
CID:IQ	0.248	1.104	0.292	0.202	0.326	0.022	1.408	2.568	1.851	0.308	2.172
VDID2014	0.038	0.492	1.517	1.444	0.691	0.027	0.073	1.539	1.998	1.264	0.718

the scenes and proves the inappropriateness of assigning equal weights to different components at different scales in MS-SSIM. Moreover, one can observe that database-wise, the largest weights for contrast (β_j) can be observed at scale $j = 2$, except for VDID2014 where the largest values can be observed for $j \in \{3,\ 4\}$; the largest contrast magnitudes (γ_j) can be seen at deeper scales: $j \in \{3,\ 4,\ 5\}$.

Table 11 shows the set of parameters learned for MS-SSIM$^{(0,\ 1)}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$ on each IQA database. Similarly to Table 9, we averaged the parameters at the end of the evolutionary process, across all the runs. See the following two paragraphs for the main findings. The column *Scale* uniquely identifies the five spatial-scales embedded into MS-SSIM. The column *Prop* represents, at the jth scale, the proportion of runs where the GA's final solution (proposed set of MS-SSIM's parameters) included the jth spatial scale in the calculations; notice that, when building Table 7, MS-SSIM$^{(0,\ 1)}_{(\alpha_j=\beta_j=\gamma_j)}$ and MS-SSIM$^{(0,\ 1)}_{(\alpha_M,\ \beta_j \neq \gamma_j)}$ did not include the jth spatial-scale for the *Prop* values smaller than 0.5. The naming nomenclature of the remaining columns was already presented in Tables 9 and 10.

Similarly to what we observed in Tables 9 and 10, the measure's parameters associated with structural component (γ_j) are, in general terms, suggested to be larger than those associated to contrast (β_j) and luminance α_M — which, once again, supports the aforementioned argument to favor the theory that HVS is highly adapted for extracting structural information from the scenes and proves the inappropriateness of assigning equal weights to different components at different scales in MS-SSIM. Differently from the findings of Table 9, the suggested values for the parameters K_1, K_2, sliding-window's size, stride, and dilation seem not deviate significantly from the baseline settings (except VID2014). We suspect this happen for two reasons. First, it might happen that the suggested parameters are, in fact, optimal in the context of MS-SSIM$^{(0,1)}_{full_{(\alpha_M,\ \beta_j \neq \gamma_j)}}$ search space optimization. Second, the reason might be the unappropriated algorithmic parameterization for MS-SSIM$^{(0,1)}_{full_{(\alpha_M,\ \beta_j \neq \gamma_j)}}$: given the fact that the population size is equal to 25, and the optimization algorithm is seeded the default parameter set in the initial population, the elite rapidly dominates the population, and the evolutionary process becomes limited to it.

7. Conclusions

Numerous computational systems that remove undesirable visual artifacts rely upon full-reference image quality assessment measures (FR-IQAMs). High-quality and computationally simple FR-IQAMs are in high demand, and the Structural Similarity Index Measure (SSIM) is among the most utilized. In this paper, we revised the original parameters of SSIM through a data-driven framework with the objective of increasing its similarity, measured through Spearman's Rank Correlation Coefficient (SRCC), with the subjective evaluation provided by human observers, such as the mean opinion score (MOS) and differential MOS (DMOS). The inclusion of viewing conditions in IQA is one of the central points in this paper. For this reason, we paid particular attention to other research branches in addition to MS-SSIM, which led us to confront the so-called standard scale selection (SSS), proposed by Wang et al. and the optimal scale selection (OSS), proposed by Gu et al. We exploited evolutionary computation and swarm intelligence metaheuristics on five popular IQA databases, including two dedicated distance-changed databases, to efficiently define the best combination of parameters for the application of SS-SSIM and MS-SSIM. The empirical results show that proper parameter settings allow to improve both SS-SSIM and MS-SSIM significantly in terms of correlation with human perception of visual quality; moreover, we prove that the set of optimal parameters learned on a given IQA database can be successfully transferred to other databases, different and previously unseen during the training, including distance-changed databases.

