MEASURING IMAGE COMPLEXITY AS A DISCRETE HIERARCHY USING MDL CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Being able to quantify the complexity of data is an important question in machine learning, computer science, and data science. In the case of image data, a number of methods have been proposed. However, existing methods are based only on the degree of variation across the image, and cannot distinguish meaningful content from noise. In particular, existing methods assign a very high complexity to white noise images, despite such images containing no meaningful information. In this paper, we present a method to measure the complexity of images by analyzing them has a discrete hierarchy of patches, using MDL clustering. Beginning with individual pixels, each level of the hierarchy is formed using the cluster labels from the level below. The complexity is the sum, across all levels, of the entropy of cluster labels inside all patches on that level. Clustering is performed using the minimum description length principle (MDL), which we leverage in a novel way to distinguish signal from noise. We test against existing methods on seven different sets of images, four from public image datasets and three synthetic, and show that ours is the only method that can assign an accurate measure of complexity to all images considered. Every other method measures white noise as highly complex, while our method gives it zero complexity. We then present ablation studies showing the contribution of the components of our method, and further experiments showing robustness to image quality.

1 INTRODUCTION

There is unavoidable subjectivity in trying to quantify the notion of complexity. This is always the case when defining a new metric. We cannot begin the investigation of a complexity metric by defining what we take complexity to mean, that would be to put the cart before the horse. Inevitably, the investigation involves exploring what complexity is, not just how to measure it, that is, the definition of complexity and the specification of a complexity metric are two sides of the same thing, the latter is really an instantiation of the former. For example, if one defines complex images as those in which there is high variation between all the pixel values, then it is natural to use the entropy of pixel values as a complexity metric; or if one defines complex images as those in which nearby pixel intensities tend to be very different from each other, then another metric is the obvious choice (grey-level co-occurrence matrix, (Section 4). This renders unavailable the standard blueprint for applied machine learning research of showing that a novel method outperforms existing methods on some quantifiable task or benchmark, because the field does not have such a benchmark for measuring image complexity. What we do have is a vague idea of what complexity is, vague but still powerful and important. The task is to translate this vague idea into something computable.

Many existing techniques for quantifying and measuring image complexity (discussed further in Section 2) are based on measuring intricacy, the idea being that the more intricate it is and the more dissimilar its parts, the more complex it is. This is relatively easy to measure, but it is incomplete for two reasons. Firstly, and most importantly, it does not distinguish between meaningful intricacy (signal) and meaningless intricacy (noise). Using intricacy as a measure of complexity means that a white-noise image, where the pixel values are chosen independently at random, is measured as highly, perhaps even maximally, complex, because there is a high degree of difference between neighbouring pixels. Secondly, it cannot capture the fact that images can have a different complexity at different scales. A blurry photograph of a complex scene, for example, is locally simple but globally complex, while a finely-detailed but repetitive pattern is the opposite.

Rather than meaning a high degree of variation, we instead conceive of complexity as 'taking a large number of steps to assemble'. An image can be thought of as being built out of pixels, local groups of pixels are combined to form patches, groups of neighbouring patches are combined to form super-patches etc. Quantifying complexity based on the assembly process is the approach taken in the theory of assembly pathways Cronin et al. (2006); Marshall et al. (2019), originally for the purpose of quantifying the complexity of molecules to aid in the search for extraterrestrial life Marshall et al. (2021); Schwieterman et al. (2018). The pathway assembly index of an object is the minimum number of combinations needed to produce it from simple parts, where repeated components can be reused without adding to the count. In order to discretize the structure of the image and allow assembly index to be applied, we employ clustering. For the first level of the hierarchy, we cluster the pixel values and replace them with their cluster index. For higher levels, we cluster the multisets of cluster indices in from the level below. This is a similar idea to that used by convolutional neural networks, which also process an image patch-wise and hierarchically. Another advantage of discretizing is that we can then easily compute entropy. Taking entropy of a continuous image is difficult, we must use some approximation of differential entropy Hulle (2005); Pichler et al. (2022). In our case, however, we are dealing with discrete cluster labels, so we need only compute the entropy of a categorical distribution, which is easy. At each scale (i.e. hierarchy level), we compute the entropy of the multisets of cluster indices across the image to quantify complexity. The total complexity score is the sum of this entropy at each scale. We can also examine the entropy for individual scales to get an indication of the local vs. global complexity in the image: low scales (i.e., small patch sizes) measure local complexity, whereas higher scales capture more global structure (as shown in Section 4).

