# Genetic Programming for Structural Similarity Design at Multiple Spatial Scales

Illya Bakurov
ibakurov@novaims.unl.pt
Universidade Nova de Lisboa
Lisbon, Portugal

Marco Buzzelli
marco.buzzelli@unimib.it
University of Milano - Bicocca
Milan, Italy

Mauro Castelli
mcastelli@novaims.unl.pt
Universidade Nova de Lisboa
Lisbon, Portugal

Raimondo Schettini
raimondo.schettini@unimib.it
University of Milano - Bicocca
Milan, Italy

Leonardo Vanneschi
lvanneschi@novaims.unl.pt
Universidade Nova de Lisboa
Lisbon, Portugal

## ABSTRACT

The growing production of digital content and its dissemination across the worldwide web require efficient and precise management. In this context, image quality assessment measures (IQAMs) play a pivotal role in guiding the development of numerous image processing systems for compression, enhancement, and restoration. The structural similarity index (SSIM) is one of the most common IQAMs for estimating the similarity between a pristine reference image and its corrupted variant. The multi-scale SSIM is one of its most popular variants that allows assessing image quality at multiple spatial scales. This paper proposes a two-stage genetic programming (GP) approach to evolve novel multi-scale IQAMs, that are simultaneously more effective and efficient. We use GP to perform feature selection in the first stage, while the second stage generates the final solutions. The experimental results show that the proposed approach outperforms the existing MS-SSIM. A comprehensive analysis of the feature selection indicates that, for extracting multi-scale similarities, spatially-varying convolutions are more effective than dilated convolutions. Moreover, we provide evidence that the IQAMs learned for one database can be successfully transferred to previously unseen databases. We conclude the paper by presenting a set of evolved multi-scale IQAMs and providing their interpretation.

## CCS CONCEPTS

• **Computing methodologies** → **Matching**.

## KEYWORDS

Genetic Programming, Image Quality Assessment, Structural Similarity, Multi-Scale Structural Similarity Index, Dilated Convolutions, Spatially-Varying Kernels, Multi-Scale Context, Multi-Scale Processing, Evolutionary Computation, Image Processing


*GECCO '22, July 9–13, 2022, Boston, MA, USA*

ACM ISBN 978-1-4503-9237-2/22/07…$15.00
https://doi.org/10.1145/3512290.3528783

**ACM Reference Format:**
Illya Bakurov, Marco Buzzelli, Mauro Castelli, Raimondo Schettini, and Leonardo Vanneschi. 2022. Genetic Programming for Structural Similarity Design at Multiple Spatial Scales. In *Genetic and Evolutionary Computation Conference (GECCO '22), July 9–13, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512290.3528783

## 1 INTRODUCTION

The dynamic rise of social media and its penetration into our daily lives causes a continuously growing production of digital content and its dissemination across the world wide web. About 58% of the world's population uses social media, and the average daily usage is about two and a half hours [1]. As of 2021, there are over 1 billion active Instagrammers, half of whom share their life moments daily through the *stories* feature [2]. Digital images and video happen to emerge as the most favored medium for communication and information sharing: recent data suggest that almost 70% of the users rated video as their number one source of information, and 75% of all video views come from mobile devices [3]. These facts are not surprising - taking a photo, recording a video, or performing a live stream to any audience, no matter the scale, is now a matter of a few clicks. Forbes estimates that online videos will make up more than 82% of all consumer internet traffic by 2022 [4]. Therefore, efficient and precise methods for digital imagery management are in high demand. In this context, compression algorithms play an essential role in saving storage space and bandwidth for transferring large amounts of imagery and video information (from now on, called *media*) through the Internet. However, the compression usually comes at the cost of visual quality degradation. For example, compression artifacts produce an unpleasant viewing experience and can deteriorate computer-vision (CV) systems' performance. To counteract this problem, the image processing (IP) community has developed numerous techniques to remove the undesirable artifacts from images that degrade the visual quality and reduce their usefulness for the underlying tasks [5, 16, 43, 45].

Both input and output media need to be evaluated for the perceived visual quality to design compression and restoration algorithms properly. Moreover, such evaluation must be ideally automatic (i.e., without human intervention), precise and efficient. The research community proposed several image quality assessment measures (IQAMs) to solve this problem [12, 17, 19, 20, 25, 27, 35–38, 42], notable among which is the structural similarity index

(SSIM) inspired by the intrinsic functioning of human visual system (HVS). Notice that IQAMs' performance is usually assessed by correlating their similarity scores with those provided by human observers; to this extent, Spearman's rank correlation coefficient (SRCC) is preferred [36]. Due to its precision, computational efficiency, and mathematical formulation's interpretability, the SSIM considerable large attention in the community and motivated several related measures [20, 37, 38, 42]; a detailed review of SSIM can be found in [33]. The multi-scale SSIM (MS-SSIM) is one of the most popular and commonly used among the existing measures. The MS-SSIM consists of aggregating the similarity components at different spatial scales (a.k.a. multi-scale context), the latter being simulated through a pooling operator. From another perspective, MS-SSIM can be seen as a *compact* deep convolutional neural network (DCNN) made of 5 sequential blocks of convolutional and pooling layers. However, recent empirical evidence in the CV field indicates that pooling layers can cause a significant loss of potentially relevant spatial information when performing multi-scale context aggregation. This produces a degradation of the systems' performance on different tasks [23, 24, 40] and can affect image quality assessment (IQA). In this study, genetic programming (GP) [21] is proposed to generate a novel MS-SSIM. Such an MS-SSIM should be more precise while maintaining the computational efficiency and the simplicity of mathematical formulation. Moreover, motivated by the up-to-date findings in CV about the harmful effects of pooling layers and to simplify the measure, we redesign the transition mechanism between different spatial scales by replacing the pooling layers with dilated convolutions and spatially-varying convolutional kernels. Particularly, in this work, we show that:

- GP can be successfully applied to evolve SSIM-like measures;
- the evolved measures can improve the correlation with the subjective evaluation provided by human observers;
- the evolved measures can be generalized to other problems;
- the evolved measures can achieve a lower computational complexity when compared to the original MS-SSIM;
- the mathematical formulation of the evolved measures allows for human-interpretability;
- the most effective approach to aggregate multi-scale spatial information is through spatially-varying convolutional kernels.