Among the original motivations of this study was the intention to challenge a set of assumptions and implications commonly accepted in image quality assessment. The first of such assumptions is the equally-weighted importance of luminance, structure, and contrast similarities. Our data-driven method suggests to prioritizing structure over contrast, these being the most critical components of SS-SSIM, whereas luminance emerges as the least important by correlation with responses from the human visual system. Moreover, we prove that the conventional values of normalization constants K1 and K2, the sliding-window's size, stride, and dilation factor have to be revised – our optimizations generally suggest significantly larger values – and better adjusted with respect to the viewing conditions used during databases subjective evaluation of the underlying visual data. Another observation is related to the role of scale selection, and its relationship to viewing distance: by comparing SS-SSIM with its multi-scale counterpart, we show that proper fine-tuning of SS-SSIM can be as good or even better than MS-SSIM (even when the latter is also fine-tuned). By comparing the results obtained from optimizing SS-SSIM after SSS with those after OSS, we concluded that the advantage of applying the more complex OSS is not granted; although, in general terms, the latter exhibits higher generalization ability, such a difference is not statistically substantiated. Finally, by allowing the estimation of the relative importance of each similarity component at each scale, instead of assuming their equality as is done in the literature, the measure achieves better performance.

With this work, we have proposed and interpreted a new set of reference parameters for SSIM variants. These parameters can be effortlessly embedded and exploited in any existing implementation of SSIM without additional overhead in terms of computational resources. The practical applications include all fields where a proxy for human judgment is needed to either evaluate a solution or provide feedback during its training. In particular, it is worth noting that both the original SSIM and our proposed parametrizations are entirely differentiable, and as such, can be used as loss functions for backpropagation-based training of convolutional neural networks.

In terms of the exploited optimization metaheuristics – Genetic Algorithms and Synchronous Particle Swarm Optimization – we found that, in general terms, they exhibit comparable behaviors; such a level of agreement between two conceptually distinct techniques suggests a fair level of convergence on the underlying optimization problems and gives us more confidence about the precision of our results. Nonetheless, we might consider the exploitation of alternative metaheuristics in the future, such as Differential Evolution (Storn & Price, 1995) and Salp Swarm Algorithm (Mirjalili et al., 2017), in order to either consolidate the findings that emerged from our current analysis or potentially discover unexplored areas of the parameters search space.

The work presented in this paper solves the dependence on viewing distance by resorting to image downscaling, as it is commonly done by the approaches of MS-SSIM, OSS-SSIM, and SSS-SSIM. Although proven effective in terms of MOS correlation, the resulting similarity measure,

Table 11
Table of aggregated (by average) parameters learned for $\text{MS-SSIM}^{(0,1)}_{full_{(\alpha_M,\ \beta_j \neq \gamma_j)}}$ at each IQA database.

IQA DB	Scale	Prop	β_j	γ_j	α_M	K_1	K_2	ws	stride	dilation
CSIQ	1	0.5	0.301	0.630	0.247	0.053	0.059	11 × 11	(2, 2)	1
	2	1.0	0.318	0.823		0.037	0.043	11 × 11	(1, 1)	1
	3	1.0	0.363	0.585		0.041	0.028	11 × 11	(1, 1)	1
	4	0.0	0.420	0.635		0.042	0.047	11 × 11	(1, 1)	1
	5	1.0	0.062	1.117		0.029	0.046	11 × 11	(1, 1)	1
TID2008	1	1.0	0.222	0.120	0.142	0.038	0.045	11 × 11	(1, 1)	1
	2	1.0	0.393	0.374		0.048	0.037	11 × 11	(1, 1)	1
	3	1.0	0.301	0.388		0.031	0.029	11 × 11	(1, 1)	1
	4	1.0	0.202	0.498		0.046	0.035	11 × 11	(1, 1)	1
	5	0.5	0.140	0.467		0.025	0.033	11 × 11	(1, 1)	1
TID2013	1	1.0	0.548	0.222	0.279	0.025	0.068	11 × 11	(1, 1)	1
	2	1.0	0.432	0.379		0.038	0.051	11 × 11	(1, 1)	1
	3	1.0	0.300	0.429		0.049	0.02	11 × 11	(1, 1)	1
	4	1.0	0.166	0.332		0.029	0.029	11 × 11	(1, 1)	1
	5	1.0	0.201	0.826		0.03	0.033	11 × 11	(2, 2)	1
CIDIQ	1	0.0	0.395	0.199	0.316	0.045	0.032	11 × 11	(1, 1)	1
	2	1.0	0.300	0.332		0.038	0.03	11 × 11	(1, 1)	1
	3	1.0	0.276	0.305		0.042	0.036	11 × 11	(1, 1)	1
	4	1.0	0.254	0.286		0.024	0.014	11 × 11	(1, 1)	1
	5	1.0	0.157	0.140		0.02	0.03	11 × 11	(1, 1)	1
VDID2014	1	0.0	0.309	0.786	0.653	0.126	0.06	13 × 13	(2, 2)	2
	2	0.0	0.671	0.660		0.08	0.073	11 × 11	(3, 3)	1
	3	1.0	0.609	1.419		0.091	0.138	13 × 13	(2, 2)	1
	4	1.0	0.833	1.371		0.074	0.075	15 × 15	(2, 2)	2
	5	1.0	0.410	0.657		0.07	0.108	11 × 11	(3, 3)	2