At each level of the hierarchy, the cluster indices produced depend on K, the number of clusters in the clustering the model. We choose K in a sound way by employing the minimum description length (MDL) principle Rissanen (1983). MDL says that we should choose the model that can completely represent the given data in the fewest number of bits. Clustering can be interpreted as compression, where we encode each point by its cluster index, along with the residual error of how it differs from the centroid of that cluster. Treating each cluster as a probability distribution, and employing the Kraft-McMillan inequality, we see that the residual error for a point x under the cluster probability distribution p can be represented using $-\log p(x)$ bits. Representing the data under the clustering model takes $-\sum_x \log p(x)$ bits, plus the number of bits to represent the cluster indices and the model itself. Increasing K reduces the average residual error, but increases the size of the indices and the model itself. By MDL, we choose K so as to minimize the total size. MDL is a key component in filtering out noise from our complexity measure. In white noise images, where there is no meaningful or consistent pattern between different points, MDL finds only one cluster, because the small reduction in residual error from encoding more is not worth the extra cost, so the image ends up with a very low complexity score. With respect to real-world applications of image complexity metrics, being able to distinguish signal from noise is especially relevant to remote sensing, where images often become corrupted by noise due to the sensing equipment or various post-processing steps Chioukh et al. (2014); Narayanan et al. (2003); Landgrebe & Malaret (1986). Much work has been done to reduce noise in remote sensing images and to improve the robustness of image processing methods to noise Chang et al. (2016); Rasti et al. (2018); Huang et al. (2020); Duan et al. (2019); Chen et al. (2014).

The contributions of this paper are briefly summarized below.

- We propose a novel measure of image complexity, which has a clear theoretical interpretation and is rigorously grounded in information theory.
- We test our method empirically on seven image datasets, four public and three synthetic datasets that we created. We show that our method performs as desired in distinguishing images from different datasets. In particular, our method is able to correctly assign a low complexity to white noise, in contrast to existing methods, which assign it a high complexity.
- In a further empirical analysis, we show how our method can measure complexity at different scales in the image, and show its robustness to changes in image quality.
- We release the code (on publication) for our method, for the creation of our synthetic datasets and for our testing procedure.

The rest of this paper is organized as follows. Section 2 gives an overview of related work. Section 3 describes our method, and Section 4 presents our empirical evaluation. Finally, Section 5 summarizes our findings and suggests directions for future work.

2 Related Work

Image complexity measures are used in a number of different tasks: remote sensing Falconer (2004); Sun et al. (2006); Yang & Zhou (2000), automatic target recognition Peters & Strickland (1990); Wang et al. (2018), psychometrics Forsythe et al. (2008), user interface design Stickel et al. (2010), and measuring aesthetic properties of art Forsythe et al. (2011); Carballal et al. (2020). Existing works fall into one of a several approaches.

Fractal dimension is a property of curves, which in some sense measures their complexity Mandelbrot (1967); Falconer (2004). It can be applied to an image by first binarizing with a threshold, then taking the boundary between white and black pixels as a curve and computing its Minkowski-Bouligand dimension. Lam et al. (2002) explores the use of fractal dimension to measure the complexity of satellite images, and Sun et al. (2006) considers the application to remote sensing images more generally. Both also contain a detailed account of methods that use fractal dimension for image complexity. In Forsythe et al. (2011), fractal dimension is compared against human judgements of the complexity and beauty of visual art.

The **file compression ratio** is the ratio between the size of a compressed file under a chosen compression algorithm, and the size of the uncompressed original. In Marin & Leder (2013), image complexity is measured using the file compression ratio, under two compression algorithms: GIF, which is lossy, and TIFF, which is lossless. The compression ratio was compared to human judgements of complexity, on the International Affective Picture System. It is also used as a complexity measure in Forsythe et al. (2011) and in Machado et al. (2015). The former investigates the ability of JPEG, GIF, and a novel ‘perimeter detection’ method to predict human judgements of complexity in visual art. The latter explores various combinations of compression algorithms with automated edge detection, and compares the results to human judgements of complexity. The authors find the best results using Sobel and Canny filters, followed by JPEG compression.

An alternative method is to use the **gradient of pixel intensities** across the image. This is the approach taken by Redies et al. (2012). The gradient is computed separately for each of the RGB channels, and the gradient at a pixel is taken to be the maximum across the three channels. The average gradient across the entire image is then taken as a measure of complexity. This is again applied to quantifying aesthetic judgement of visual art, this time as part of the Birkhoff-like measure Birkhoff (1933), which characterizes beauty as the ratio of order and complexity.

A final method to consider is **the Fourier transform**, as used by Khan et al. (2021). The idea is that the more high-frequency components present in the power spectrum, the more complex the image. The authors investigate using both the mean and the median of the power spectrum. The best results are found for the median power spectrum. The application in this case is guiding neural architecture search, the claim being that one should first measure the complexity of a given image dataset, and then use the result to inform architecture design.