The paper is organized as follows: Section 2 introduces the necessary theoretical background by providing an overview of IQA, SSIM and the different approaches to aggregate multi-scale contextual information. Section 3 describes how new IQAMs based on SSIM can be formulated using GP; in particular, the considered terminal and function sets are provided and explained. Section 4 presents the research objectives, characterizes the datasets used in our study, broaches and discusses the hyper-parameters used, and shows the results obtained. Finally, Section 5 draws the main conclusions and proposes future research ideas.

## 2 BACKGROUND

### 2.1 Image quality assessment (IQA)

For those applications in which humans ultimately consume the media, the most appropriate method for media quality assessment is through human visual system (HVS). In other words, by involving people to evaluate the perceived quality of the media subjectively. However, in practice, subjective evaluation happens to be complex, time-consuming, expensive, and sensitive to the experimental design [29]. To overcome these limitations, several researchers have proposed objective IQAMs that can automatically (i.e., without human intervention) estimate the perceived visual quality.

Typically, FR-IQAMs (like SSIM) estimate a standardized similarity score given a pair of reference-distortion images. The precision of a given FR-IQAM is defined as the correlation between these scores and the subjective evaluation provided by human observers - such as the mean opinion score (MOS) and differential MOS (DMOS) - on different IQA databases. A larger correlation with the subjective evaluation is desirable [29]. Additionally, other concerns, like measures' complexity, must be considered.

*2.1.1 Single-scale structural similarity index (SS-SSIM).* The structural similarity index (SSIM) [36] is by far the most popular. It is classified as a *full-reference* IQAM (FR-IQAM) since it directly compares a pristine reference image with its potentially corrupted variant (like an image subject to JPEG compression). The SSIM was inspired by the theory that HVS is particularly suited for extracting structural information from the scenes. By explicitly including HVS's characteristics, the authors were able to introduce a whole new paradigm in the IQAMs field. The high precision, allied to a simple mathematical formulation, assured SSIM's preeminence as a proxy evaluation for human assessment in different IP and CV applications [10, 11, 34, 44].

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 similarity components extracted at a single spatial scale (resolution): luminance, contrast, and structure. Components' extraction is performed by sliding an 11x11 symmetric Gaussian kernel with a standard deviation of 1.5 across reference-distortion pairs. The SSIM is computed as an aggregation of locally estimated components. Each image's patch average $\mu$ 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): $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$, $C_3 = C_2/2$. Contrast is represented by using each image's patch standard deviation $\sigma$. 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:

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

Finally, the three components are combined into a unique expression that is weighted with exponents $\alpha$, $\beta$, and $\gamma$:

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

*2.1.2 Multi-scale structural similarity index (MS-SSIM).* Initially, the SSIM was mainly used at a single spatial scale (SS-SSIM). In practice, however, subjective evaluation is highly dependent on the numerous viewing conditions. These include the environment illumination, the conditions of the displaying device, such as display's resolution and the response time, the distance from the display to the observer, network bandwidth and latency, etc. The single-scale approach may only be appropriate for specific settings in this context. Notably, SSIM, like several other IQAMs, is significantly sensitive to the spatial scale selection [18, 38]. To incorporate viewing conditions' diversity, Wang et al. [38] proposed a multi-scale extension of SSIM (MS-SSIM). MS-SSIM aggregates the inner similarity indexes calculated from a range of 5 different spatial scales (resolutions). To transit between spatial scales, MS-SSIM down-samples the reference-distortion pairs through pooling. Empirical findings constantly show that MS-SSIM tends to outperform its single-scale counterpart on several tasks by a significant margin. Formally MS-SSIM is defined as:

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

By 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 the last scale (i.e., $M = 5$). The exponents $\alpha_M$, $\beta_j$, and $\gamma_j$ are used to adjust the relative importance of different components. The overall cross-scale evaluation is then given by a weighted product of the above-mentioned components extracted at different scales.

For a reader familiar with the field of deep learning (DL), it becomes clear that MS-SSIM is, in practical terms, a *compact* human-interpretable DCNN made of 5 sequential blocks of convolutional and average pooling layers. This observation motivated us to bring state-of-the-art findings in the CV field to design a novel MS-SSIM. Section 2.2 provides the necessary background for that.

## 2.2 Multi-scale spatial information

Current deep learning (DL) systems are made of sequential stacks of similar blocks. These typically comprise convolution, normalization, and activation layers, interleaved with pooling layers that perform spatial down-sampling of the input feature maps [22, 28]. The latter are necessary to (i) reduce the feature maps' spatial dimensionality in such a way enabling a significant improvement in terms of systems' computational efficiency, and (ii) integrate the multi-scale contextual information by enlarging systems' receptive field, therefore embracing greater global context in the scene. However, recent evidence suggests that pooling can cause a significant loss of potentially relevant spatial information [23, 24, 39, 40], therefore degenerating systems' performance on the underlying CV tasks. This facet was found to be particularly harmful to classifying natural images, which tend to exhibit many objects whose identities and relative configurations are important for understanding the scene.

Dilated (a.k.a. atrous) convolutions can be employed to avoid the destructive effects of pooling while maintaining the ability to aggregate multi-scale contextual information without losing resolution. In simple terms, a dilated convolution is a traditional convolution except the filter's resolution is increased by inserting zeros between two successive values along each spatial dimension. It is a simple yet powerful technique to make filters' receptive field larger without impacting computation or the number of parameters. Several studies point out that dilated convolutions outperform their non-dilated counterparts on several complex CV tasks without increasing the model's depth or complexity [13, 14, 39, 40].

Another approach to aggregating the multi-scale contextual information without losing resolution involves processing the input feature map using convolution kernels with different spatial dimensions [15, 41]. The experimental evidence shows that such an approach to learning features, called PyConv, has the potential to impact nearly every CV task. In particular, Duta et al. [15] demonstrated that PyConv significantly improve image classification, video action classification/recognition, object detection, and semantic image segmentation/parsing.