does not explicitly handle the viewing distance, relying instead upon data-driven statistical optimization. It would be interesting to consider the inclusion of the viewing distance as an additional input to the similarity measure formulation. To this extent, the main limitation is currently the scarcity of appropriate MOS-annotated datasets with multiple viewing distances, which is why an investigation in merging existing datasets could be considered for future developments. Additionally, the ever-increasing success of deep learning-based solutions in computer vision suggests the possibility of extending our parameter optimization methodology to the combination of higher-abstraction similarity components derived from existing solutions for convolutional neural networks for image-quality assessment.

CRediT authorship contribution statement

Illya Bakurov: Conceptualization, Methodology, Development, Software, Programming, Validation, Verification, Formal analysis, Writing - original draft, Writing - review & editing. **Marco Buzzelli:** Conceptualization, Methodology, Development, Software, Programming, Validation, Verification, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Raimondo Schettini:** Conceptualization, Validation, Verification, Formal analysis, Investigation, Writing - review & editing, Supervision. **Mauro Castelli:** Conceptualization, Validation, Verification, Formal analysis, Investigation, Writing - review & editing, Supervision, Funding acquisition. **Leonardo Vanneschi:** Conceptualization, Validation, Verification, Formal analysis, Investigation, Writing - review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by national funds through the FCT (Fundação para a Ciência e a Tecnologia), Portugal by the projects GADgET (DSAIPA/DS/0022/2018), BINDER (PTDC/CCIINF/29168/2017), and AICE (DSAIPA/DS/0113/2019). Mauro Castelli acknowledges the financial support from the Slovenian Research Agency, Slovenia (research core funding no. P5-0410).

References

Bakurov, I., Buzzelli, M., Castelli, M., Vanneschi, L., & Schettini, R. (2020). Parameters optimization of the structural similarity index. In *London imaging meeting 2020: future colour imaging* (pp. 19–23). http://dx.doi.org/10.2352/issn.2694-118X.2020.LIM-13.

Bakurov, I., Buzzelli, M., Castelli, M., Vanneschi, L., & Schettini, R. (2021). General purpose optimization library (GPOL): A flexible and efficient multi-purpose optimization library in python. *Applied Sciences*, *11*(11), http://dx.doi.org/10.3390/app11114774, URL https://www.mdpi.com/2076-3417/11/11/4774.

Bartashevich, P., Grimaldi, L., & Mostaghim, S. (2017). Pso-based search mechanism in dynamic environments: Swarms in vector fields. In *2017 IEEE congress on evolutionary computation (CEC)* (pp. 1263–1270). http://dx.doi.org/10.1109/CEC.2017.7969450.

Bartashevich, P., Koerte, D., & Mostaghim, S. (2020). Impact of communication topology on PSO-based swarms in vector fields. In *2020 IEEE symposium series on computational intelligence (SSCI)* (pp. 497–504). http://dx.doi.org/10.1109/SSCI47803.2020.9308526.

Bianco, S., Celona, L., & Napoletano, P. (2021). Disentangling image distortions in deep feature space. *Pattern Recognition Letters*, *148*, 128–135. http://dx.doi.org/10.1016/j.patrec.2021.05.008.