3 Method

3.1 Minimum Description Length Patch Clustering

The clustering of patches, which is an important component of our complexity metric, is based on description length, i.e., the number of bits needed to specify the given data. Description length is relative to an encoding scheme, and via the Kraft-MacMillan inequality, this corresponds to a probability distribution. Specifically, the Kraft-MacMillan inequality says that, under the optimal encoding scheme (optimal in the sense of being shortest on average) of a probability distribution $p(\cdot)$, the description length of a point x is $-\log p(x)$. We model the probability distribution with a Gaussian mixture model (GMM) because (a) we seek a distribution-based clustering model, and a GMM is by far the most commonly used distribution-based cluster model, (b) choosing a GMM is equivalent to simply modelling the distribution within each cluster as Normal, and this has theoretical justifications in the central limit theorem and maximization of differential entropy Thomas & Joy (2006). The description length is then relative to its means $\mu = (\mu_i)_{1 \leq i \leq K}$ and covariances $\Sigma = (\Sigma_i)_{1 \leq i \leq K}$ of the GMM. The probability density of a point x is given by

$$p(x, \mu, \Sigma) = \max_{1 \leq k \leq K} \frac{\exp(-\frac{1}{2}(x - \mu_k)\Sigma_k^{-1}(x - \mu_k))}{\sqrt{(2\pi)^d |\Sigma_k|}} \,, \tag{1}$$

where μ_k and σ_k are, respectively, the mean and covariance of the kth component, and d is the dimensionality of the data. Specifying x under p requires first indexing the cluster to which x belongs and then encoding x under the probability distribution of that cluster, which we refer to as the residual error. The latter we have just shown to take $-\log p(x, \mu, \Sigma)$ bits. Similarly, the length of the former depends on the encoding scheme for, and equivalently the probability distribution over, the indices $1, \ldots, K$, which can be taken empirically from the data. Specifically, the length of encoding which cluster x belongs to is $-\log n_k/N$, where k is the index of the cluster that it belongs to, n_k is the number of points belonging to cluster k, and N is the total number of data points.

Discretizing the Probability Density Function Because the multivariate normal distributions composing the GMM are continuous probability density functions, it is possible that $p(x, \mu, \Sigma) > 1$. The Kraft-MacMillan inequality would then seem to suggest that the corresponding encoding scheme can represent x with a strictly negative number of bits, which of course is not possible. The apparent contradiction is resolved by making explicit the precision with which we want to encode x. Completely specifying any real number is not possible with a finite number of bits, instead we can only specify an extended region $D_x \subset \mathbb{R}^n$, which is thought to contain x. The number of required bits is then determined by the probability mass inside D_x, which is given by

$$p_m(D_x, \mu, \Sigma) = \int_{D_x} p(z, \mu, \Sigma) dz. \tag{2}$$

Let ϵ be the coordinate-wise precision for specifying x, i.e., set D_x to be a hypercube of side-length ϵ. The probability mass in D_x is then approximated as $p(x, \mu, \Sigma)\epsilon^d$, giving description length

$$-d \log \epsilon - \log(p(x, \mu, \Sigma) + \log K \,. \tag{3}$$

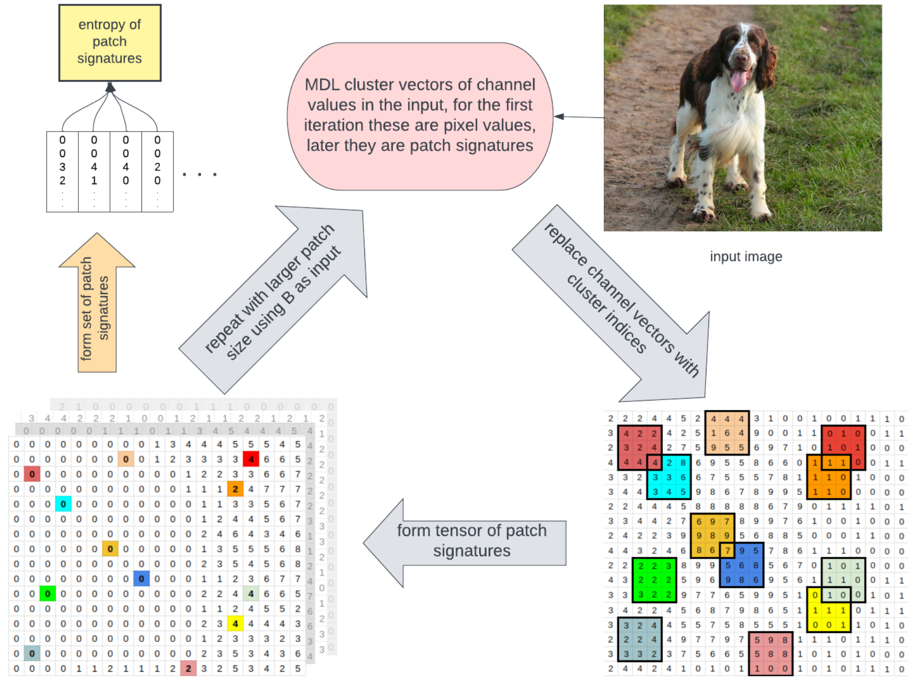
So, even if $-\log(p(x, \mu, \Sigma) < 0$, the total description length is still positive, as the probability mass in $equation$ 2 is always at most 1. In our experiments, we set $\epsilon = 2^{-32}$, corresponding to the maximum precision possible for a 32-bit float.