# 3 THE PROPOSED APPROACH

In this study, GP is used to evolve novel multi-scale IQAM, which uses dilated convolutions and spatially varying kernels to aggregate multi-scale similarities. Specifically, we propose to evolve GP individuals using a set of terminal symbols made of aggregated inner similarity indexes, extracted at different spatial scales through dilated convolutions and spatially-varying kernels (as motivated and described in Section 2.2). In this way, the GP individuals are intended to represent potentially new IQAMs that aggregate multi-scale contextual information without losing input images' resolution. Ideally, these new measures will better reflect people's subjective evaluation and, when compared with the traditional MS-SSIM, they will not increase the overall computational complexity. In general terms, the proposed approach is divided into two sequential stages. First, we use GP to estimate the subset of the most prominent spatial scales (i.e., features). Second, we redesign the MS-SSIM by means of GP using previously estimated spatial scales.

## 3.1 Terminals

In the proposed approach, GP individuals aggregate inner similarity indexes extracted at different spatial scales, these being simulated through different dilation rates and window sizes. For the first stage, a large terminal set $T_{stage1}$ is generated to perform feature selection. The pseudo-code in 1 illustrates how $T_{stage1}$ was obtained. Each individual terminal in $T_{stage1}$ is a vector which size equals the number of reference-distortion pairs of a given IQA database, and each value represents a given spatially-aggregated similarity component, obtained using a $w_i \times w_i$ Gaussian kernel, with dilation rate

of $d_j$, for a given reference-distortion pair. From the pseudo-code, we can see that three spatially-aggregated similarity components are computed for the combination of dilation and window-size. Therefore, the cardinality of the terminal set $T_{stage1}$ is given by $\#T_{stage1} = \#F_{SSIM} \times \#(W) \times \#(D) = 3 \times 6 \times 9 = 162$. Although, in DL literature, the alternative multi-scale context aggregation strategies has been explored independently (i.e., with dilated convolutions or with spatially-varying kernels), we decided to generate terminals that blend these two approaches. The rationale behind this decision is to leave GP deciding, in a data-driven manner, which form of multi-scale context aggregation is preferred given the optimization task at hand. We use the symbol $T_{stage2}$ to distinguish the subset of $T_{stage1}$ which is used in the second stage of the proposed GP approach.

**Algorithm 1** GP terminal set composition

**Require:**
  $(X,\ Y)$ ▷ reference-distortion in a given IQA-DB
  $G(w,\ d)$ ▷ $w \times w$ kernel with dilation of $d$
  $F_{SSIM} \leftarrow [l,\ c,\ s]$ ▷ SSIM's similarity components
**Ensure:**
  $T_{stage1} \leftarrow []$ ▷ GP terminals
  $W \leftarrow [3, 5, 7, 9, 11, 13]$ ▷ window-sizes
  $D \leftarrow [1, 2, 3, 4, 5, 6, 7, 8, 9]$ ▷ dilation rates
  **for** $w_i \in W$ **do**
    **for** $d_j \in D$ **do**
      $G' = G(w_i,\ d_j)$ ▷ $w_i \times w_i$ kernel with dilation of $d_j$
      **for** $f_{ssim} \in F_{SSIM}$ **do**
        $t = f_{ssim}^{G'}(X, Y)$ ▷ similarity component using $G'$
        $t = t.mean((-2, -1))$ ▷ spatially-aggregates $t'$
        $T_{stage1}$.append$(t)$
      **end for**
    **end for**
  **end for**

## 3.2 Functions

The original formulations of SSIM and MS-SSIM can be subject to a probabilistic interpretation. In such a view, the overall similarity can be interpreted as the probability that the image pair is simultaneously similar according to three different "sub-similarities" (luminance-, structural-, and contrast- based). By assuming these similarities are independent, the overall probability is, in practice, computed through the multiplication rule. Furthermore, each sub-similarity probability is first reprocessed through an exponential function that modifies its impact: for example, given a base in the 0 to 1 range, a very low exponent will yield a sub-similarity probability that approaches 1, regardless of the starting value. As a consequence of the value being close to 1, its impact on the overall similarity is significantly reduced. In this work, we have leveraged and expanded upon such a concept by including the addition operation among the set of functions that a learned similarity expression can exploit. In the aforementioned probabilistic interpretation, using the addition means that the overall similarity can be seen as the probability that the image pair is similar according to either different similarities. In practice, the set of functions $F$ includes two arithmetic operators ({+, x}), and an exponentiation operator that raises an input terminal to the power of a given exponent ($x^i$, $i \in$ [0.05, 0.15, 0.3, 0.5, 0.8, 1/0.8, 1/0.5, 1/0.3, 1/0.15, 1/0.05]).

## 3.3 Limit complexity without performance deterioration

An extensive evolutionary search inevitably implies solutions' growth in terms of complexity and can cause overfitting [30, 32]. We use EDDA [9, 31] to seed the GP population in this context. In a nutshell, the initial population is created by using the best individuals extracted from a set of independent sub-populations (a.k.a. demes), which run under distinct evolutionary conditions (i.e., parameters). Demes are left to evolve for a few generations, such that the elite solutions do not grow significantly in size when they are extracted. The abundance of numerous independent sub-populations allows performing a broad exploration of the search space efficiently, allowing high-quality initial solutions to be found. Moreover, EDDA allows redistributing the computational effort towards initial exploration, followed by solutions' refinement in the main evolutionary process. For example, given a budget of 1M fitness evaluations and a population of 1000 solutions, GP with the traditional ramped half-and-half (RHH) initialization takes 1000 generations to complete one run. This factor implies that the final solutions will be exposed to the variation operators 1000 times. In EDDA, assuming a configuration with 1000 demes of 100 individuals each, evolved for five generations, the main evolutionary process takes just 500 generations to complete one run with 1M fitness evaluations. This aspect implies that the final solutions will be exposed to the variation roughly twice as less (505 times), therefore presenting a smaller size. Moreover, the empirical evidence shows that the solutions obtained using EDDA happen to generalize significantly better than several traditional methods [9, 31].