Bianco, S., Celona, L., & Piccoli, F. (2020). Single image dehazing by predicting atmospheric scattering parameters. *2020*, (pp. 74–77). http://dx.doi.org/10.2352/issn.2694-118X.2020.LIM-11,

Charrier, C., Knoblauch, K., Maloney, L. T., Bovik, A. C., & Moorthy, A. K. (2012). Optimizing multiscale SSIM for compression via MLDS. *IEEE Transactions on Image Processing*, *21*(12), 4682–4694.

Du, H., Wang, Z., Zhan, W., & Guo, J. (2018). Elitism and distance strategy for selection of evolutionary algorithms. *IEEE Access*, *6*, 44531–44541. http://dx.doi.org/10.1109/ACCESS.2018.2861760.

Franzen, R. (1999). Kodak lossless true color image suite. URL http://r0k.us/graphics/kodak/ Last accessed: 04.10.2019.

Gu, K., Liu, M., Zhai, G., Yang, X., & Zhang, W. (2015). Quality assessment considering viewing distance and image resolution. *IEEE Transactions on Broadcasting*, *61*(3), 520–531.

Gu, K., Zhai, G., Liu, M., Xu, Q., Yang, X., Zhou, J., & Zhang, W. (2013). Adaptive high-frequency clipping for improved image quality assessment. In *2013 visual communications and image processing (VCIP)* (pp. 1–5).

Gu, K., Zhai, G., Yang, X., & Zhang, W. (2013). Self-adaptive scale transform for IQA metric. In *2013 IEEE international symposium on circuits and systems (ISCAS)* (pp. 2365–2368).

Hamzaoui, Y., & Arellano, J. (2018). Comparison of particle swarm optimization and genetic algorithm for multiproduct batch plant design of protein production. *Journal of Analytical and Pharmaceutical Research*, *7*, 553–563. http://dx.doi.org/10.15406/japlr.2018.07.00282.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence*. Cambridge, MA, USA: MIT Press.

Jatana, N., & Suri, B. (2020). Particle swarm and genetic algorithm applied to mutation testing for test data generation: A comparative evaluation. *Journal of King Saud University - Computer and Information Sciences*, *32*(4), 514–521. http://dx.doi.org/10.1016/j.jksuci.2019.05.004, URL https://www.sciencedirect.com/science/article/pii/S1319157819301466, Emerging Software Systems.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Vol. 4*, In *Proceedings of ICNN'95 - international conference on neural networks* (pp. 1942–1948). http://dx.doi.org/10.1109/ICNN.1995.488968.

Kuo, T.-Y., Su, P.-C., & Tsai, C.-M. (2016). Improved visual information fidelity based on sensitivity characteristics of digital images. *Journal of Visual Communication and Image Representation*, *40*, 76–84. http://dx.doi.org/10.1016/j.jvcir.2016.06.010, URL http://www.sciencedirect.com/science/article/pii/S104732031630102X.

Larson, E., & Chandler, D. (2010). Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, *19*, Article 011006. http://dx.doi.org/10.1117/1.3267105.

Lin, W., & Kuo, C.-C. J. (2011). Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, *22*, 297–312.

Liu, X., Pedersen, M., & Hardeberg, J. Y. (2014). Cid:IQ – a new image quality database. In A. Elmoataz, O. Lezoray, F. Nouboud, & D. Mammass (Eds.), *Image and signal processing* (pp. 193–202). Cham: Springer International Publishing.

Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., & Mirjalili, S. M. (2017). Salp swarm algorithm: A bio-inspired optimizer for engineering design problems. *Advances in Engineering Software*, *114*, 163–191.

Mitchell, M. (1998). *An introduction to genetic algorithms*. Cambridge, MA, USA: MIT Press.

Moraglio, A., & Poli, R. (2004). Topological interpretation of crossover. http://dx.doi.org/10.1007/978-3-540-24854-5_131,

Ponomarenko, N., Battisti, F., Egiazarian, K., Astola, J., & Lukin, V. (2009). Metrics performance comparison for color image database. In *Proceedings of the 4th international workshop on video processing and quality metrics* (pp. 1–6).