Determining Outliers As well as choosing the number of clusters (see Section 3.1), we use the minimum description length (MDL) principle, to precisely determine which points are outliers with respect to our model. Recall that we apply clustering to different parts of a single image, so identifying outliers can be interpreted as identifying irrelevant parts of an image. To our knowledge, no existing works have used the MDL principle in this way. An outlier can be defined as one that takes more bits to specify under the model than it does to specify directly, independently of the model. We can always specify (up to finite precision ϵ) any point directly using the same discretizing reasoning as above. First, we restrict our attention to some bounded region of $\mathbb{R}^n$, which is large enough that we can assume it will contain all values our data could have. There are several reasonable choices for such a bounded set. We find that the exact choice does not affect results. In our implementation, we choose the hypercube whose sides, in each dimension, run from the minimum to the maximum values across all dimensions in the dataset, denoted a_{max} and a_{min}, respectively. Once this bounded region is specified, we partition it into a set of small regions–hypercubes with side-length ϵ–and then specify a point x by indexing the unique region that contains x. Then, comparing to equation 3, x is an outlier iff

$$p(x, \mu, \Sigma)\frac{n_k}{K} < (a_{max} - a_{min})^{-d} \,, \tag{4}$$

where, as above, n_k is the number of points assigned to the same cluster as x (details in appendix). We can then define the description length of x, where x can be specified either directly or using the encoding scheme from the model, as

$$d(x, \mu, \Sigma) = -d \log \epsilon + \min \left(d \log(a_{max} - a_{min}), -\log(p(x, \mu, \Sigma) + \log \frac{N}{n_k} \right) \,. \tag{5}$$


entropy of patch signatures
MDL cluster vectors of channel values in the input, for the first iteration these are pixel values, later they are patch signatures
input image
form set of patch signatures
repeat with larger patch size using B as input
replace channel vectors with cluster indices
form tensor of patch signatures

B: 3d tensor of patch signatures, visible channel is number of 0's per patch, kth channel is number of k's per patch, colours show correspondence with A

A: 2d tensor of cluster indices

Figure 1: Method for computing the entropy of patch signatures as a measure of complexity. Each patch signature is the multiset of MDL cluster indices that appear there.

Determining the Number of Clusters The description length of a given image X depends on the number of clusters in the GMM, and using the MDL principle, we can determine the optimum number of clusters as that which produces the smallest description length.

Let $\mu(X, K), \Sigma(X, K)$ denote the values of μ and Σ with K components, which maximize the probability of X:

$$\mu(X, K), \sigma(X, K) = \underset{\mu, \Sigma}{\arg\max} \prod_{x \in X} p(x, \mu, \Sigma)\,. \tag{6}$$

Finding these optimal parameters means fitting the GMM to the dataset X, and can be performed with the usual expectation-maximization algorithm. Then, using $d(\cdot)$ from equation 5, the MDL-optimal number of clusters K^* is that which minimizes the total description length:

$$K^* = \underset{1 \leq K \leq |X|}{\arg\min} \sum_{x \in X} d(x, \mu(X, K), \sigma(X, K))\,. \tag{7}$$

We need only consider K up to the size of the dataset, as adding more clusters beyond that point can only increase the total description length. In practice, we test values up to 8, we find testing more does not change results. GMMs have diagonal covariances, are initialized with k-means, have tolerance $1e-3$, and are capped at 100 iterations. Total processing time is ∼10s per image, which can be reduced to ∼4s by only checking up to 5 clusters, with essentially the same results.

3.2 Hierarchical Patch Entropy

The method described in this section is depicted graphically in Figure 1. At each level of the hierarchy, we begin with a 3d tensor X of shape (H, W, C) and will cluster the vectors of the last dimension; on the first level, this means clustering 3d vectors specifying the colour intensities for each of the three colour channels at each point. Before clustering, the model computes K^* as in equation 7, then clusters the last-dimension vectors of X using a mixture model with K^* components. From this clustering, we can form the 2d tensor A, of shape (H, W) whose (i, j)th entry is the cluster index of the (i, j)th pixel in X, and B, the 3d tensor of shape $(H - m + 1, W - m + 1, K^*)$ whose (i, j, k)th entry is the count of how many times the kth cluster appears in the $m \times m$ patch beginning at (i, j) in A. The patch size m is a user-set parameter. We refer to the vector at (i, j) in B as the signature of the (i, j)th patch. Our measure of complexity at this level is the entropy of all signatures that appear in B. As we are dealing with discrete data, in the form of cluster labels, computing entropy is easy.