Another approach for limiting the complexity of the final solutions is to use hoist mutation as a pruning method. Specifically, we propose to split the main evolutionary process (i.e., after EDDA initialization) into two halves: the first uses a combination of swap crossover and subtree mutation, whereas the second uses hoist mutation only. The experimental results show that this separation allows for reducing individuals' size while improving their generalization ability.

## 3.4 Fitness function

Given that we want the GP system to maximize solutions' association with the subjective evaluation provided by human observers, we formalized the fitness function $f$ as the modulus Spearman's rank correlation coefficient (SRCC) between both measures - the subjective evaluation and the respective outcome of GP individuals. SRCC is a widely accepted and used evaluation measure for IQA metrics in the community [29, 36]. In such a way, $f : S \rightarrow [0, 1]$, with higher values representing higher similarity with the subjective evaluation. Specifically, let $msssim_{gp}$ be a candidate multi-scale IQAM formulated by means of GP, $(X,\ Y)$ a tuple of reference-distortion pairs and $MOS$ the respective target (mean opinion score), the fitness of $msssim_{gp}$ is computed as $f\Big(msssim_{gp}(X,\ Y),\ MOS\Big)$.

**Table 1: Summary features of the considered IQA databases. The columns $W_i$x$H_i$ and D/$H_i$ stand for image resolution and viewing distance in terms of the image height, respectively. *References, Distortions, and Pairs* refer to the number of reference images, distortion types, and resulting reference-distortion pairs, respectively.**

| Features | TID2013 [26] | VDID2014 [18] |
|---|---|---|
| $W_i$x$H_i$ | 512×384 | 768×512, 512×512 |
| D/$H_i$ | 3 | 4, 6 |
| #References | 25 | 8 |
| #Distortions | 24 | 3 |
| #Pairs | 3000 | 160 |

# 4 EXPERIMENTS

## 4.1 Data

The proposed approach is assessed on two well-known databases to assess image quality aspects. In this sub-section, the reader can find their detailed description and instructions on how these were used. Both databases were created using reference images that account for diverse visual scenes and contain several different types of distortion with different intensity degrees. The reference-distortion pairs were subjectively evaluated by involving hundreds of volunteers in a controlled experimental environment.

We train our approach using one of the biggest and most popular databases for image quality assessment (IQA) - TID2013 [26]. In the first stage, the input reference-distortion pairs are partitioned by reference images into a training, validation, and test sets. The most prominent spatial scales are chosen based on the validation partition. In the second step, the training and validation partitions are merged. Moreover, to assess the generalization ability of the evolved measures to previously unseen viewing conditions, we use a popular dedicated viewing distance-changed image database VDID2014 [18]. Unlike in TID2013, where a convenient (preferred) distance to a monitor was used to compute the similarities between the reference-distortion pairs, the subjective assessment in VDID2014 was performed at two groups of typical viewing distances and image resolutions. The latter is motivated by the fact that the amount of extractable information in a visual scene depends on the viewing distance and image resolution. In this sense, we explicitly incorporate such an important aspect as the varying viewing distance in our study. Table 1 provides a detailed description of the IQA databases that were used.

## 4.2 Experimental settings

The experiments were repeated 30 times (runs) to provide an outright statistical analysis. A different seed for the pseudo-random numbers generator was used in each run to partition the data and initialize and execute the algorithm. The EDDA initialization technique was used with 200 individuals per deme ($200_{EDDA}$), each left to evolve for five generations ($EDDA_5$) [31]. During the EDDA initialization, subtree mutation was the only variation operator being used to foster the search space's exploration. The selection was the tournament. During the EDDA initialization, random selection pressure in [0.05, 0.2] interval was generated for every deme, whereas during the main evolutionary process (i.e., after the initialization), a pressure of 10% was used. Contrarily to the traditional 5%, a higher pressure was considered to foster convergence given the relatively small amount of generations. Once the population was initialized, the first half of GP's main evolutionary process (i.e., the first ten generations) was conducted using the swap crossover and the subtree mutation, with probabilities of 0.7 and 0.3, respectively. In the second half of the process (i.e., the remaining ten generations), the hoist mutation was used, and the mutation's probability was increased to 1.0 to foster trees' pruning (i.e., no crossover was used). Table 2 provides a complete enumeration of the parameters used in this study. The hyper-parameters were selected following the findings from the literature to avoid a computationally demanding tuning phase.

We train our system using the TID2013 [26] IQA database. In the first stage, the input reference-distortion pairs are partitioned by the reference images into train, validation, and test sets containing 64%, 16%, and 20%, respectively. The partition is 80% and 20% for training and test sets in the second stage, respectively. Moreover, the evolved solutions are assessed for generalization ability on a previously unseen database VDID2014 (which is also distance-changed).

We rely on GPOL [7] to conduct our experiments. GPOL is a felxible and efficient multipurpose optimization library in Python that covers a wide range of stochastic iterative search algorithms, including GP. Its flexible and modular implementation allows for solving optimization problems like the one in this study.

## 4.3 Experimental findings

Section 4.3.1 shows the findings regarding the first stage, which aims at estimating the subset of the most prominent spatial scales (i.e., features). The findings regarding the second stage, consisting of a final estimation of MS-SSIM by means of GP, can be found in Section 4.3.2.

*4.3.1 Stage 1.* We computed the frequency of the input terminals in the elite GP trees as a proxy for their worth to identify the most prominent approach for multi-scale contextual information extraction. Figure 1 shows the most relevant combinations of the dilation rates and the window sizes, regardless of the similarity component. From the figure, it becomes clear that the most prominent approach for extracting the multi-scale contextual information for FR-IQAMs consists of using spatially-varying Gaussian kernels with a dilation rate of 1 (consider the top 5 bars). Moreover, there is an indication that smaller window sizes are preferred. These findings appear consistent with the nature of artifacts introduced in the IQAM database, which consists of local degradations whose appearance is well described by convolutional kernels with smaller spatial resolution. To dig deeper, we analyze the frequency distribution of different dilation rates at each window size. Figure 2 shows that, among all the possible combinations, a dilation rate of 1 happens to be the most frequent for every window size considered in our experiments, except for the largest window (13). Interestingly, the greater the spatial size of the Gaussian kernel, the higher is the frequency of the two largest dilation rates (8 and 9).