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., & Kuo, C.-C. J. (2015). Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, *30*, 57–77. http://dx.doi.org/10.1016/j.image.2014.10.009.

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., & Battisti, F. (2009). Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, *10*(4), 30–45.

Qian, R., Tan, R., Yang, W., Su, J., & Liu, J. (2018). *Attentive generative adversarial network for raindrop removal from A single image* (pp. 2482–2491). http://dx.doi.org/10.1109/CVPR.2018.00263.

Rhodes, A. D. (2019). Evolving order and chaos: Comparing particle swarm optimization and genetic algorithms for global coordination of cellular automata. CoRR http://arxiv.org/abs/1909.03560.

Rundo, L., Tangherloni, A., Cazzaniga, P., Nobile, M. S., Russo, G., Gilardi, M. C., Vitabile, S., Mauri, G., Besozzi, D., & Militello, C. (2019). A novel framework for MR image segmentation and quantification by using medga. *Computer Methods and Programs in Biomedicine*, *176*, 159–172. http://dx.doi.org/10.1016/j.cmpb.2019.04.016, URL http://www.sciencedirect.com/science/article/pii/S0169260718317565.

Shao, Y., Li, L., Ren, W., Gao, C., & Sang, N. (2020). Domain adaptation for image dehazing. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2805–2814). http://dx.doi.org/10.1109/CVPR42600.2020.00288.

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

Silvestre-Blanes, J. (2011). Structural similarity image quality reliability: Determining parameters and window size. *Signal Processing*, *91*(4), 1012–1020.

Skurowski, P., & Janiak, M. (2014). Component weight tuning of SSIM image quality assessment measure. In *International conference on computer vision and graphics* (pp. 57–65). Springer.

Smith, C. (2013). Facebook users are uploading 350 million new photos each day. URL https://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9, Last accessed: 04.10.2019.

Storn, R., & Price, K. (1995). Differential evolution: A simple and efficient adaptive scheme for global optimization over continuous spaces. *Technical Report TR-95-012, ICSI*, (p. 23).

Streijl, R., Winkler, S., & Hands, D. (2016). Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, *22*, 213–227. http://dx.doi.org/10.1007/s00530-014-0446-1.

Toizumi, T., Zini, S., Sagi, K., Kaneko, E., Tsukada, M., & Schettini, R. (2019). *Artifact-free thin cloud removal using gans* (pp. 3596–3600). http://dx.doi.org/10.1109/ICIP.2019.8803652.

Venkataramanan, A. K., Wu, C., Bovik, A., Katsavounidis, I., & Shahid, Z. (2021). A hitchhiker's guide to structural similarity. *IEEE Access*, *9*, 28872–28896.

Wang, Z., & Bovik, A. C. (2006). Modern image quality assessment. In *Modern image quality assessment*.

Wang, Z., & Bovik, A. (2009). Mean squared error: love it or leave it? - a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, *26*, 98–117. http://dx.doi.org/10.1109/MSP.2008.930649.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, C., Fan, W., Wu, Y., & Su, Z. (2020). Weakly supervised single image dehazing. *Journal of Visual Communication and Image Representation*, *72*, Article 102897. http://dx.doi.org/10.1016/j.jvcir.2020.102897, URL https://www.sciencedirect.com/science/article/pii/S1047320320301395.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multi-scale structural similarity for image quality assessment.

Wihartiko, F. D., Wijayanti, H., & Virgantari, F. (2018). Performance comparison of genetic algorithms and particle swarm optimization for model integer programming bus timetabling problem. *IOP Conference Series: Materials Science and Engineering*, *332*, Article 012020. http://dx.doi.org/10.1088/1757-899x/332/1/012020, URL https://doi.org/10.1088/1757-899x/332/1/012020.

Wolpert, D., & Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. http://dx.doi.org/10.1109/4235.585893.

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. CoRR http://arxiv.org/abs/1511.07122.

Zhao, J., Xiong, R., Xu, J., & Huang, T. (2019). *Learning a deep convolutional network for subband image denoising* (pp. 1420–1425). http://dx.doi.org/10.1109/ICME.2019.00246.

Zini, S., Bianco, S., & Schettini, R. (2020). Deep residual autoencoder for blind universal JPEG restoration. *IEEE Access*, *8*, 63283–63294.