To measure complexity at larger scales, we repeat the above procedure, this time beginning with B instead of X, and interpreting the counts as points in continuous space. Let A_i and B_i be the tensors formed, as just described, on the ith level of the hierarchy. Then B_i contains the signatures (i.e. counts vectors) of the patches in A_i, and A_i contains the MDL-cluster indices of the last-dimension vectors in B_{i-1}. To begin the iteration, B_0 is set to X, the input image.

The present implementation computes up to B_4, and uses larger patch sizes for each level: 4, 8, 16, and 32. Note, however, that this is not the same as simply clustering larger patches of an image. What is clustered at each level is the cluster indices from the level below, so, apart from the first level, it is quite different from the input image. The full method is described in Algorithm 1.

Algorithm 1: Algorithm for computing the complexity of an image.

```
function MDL_CLUSTER(D)
    $best\_DL \leftarrow \infty$
    $A \leftarrow$ cluster indices of MDL of $D$, initialized randomly
    for $K \in \{1, \ldots, K\_max\}$ do
        fit a GMM with $K$ components to $D$
        $DL \leftarrow$ description length of $D$ under this fit GMM
        if DL ¡ best_DL then
            $A \leftarrow$ cluster indices of $D$ under this fit GMM
            $best\_DL \leftarrow DL$
    return $A$
function SIGNATURES_ENTROPY(S)
    $bin\_counts \leftarrow$ hash table whose keys are the unique elements in $S$, and whose values are the number of times that element occurs in $S$
    return $-\sum_{b \in bin\_counts} \frac{bin\_counts[x]}{|S|} \log \frac{bin\_counts[x]}{|S|}$
function COMPUTE_PATCH_SIGNATURES(X,m)
    $A \leftarrow$ MDL_CLUSTER($X$)
    $B \leftarrow$ multisets of cluster indices appearing in all $m \times m$ patches of $A$ (including overlapping)
    return $B$
function COMPLEXITY(X,scales)
    $total\_complexity \leftarrow 0$
    for $m \in scales$ do
        $X \leftarrow$ COMPUTE_PATCH_SIGNATURES($X, m$)
        $total\_complexity \leftarrow total\_complexity +$ SIGNATURES_ENTROPY($X$)
    return $total\_complexity$
```

4 Experimental Evaluation

It is difficult to assess the performance of an image complexity measure. Some works gather human subjective judgements on a particular distribution of images (e.g., European renaissance paintings) and report accuracy/correlation, often also training a supervised model on these human judgements Machado et al. (2015); Nagle & Lavie (2020). Aside from the practical difficulties of running these psychological studies, evaluating a model on a single distribution does not give a rounded indication of its accuracy, it is unclear how such models will perform when presented with a more diverse set of images. Additionally, collecting human judgements of complexity in this way may not be reliable, they have been shown to be influenced by the presentation of the image as well as cognitive factors such as visual working memory Sherman et al. (2013), and show high inter-subject variability Madrid-Herrera et al. (2019). There is also EEG evidence suggesting that humans use different cognitive processes to judge an image's complexity depending on its degree of naturalness/familiarity Nicolae & Ivanovici (2020).

We instead evaluate our method by presenting the scores that it produces for a diverse set of images of different types. Comparing sets/types of images, rather than individual images, has the advantage of reducing subjectivity. One can say with reasonable objectivity that ImageNet images are more complex than MNIST images, whereas trying to compare the complexity of two different images of the same type, such as renaissance paintings or ImageNet images, may be more subjective.

Datasets We present the average score of our method on seven different sets of images, four popular image datasets and three synthetic datasets that we created: **ImageNet** and **CIFAR** are datasets with high complexity, depicting real-world objects in context, of resolutions 224×224 and 32×32 respectively. **MNIST** depicts low-resolution greyscale digits. Its images are simple in that they can be represented with a few bits, but still have meaningful semantic content. **DTD2** is a dataset we created by manually searching through the Describable Textures Dataset Cimpoi et al. (2014) for all images of fine-detailed repeating textures (full list in appendix). **Stripes**, **Halves**, and **Rand** are our synthetic datasets of greyscale images of stripes of varying thickness and orientation, cleanly divided white-black halves, and independent uniformly random pixel values (i.e., white noise), respectively (details in appendix). For DTD2, we find 341 suitable images. For all other datasets, we report the average for 1500 randomly sampled images, all resized to 224×224.