**Table 2: Summary of the hyper-parameters. Note that $P(C)$ and $P(M)$ indicate the crossover and the mutation probabilities.**

| Parameters | Values |
|---|---|
| №runs | 30 |
| № generations | $\{5_{EDDA}, 20\}$ |
| Population's size | $\{200_{EDDA}, 1000\}$ |
| Terminals | $\{T_{stage1}, T_{stage2}\}$ |
| #Terminals | $\{\text{stage}_1 = 162, \text{stage}_2 = 15\}$ |
| Functions ($F$) | $\{+, \text{x}, x^i\}, i \in [0.05, 0.15, 0.3, 0.5, 0.8, 1/0.8, 1/0.5, 1/0.3, 1/0.15, 1/0.05]$ |
| Initialization | $EDDA_5$ subtree mutation and swap crossover |
| Selection | tournament with 10% pressure |
| Crossover | swap crossover |
| Mutation | {subtree mutation, hoist mutation} |
| $P(C)$ | {0.7, 0.0} |
| $P(M)$ | {0.3, 1.0} |
| Stopping criteria | № generations |

The analysis of the two figures allows us to conclude that we can significantly simplify the terminal set by using only five spatially-varying Gaussian kernels with a dilation rate of 1, whose window sizes are 3, 5, 7, 9, and 11. Given that there are three similarity components to be extracted at each spatial scale, the terminal set for the second stage $T_{stage2}$ will be reduced, therefore, to 15 terminals ($5 \times 3$).

Figure 3 shows the learning curves during the first stage. The sub-figure on the left shows elite individuals' averaged fitness as the SRCC (on the vertical axis) across the generations (on the horizontal axis). The sub-figure on the right shows how elites' length (on the vertical axis) progresses through the generations (on the horizontal axis). The gray vertical line divides the evolutionary process into two parts. The first part (on the left) uses a combination of swap crossover and subtree mutation (with probabilities of 70 and 30%, respectively). The second part only uses the hoist mutation to prune the trees. The red line regards the traditional MS-SSIM's test fitness and length (in the left and right sub-figures, respectively). By looking at the figure, we conclude that:

- the proposed GP system outperforms standard MS-SSIM at modeling the perceived visual quality of the media;
- the proposed GP system converges to a point of stability and presents a small gap with the training loss;
- the use of hoist mutation allows to marginally improve the individuals' generalization ability while noticeably reducing their size;
- the number of generations we defined seems to be a good compromise for the underlying optimization task; therefore, we will stick to them in the second stage;
- although elite individuals' length appears to be significantly higher, the mathematical simplification of evolved solutions shows that GP can effectively evolve solutions with a lower complexity (to be shown in Section 4.3.3)

To support the findings above, we used the Wilcoxon rank-sum test for pairwise data comparison under the null hypothesis that the differences between two related paired samples are symmetrically about zero. It is worth pointing out that we reject the null hypothesis when the p-value of the test is smaller or equal to a 5% significance level. Concretely, we compared the SRCC achieved by genetically-evolved multi-scale IQAMs against MS-SSIM on unseen data partitions of TID2013. The records for the proposed approach were taken on the last generation of every run. The p-value of the test is less than 0.05, which demonstrates that our approach statistically outperforms MS-SSIM.

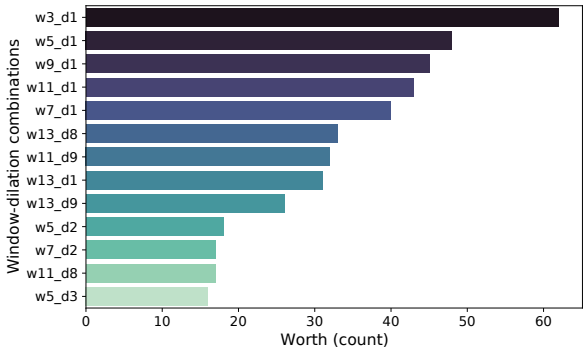


**Figure 1: Most frequently used combinations of the dilation rate and the window size in elite GP individuals during the first stage.**

*4.3.2 Stage 2.* The experimental findings of the first stage allowed us to conclude that GP can effectively aggregate similarity components at multiple scales using mainly five spatially-varying Gaussian kernels with a dilation rate of 1. In this stage, we repeat the experiments using only the most prominent features that model multi-scale contextual information.

Figure 5 shows the learning curves of the second stage. Like in Figure 3, the sub-figure on the left shows elite individuals' fitness, while the sub-figure on the right shows how elites' length grows. The gray vertical line denotes the transition point when only the hoist mutation is used. The red line indicates the baseline MS-SSIM. By looking at the figure, we could reinforce the observations found in Figure 3. The p-value of the Wilcoxon rank-sum test for pairwise

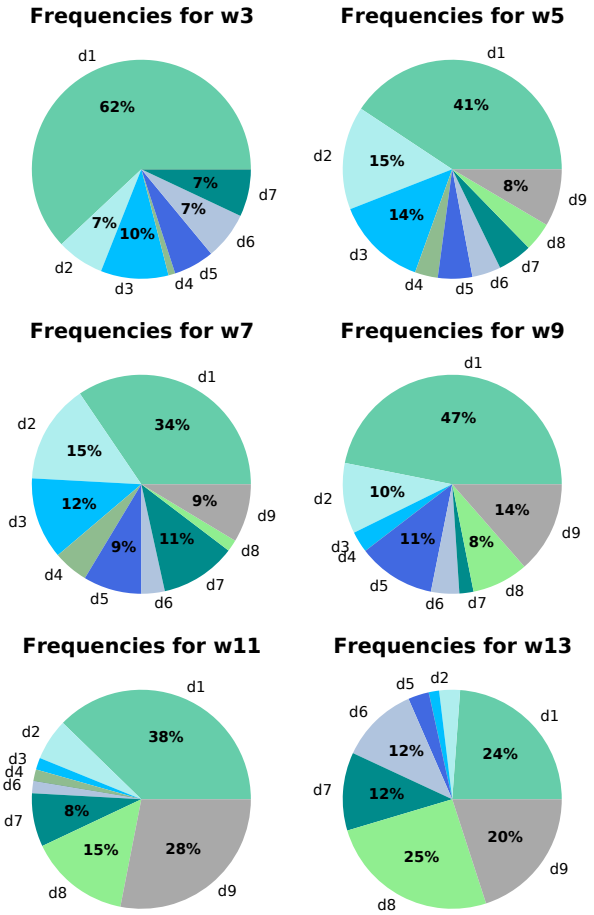


**Figure 2: Distribution of different dilation rates for each window size during the first stage.**

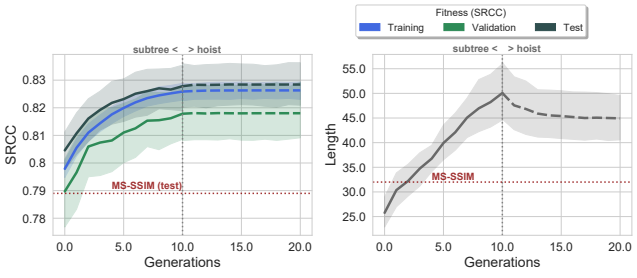


**Figure 3: Elite individuals' fitness (on the left) and complexity (on the right) during the first stage. The gray vertical line divides the main evolutionary process according to the variation strategy (subtree mutation and swap crossover on the left, hoist mutation on the right). The red line regards the traditional MS-SSIM.**

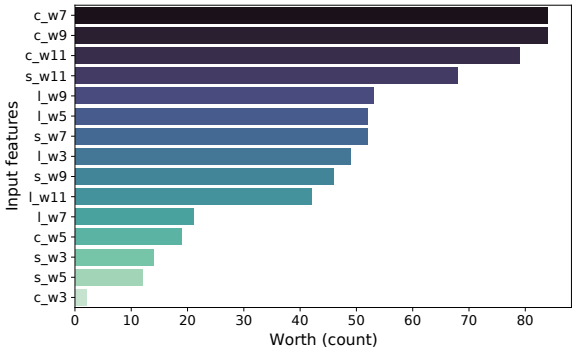


**Figure 4: Most frequently used terminals in elite GP individuals during the second stage.**

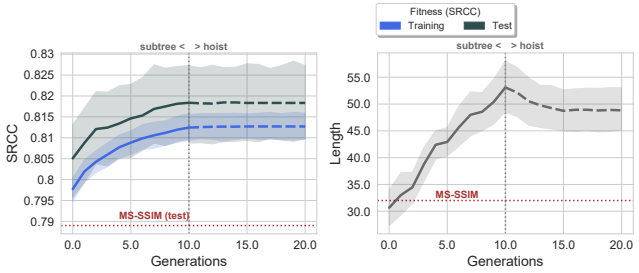


**Figure 5: Elite individuals' fitness (on the left) and complexity (on the right) during the second stage.**

data comparison demonstrated, once again, that our approach is statistically better than MS-SSIM. Nonetheless, it is necessary to note a slight decrease in elites' generalization ability and a mild increase in their length. Such a phenomenon has an explanation. The selection of the most prominent features based on frequency is a mere approximation and is, therefore, incomplete. Although the feature selection in stage 1 was necessary to brace final solutions' interpretability under the light of full-reference image quality assessment, it deprived GP of potentially useful, although dispensable, building blocks to achieve superior performance. Therefore, the evolution created individuals of relatively higher complexity (assessed by length) than in the first stage.

Figure 4 ranks the features in $T_{stage2}$ by their frequency. From the figure, it becomes clear that the potentially most relevant features in the GP-reformulated MS-SSIM comprise contrast and structural similarities in the first place. These data-driven findings are consistent with other works in the scientific literature, which analyzed the individual contribution of luminance, contrast, and structure for an effective SSIM-based IQAM [6, 8].

*4.3.3 Evolved solutions and generalization.* By following the probabilistic interpretation introduced in Section 3.2, we can provide some additional insights about the IQAMs evolved through GP. In practice, our optimization process appears to favor describing the overall similarity as the probability that the image pair is similar

according to either similarities, instead of being simultaneously similar according to different similarities. The addition operator, in fact, appears in the final solutions with a significantly higher frequency than the multiplication operator (more than three times as likely). A couple of example solutions, after mathematical simplification, are as follows:

$$MS-SSIM_1 = {s_7}^{\frac{100}{3}} + \left({s_7}^{0.8} + {l_9}^{\frac{8000}{3}} + {c_7}^{\frac{160}{3}}\right)^{133.3} + {l_9}^{\frac{8000}{3}} \quad (7)$$

$$MS-SSIM_2 = {s_5}^{400} + \left({l_3}^{320} + {c_5}^{\frac{320}{3}}\right)^{8888.8} + {l_{11}}^{\frac{8000}{9}} {l_3}^{\frac{1000}{3}} + {c_5}^{320} + {c_5}^{\frac{400}{9}} \quad (8)$$

Here, we refer to luminance, structure, and contrast similarity between the reference and the distorted image as, respectively, $l$, $s$, $c$, and the subscript refers to window size. The first thing we can observe by analyzing Equations 7 and 8 is that their structural form happens to be simpler than MS-SSIM's, here reported:

$$MS-SSIM = {l_{11}^{(5)}}^{\alpha} {c_{11}^{(1)}}^{\beta_1} {s_{11}^{(1)}}^{\gamma_1} {c_{11}^{(2)}}^{\beta_2} {s_{11}^{(2)}}^{\gamma_2} {c_{11}^{(3)}}^{\beta_3} {s_{11}^{(3)}}^{\gamma_3} {c_{11}^{(4)}}^{\beta_4} {s_{11}^{(4)}}^{\gamma_4} {c_{11}^{(5)}}^{\beta_5} {s_{11}^{(5)}}^{\gamma_5} \quad (9)$$

The superscript number in parentheses indicates the image scale at which the similarity component is processed (in our case this is always 1, and thus omitted from the formulation for better readability). The number of individual components is, respectively, 11 for the traditional MS-SSIM, and 5 and 7 for our example solutions, thus suggesting faster execution times. The actual processing cost of each component is, however, a complex factor determined by the downscaling operation (not necessary in our solutions), and the convolution with a Gaussian filter of a given size on an input whose size is determined by the downscaling operation itself. For these reasons, we reserve for future work the direct measurement of inference time of a direct implementation of our solutions.