Comparison with Existing Methods Table 1 compares our method to seven others: 'khan2021' Khan et al. (2021), 'machado2015' Machado et al. (2015), and 'redies2012' Redies et al. (2012) are as described in Section 2; 'entropy' converts the image to greyscale, discretizes the values into 256 bins, and then computes the Shannon entropy of the bin counts; 'fractal dim.' converts the image to greyscale, then binarizes it to 0 or 1, and computes the fractal dimension of the resulting shape using the box-counting method; 'jpg-ratio' measures the ratio of the JPEG-compressed file size to that of the original; and 'GLCM' computes the the average entropy of the grey-level co-occurrence matrix, at offsets 1, 4, 8, 16 32 (see Sebastian V. et al. (2012) for details of GLCM in image complexity). The most striking result is that our method assigns zero complexity to white-noise images, while every other method assigns them high complexity, with many assigning maximum complexity. White noise images are not at all meaningful or interesting to humans, and it is a significant finding that our method is the first to reflect this. It suggests that, while existing methods are based only on the variation across the image, our method is able to measure the degree of *meaningful* variation. The only two existing methods not to measure white noise as maximally complex are 'machado2015' and 'redies20212', though they still give it a high score. Instead, they give their max score to Stripes. This is also undesirable, because the simple repeating black and white stripes are not intuitively complex or meaningful either. Stripes is also given a high score by the fractal dimension and JPEG-ratio methods, both assigning it only slightly less than white noise and significantly more than any other dataset, including ImageNet. Our method agrees much more closely with the intuitive notion of complexity: it assigns the highest complexity to ImageNet; it puts CIFAR ahead of DTD2 even though the latter is of higher resolution and has a complex texture, which shows that it recognizes CIFAR to have more semantically meaningful content; and it assigns MNIST a reasonably high complexity, despite it being the smallest in terms of file size, again showing that it can recognize global structure. Even aside from the white noise, no method but ours correctly places the remaining six datasets in order of complexity (left-to-right, as they appear in Table 1). This highlights the superior ability of our method to capture meaningful complexity across a variety of image types.

Complexity at Different Scales The results from Section 4 suggest that, unlike existing methods, which focus only on detailed textures, ours is able to recognize complexity at a global level. Figure 2 provides further support for this claim by showing the breakdown of our complexity measure at the four different scales (that is, four different patch sizes; see Section 3.2). Smaller scales respond to local complexity, and as the process is iterated to larger scales, global structure can be detected. The first plot shows MNIST and our synthetic images. While MNIST has a similar local complexity score to Stripes, it has a much higher global complexity score, indicating that the more meaningful global structure in MNIST images can be detected. Halves, which is almost uniform locally but shows some variation globally, is given a very low local complexity but a small amount of global complexity. The second plot compares real-world images. CIFAR has the lowest local complexity

Table 1: Comparison of our method with existing methods. The figures for each dataset are the mean across all images from that dataset, with std. dev. from batches of 25 in parentheses. All methods are normalized, so the maximum score that they assign is 1. Ours is the only method that does not assign white noise images high complexity, and gives the most reasonable results on all other datasets.

	Dataset						
	ImageNet	**CIFAR**	**DTD2**	**MNIST**	**Stripes**	**Halves**	**white-noise**
ours	0.80 (.10)	0.74 (.06)	0.62 (.29)	0.50 (.08)	0.36 (.11)	0.26 (.01)	0.00 (.00)
khan2021	0.09 (.05)	0.01 (.01)	0.07 (.06)	0.00 (.00)	0.00 (.00)	0.00 (.00)	0.99 (.00)
machado2015	0.23 (.08)	0.15 (.02)	0.38 (.08)	0.21 (.01)	0.53 (.02)	0.06 (.00)	0.87 (.00)
redies2012	0.13 (.05)	0.04 (.01)	0.21 (.11)	0.00 (.00)	0.66 (.34)	0.01 (.00)	0.59 (.00)
entropy	0.89 (.10)	0.89 (.07)	0.83 (.13)	0.30 (.06)	0.13 (.00)	0.13 (.00)	0.96 (.00)
fractal dim.	0.74 (.09)	0.61 (.08)	0.86 (.16)	0.45 (.06)	0.98 (.02)	0.44 (.02)	1.00 (.00)
jpg-ratio	0.22 (.08)	0.09 (.0)	0.29 (.09)	0.06 (.01)	0.57 (.01)	0.06 (.00)	0.57 (.00)
GLCM	0.84 (.11)	0.80 (.08)	0.83 (.14)	0.27 (.05)	0.11 (.02)	0.08 (.00)	0.98 (.00)

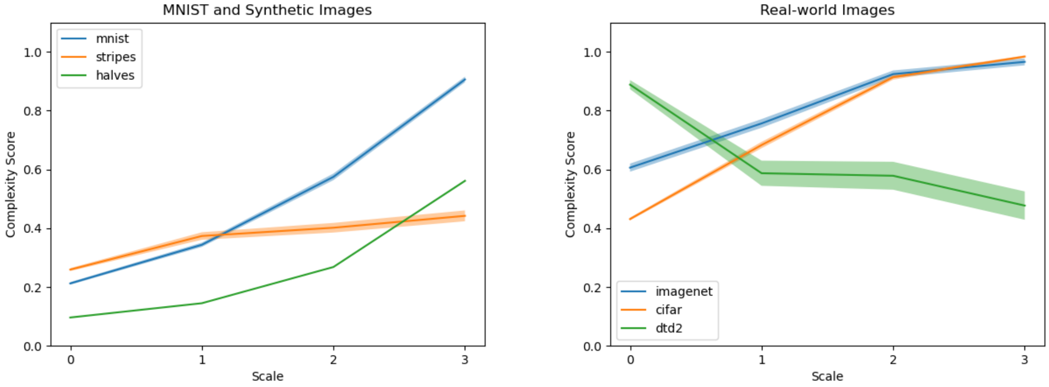


Figure 2: Our complexity measure for different scales. The x-axis depicts patch size, on a log scale. Plots show mean score for all images of that type. Shaded regions are std dev from batches of 25 images.