Moreover, one can notice that the involved exponents are numerically very high when compared to the values proposed by the original SSIM authors [38] and by other SSIM-related optimization approaches [6, 7]. When combining probabilities through the addition rule, however, the interpretation of the role of exponents is inverted with respect to the use of multiplication. Suppose we see the addition as an operator stacking individual contributions of sub-similarities. In that case, a low exponent will still yield a value approaching maximum probability 1, thus providing a significant contribution to reaching the conclusion of high overall similarity. Finally, it should be noted that, despite this probabilistic interpretation, no explicit constraint was put towards having an upper bound on the overall similarity. In fact, we are mainly interested in the ordinal relationship produced by the similarity expression, as well represented by the Spearman rank correlation.

Table 3 shows the performance of the above-mentioned equations 7 and 8 on both training and test partitions of TID2013 and the whole set of reference-distortion pairs of VDID2014; recall that the latter was not used for training. The solutions were extracted at the end of two independent evolutionary processes (i.e., two different runs). The performance is reported as the SRCC between solutions' similarity scores and human observers' subjective assessment. In parenthesis, the performance of traditional MS-SSIM is reported. The experimental results indicate that GP can evolve multi-scale IQAMs that demonstrate better performance, besides showing a low complexity. Moreover, the empirical evidence indicates that the evolved measures can be successfully transferred to other previously unseen databases (such as VDID2014).

**Table 3: Performance of example individuals on TID2013 (both training and test partitions) and VDID2014. Recall that the latter was not used in training. The performance of traditional MS-SSIM is reported in parenthesis.**

| Individuals | $TID2013_{train}$ | $TID2013_{test}$ | VDID2014 |
|---|---|---|---|
| $MS-SSIM_1$ | **0.8034** (0.7856) | **0.8043** (0.7634) | 0.8962 **(0.8995)** |
| $MS-SSIM_2$ | 0.7943 **(0.7947)** | **0.8284** (0.7486) | **0.9007** (0.8995) |

## 5 CONCLUSIONS

MS-SSIM is among the most utilized full-reference image quality assessment measures. It aggregates diverse similarity statistics at multiple spatial scales, similar to how DCNNs aggregate multi-scale context when solving a given CV task. This work proposed the use of GP to generate novel multi-scale FR-IQAMs based on MS-SSIM. Specifically, we define a new set of terminal symbols to represent solutions, allowing us to foster a precise multi-scale quality assessment. The approach is partitioned into two stages: (i) use GP to estimate the subset of the most prominent spatial scales (i.e., features), and (ii) redesign the MS-SSIM through GP using previously estimated scales. To encourage small solutions, we initialize GP's population using the EDDA technique, and we perform trees' pruning through hoist mutation. The approach is trained and assessed using a sizeable real-world database for IQA. The evolved solutions are additionally evaluated for their capability to generalize on other problems, using a dedicated distance-changed IQA database. The experimental results show that the proposed approach outperforms the traditional MS-SSIM. Also, the evolved individuals present a lower degree of complexity. A comprehensive analysis of the feature selection indicates that, for extracting multi-scale similarities, spatially-varying convolutions are more effective than dilated convolutions. Empirical evidence also shows that the evolved measures can be successfully transferred to previously unseen databases. The current work paves the way for a brand new application area for evolutionary computation and GP in particular to IP and CV: given the urge for precise and simultaneously efficient IQAMs, several GP techniques can be applied to improve upon this trade-off, such as bloat control methods and convergence-based stopping criteria [30, 32]. Future work directions include exploring other categories of feature-selection algorithms besides filters.

## ACKNOWLEDGEMENTS

FCT Portugal partially supported this work, under the grand SFRH/BD/137277/2018, and through projects BINDER (PTDC/CCI-INF/29168/2017) and AICE (DSAIPA/DS/ 0113/2019).

# REFERENCES


[1] 2022. 24 Noteworthy Video Consumption Statistics [2021 Edition]. https://techjury.net/blog/video-consumption-statistics/#gref. Accessed: 19.01.2022.

[2] 2022. 31 Mind-Boggling Instagram Stats & Facts for 2022. https://www.wordstream.com/blog/ws/2017/04/20/instagram-statistics. Accessed: 19.01.2022.

[3] 2022. Global social media statistics research summary 2022. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. Accessed: 19.01.2022.

[4] 2022. The State Of Online Video For 2020. https://www.forbes.com/sites/tjmccue/2020/02/05/looking-deep-into-the-state-of-online-video-for-2020/?sh=73ed21902eac. Accessed: 19.01.2022.

[5] Yuval Bahat and Tomer Michaeli. 2021. What's in the Image? Explorable Decoding of Compressed Images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 2907–2916.

[6] Illya Bakurov, Marco Buzzelli, Mauro Castelli, Leonardo Vanneschi, and Raimondo Schettini. 2020. Parameters optimization of the Structural Similarity Index. *London Imaging Meeting 2020: Future Colour Imaging* 2020, 19–23. https://doi.org/10.2352/issn.2694-118X.2020.LIM-13

[7] Illya Bakurov, Marco Buzzelli, Mauro Castelli, Leonardo Vanneschi, and Raimondo Schettini. 2021. General Purpose Optimization Library (GPOL): A Flexible and Efficient Multi-Purpose Optimization Library in Python. *Applied Sciences* 11, 11 (2021). https://doi.org/10.3390/app11114774

[8] Illya Bakurov, Marco Buzzelli, Raimondo Schettini, Mauro Castelli, and Leonardo Vanneschi. 2022. Structural similarity index (SSIM) revisited: A data-driven approach. *Expert Systems with Applications* 189 (2022), 116087. https://doi.org/10.1016/j.eswa.2021.116087

[9] Illya Bakurov, Leonardo Vanneschi, Mauro Castelli, and Francesco Fontanella. 2018. EDDA-V2–An Improvement of the Evolutionary Demes Despeciation Algorithm. In *International Conference on Parallel Problem Solving from Nature*. Springer, 185–196.