because it is low resolution, having been resized from 32×32 so neighbouring pixels are all similar, but this does not affect its global complexity, which is as high as that of Imagenet. DTD2, on the other hand, has the highest local complexity, because it depicts detailed textures, but the lowest global complexity, because the textures are uniform across different regions of the image.

Ablation Studies Table 2 shows the effect of removing two key components of our method. In 'no mdl', we fix the number of clusters to five for all images, rather than using the minimum description length principle. This results in the same problem that existing methods suffer from: white noise is mistaken for high complexity and receives the maximum score. Also, 'no mdl' scores DTD2 too highly, showing that the method is not responding to global structure. In 'no patch', we take the entropy not of patch signatures, but of individual points in the array, i.e., of A rather than B in the

Table 2: Effect of removing two main components of our method. In 'no mdl', clustering is performed without MDL, instead simply fixing the number of clusters to 5 for all images and all scales. In 'no patch', we compute the entropy of the clusters themselves rather than of the patch signatures.

	Dataset						
	ImageNet	**CIFAR**	**DTD2**	**MNIST**	**Stripes**	**Halves**	**white-noise**
main	0.80 (.10)	0.74 (.06)	0.62 (.29)	0.50 (.08)	0.36 (.11)	0.26 (.01)	0.00 (.00)
no mdl	0.73 (.09)	0.66 (.06)	0.90 (.11)	0.40 (.07)	0.35 (.13)	0.27 (.01)	0.98 (.00)
no patch	0.92 (.09)	94 (.04)	0.62 (.28)	0.61 (.1)	0.74 (.09)	0.50 (.01)	0.00 (.00)

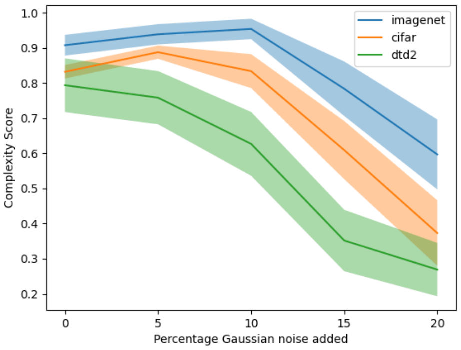


Table 3: Effect of adding Gaussian noise. Shaded regions are std. dev. from batches of 25 images.

Table 4: Effect of reducing image resolution of ImageNet images. Shown resolution is for both axes. Full resolution is 224.

	full	128	64	32
L1	7.70	7.01	6.32	5.49
L2	9.71	9.49	9.31	8.86
L3	11.93	11.69	11.75	11.56
L4	12.43	12.26	12.31	12.43
Total	41.77	40.46	39.69	38.33

terminology of Section 3.2. This setting still performs reasonably well, but it gives too high a score to Stripes and a higher score to CIFAR than to ImageNet.

Addition of Gaussian Noise As our method so consistently assigns zero complexity to white noise, one may wonder whether it just searches for randomness in the image, and assigns zero if it finds any. To check this, we progressively add Gaussian noise to the three real-world datasets. The results are shown in Figure 3. Noise is sampled independently from a standard Normal distribution for each pixel, and a fraction of this noise is added to the image. Up until 10%, the scores are largely unchanged (DTD drops slightly), and then the scores for all three datasets steadily decrease with further noise. If the method was simply assigning low complexity in response to any randomness in the image, then we would see a sharp decline as soon as a small amount of noise is added. The results suggest that the method is instead responding to the amount of meaningful content in the image. A gradual decline in complexity is precisely what we would expect as the image quality deteriorates.

Reducing Image Resolution The effect of resolution on our method is already somewhat apparent from Table 1. CIFAR gets a higher score than DTD despite being lower resolution, which suggests that our method is not responding to high resolution. Specifically, it is not just giving ImageNet a high score because its images are high-resolution. Here, we provide further support for this fact by directly manipulating the resolution of ImageNet images. Table 4 shows the score at four different reduced resolutions. Similar to the addition of Gaussian noise, a small reduction in resolution does not change the semantic contents, so should have little impact on complexity, and further reuctions should show a gradual decline as meaningful information begins to be lost. Additionally, we should see the largest effect on the lowest level of the hierarchy, as the higher levels do not respond to local detail anyway. This is precisely the case for our method. There is a slight drop on the lower levels, corresponding to greater uniformity at the local scale in the blurry, low-resolution images. The scores at the higher levels are essentially identical, and overall the scores are almost the same as for the full-resolution images. This shows our method to be robust to changes in resolution, responding more to the contents of the image than to the resolution it is depicted at.