[10] Simone Bianco, Luigi Celona, and Paolo Napoletano. 2021. Disentangling Image distortions in deep feature space. *Pattern Recognition Letters* 148 (2021), 128–135. https://doi.org/10.1016/j.patrec.2021.05.008

[11] Simone Bianco, Claudio Cusano, Flavio Piccoli, and Raimondo Schettini. 2020. Personalized Image Enhancement Using Neural Spline Color Transforms. *IEEE Transactions on Image Processing* 29 (2020), 6223–6236. https://doi.org/10.1109/TIP.2020.2989584

[12] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek. 2018. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. *IEEE Transactions on Image Processing* 27, 1 (2018), 206–219.

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018), 834–848.

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *ArXiv* abs/1706.05587 (2017).

[15] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. 2020. Pyramidal Convolution: Rethinking Convolutional Neural Networks for Visual Recognition. *ArXiv*.

[16] Max Ehrlich, Ser-Nam Lim, Larry S. Davis, and Abhinav Shrivastava. 2020. Quantization Guided JPEG Artifact Correction. In *ECCV*.

[17] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. 2017. DeepSim: Deep similarity for image quality assessment. *Neurocomputing* 257 (2017), 104–114.

[18] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang. 2015. Quality Assessment Considering Viewing Distance and Image Resolution. *IEEE Transactions on Broadcasting* 61, 3 (2015), 520–531.

[19] Justin Johnson, Alexandre Alahi, and Fei Fei Li. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution, Vol. 9906. 694–711. https://doi.org/10.1007/978-3-319-46475-6_43

[20] Ke Gu, Guangtao Zhai, Xiaokang Yang, Wenjun Zhang, and Min Liu. 2013. Structural similarity weighting for image quality assessment. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 1–6.

[21] J.R. Koza. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. [n. d.]. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 2012.

[23] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. 2016. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4905–4913.

[24] Davide Mazzini. 2018. Guided Upsampling Network for Real-Time Semantic Segmentation. *ArXiv* abs/1807.07466 (2018).

[25] Anush Krishna Moorthy and Alan Conrad Bovik. 2010. A Two-Step Framework for Constructing Blind Image Quality Indices. *IEEE Signal Processing Letters* 17, 5 (2010), 513–516. https://doi.org/10.1109/LSP.2010.2043888

[26] Nikolay Ponomarenko, Lina Jin, O. Ieremeiev, Vladimir Lukin, Karen Egiazarian, J. Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* 30 (01 2015), 57–77. https://doi.org/10.1016/j.image.2014.10.009

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

[28] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.1556

[29] Robert Streijl, Stefan Winkler, and David Hands. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* 22 (03 2016), 213–227. https://doi.org/10.1007/s00530-014-0446-1

[30] Leonardo Trujillo, Luis Muñoz, Edgar Galván-López, and Sara Silva. 2016. neat Genetic Programming: Controlling bloat naturally. *Information Sciences* 333 (2016), 21–43. https://doi.org/10.1016/j.ins.2015.11.010

[31] Leonardo Vanneschi, Illya Bakurov, and Mauro Castelli. 2017. An initialization technique for geometric semantic GP based on demes evolution and despeciation. In *Evolutionary Computation (CEC), 2017 IEEE Congress on*. IEEE, 113–120.

[32] Leonardo Vanneschi, Mauro Castelli, and Sara Silva. 2010. Measuring Bloat, Overfitting and Functional Complexity in Genetic Programming. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation* (Portland, Oregon, USA) *(GECCO '10)*. Association for Computing Machinery, New York, NY, USA, 877–884. https://doi.org/10.1145/1830483.1830643

[33] Abhinau K. Venkataramanan, Chengyang Wu, Alan Conrad Bovik, Ioannis Katsavounidis, and Zafar Shahid. 2021. A Hitchhiker's Guide to Structural Similarity. *IEEE Access* 9 (2021), 28872–28896.

[34] Cong Wang, Wanshu Fan, Yutong Wu, and Zhixun Su. 2020. Weakly supervised single image dehazing. *Journal of Visual Communication and Image Representation* 72 (2020), 102897. https://doi.org/10.1016/j.jvcir.2020.102897

[35] Zhou Wang and A. C. Bovik. 2002. A universal image quality index. *IEEE Signal Processing Letters* 9 (2002), 81–84.

[36] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE TRANSACTIONS ON IMAGE PROCESSING* 13, 4 (2004), 600–612.

[37] Zhou Wang and Qiang Li. 2010. Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Image Proc. 20(5), 1185-1198. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 20 (11 2010), 1185–98. https://doi.org/10.1109/TIP.2010.2092435

[38] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. 2003. Multi-Scale Structural Similarity for Image Quality Assessment.

[39] Fisher Yu and Vladlen Koltun. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1511.07122

[40] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. 2017. Dilated Residual Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[41] Hu Zhang, Keke Zu1, Jian Lu, Yuru Zou, and Deyu Meng. 2021. EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network. *ArXiv*.

[42] L. Zhang, L. Zhang, X. Mou, and D. Zhang. 2011. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing* 20, 8 (2011), 2378–2386.

[43] Xiaoshuai Zhang, Wenhan Yang, Yueyu Hu, and Jiaying Liu. 2018. Dmcnn: Dual-Domain Multi-Scale Convolutional Neural Network for Compression Artifacts Removal. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. 390–394. https://doi.org/10.1109/ICIP.2018.8451694

[44] Jing Zhao, Ruiqin Xiong, Jizheng Xu, and Tiejun Huang. 2019. Learning a Deep Convolutional Network for Subband Image Denoising. 1420–1425. https://doi.org/10.1109/ICME.2019.00246

[45] Simone Zini, Simone Bianco, and Raimondo Schettini. 2020. Deep residual autoencoder for blind universal jpeg restoration. *IEEE Access* 8 (2020), 63283–63294.