5 Conclusion

This paper presented a new method for measuring image complexity. It uses clustering to analyse an image as a hierarchy of patches, with each patch composed of the cluster indices of its sub-patches. Clustering is pppenderformed with the minimum description length principle to distinguish signal from noise. We gave a derivation of our method, and presented experimental evaluation showing that it performs better than existing measures of image complexity. It assigns zero complexity to white noise, in contrast to existing methods which all assign white noise very high complexity. We also presented ablation studies, and further experiments showing our method captures complexity at different scales, and that it is robust to degradations in image quality.

References

George David Birkhoff. *Aesthetic Measure*. Harvard University Press, Cambridge, 1933.

Adrian Carballal, Carlos Fernandez-Lozano, Nereida Rodriguez-Fernandez, Iria Santos, and Juan Romero. Comparison of outlier-tolerant models for measuring visual complexity. *Entropy*, 22(4): 488, 2020.

Yi Chang, Luxin Yan, Tao Wu, and Sheng Zhong. Remote sensing image stripe noise removal: From image decomposition perspective. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7018–7031, 2016.

Chen Chen, Wei Li, Eric W. Tramel, Minshan Cui, Saurabh Prasad, and James E. Fowler. Spectral–spatial preprocessing using multihypothesis prediction for noise-robust hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(4):1047–1059, 2014.

Lydia Chioukh, Halim Boutayeb, Dominic Deslandes, and Ke Wu. Noise and sensitivity of harmonic radar architecture for remote sensing and detection of vital signs. *IEEE Transactions on Microwave Theory and Techniques*, 62(9):1847–1855, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Leroy Cronin, Natalio Krasnogor, Benjamin G. Davis, Cameron Alexander, Neil Robertson, Joachim H. G. Steinke, Sven L. M. Schroeder, Andrei N. Khlobystov, Geoff Cooper, Paul M. Gardner, et al. The imitation game — A computational chemical approach to recognizing life. *Nature Biotechnology*, 24(10):1203–1206, 2006.

Puhong Duan, Xudong Kang, Shutao Li, and Pedram Ghamisi. Noise-robust hyperspectral image classification via multi-scale total variation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(6):1948–1962, 2019.

Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, 2004.

Alex Forsythe, Gerry Mulhern, and Martin Sawey. Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior Research Methods*, 40(1):116–129, 2008.

Alex Forsythe, Marcos Nadal, Noel Sheehy, Camilo J. Cela-Conde, and Martin Sawey. Predicting beauty: Fractal dimension and visual complexity in art. *British Journal of Psychology*, 102(1): 49–70, 2011.

Wenzhun Huang, Shanwen Zhang, and Harry Haoxiang Wang. Efficient gan-based remote sensing image change detection under noise conditions. In *Proceedings of the International Conference on Image Processing and Capsule Networks*, pp. 1–8. Springer, 2020.

Marc M. Van Hulle. Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17(9):1903–1910, 2005.

Tariq M. Khan, Syed S. Naqvi, and Erik Meijering. Leveraging image complexity in macro-level neural network design for medical image segmentation. *arXiv preprint ArXiv:2112.11065*, 2021.

Nina Siu-Ngan Lam, Hong-lie Qiu, Dale A. Quattrochi, and Charles W. Emerson. An evaluation of fractal methods for characterizing image complexity. *Cartography and Geographic Information Science*, 29(1):25–35, 2002.

David A. Landgrebe and Erick Malaret. Noise in remote-sensing systems: The effect on classification error. *IEEE Transactions on Geoscience and Remote Sensing*, 24(2):294–300, 1986.

Penousal Machado, Juan Romero, Marcos Nadal, Antonino Santos, João Correia, and Adrián Carballal. Computerized measures of visual complexity. *Acta Psychologica*, 160:43–57, 2015.

Luis Madrid-Herrera, Mario I. Chacon-Murguia, Daniel A. Posada-Urrutia, and Juan A. Ramirez-Quintana. Human image complexity analysis using a fuzzy inference system. In *Proceedings of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6. IEEE, 2019.

Benoit Mandelbrot. How long is the coast of britain? Statistical self-similarity and fractional dimension. *Science*, 156(3775):636–638, 1967.

Manuela M. Marin and Helmut Leder. Examining complexity across domains: Relating subjective and objective measures of affective environmental scenes, paintings and music. *PloS One*, 8(8): e72412, 2013.

Stuart M. Marshall, Douglas Moore, Alastair R. G. Murray, Sara I. Walker, and Leroy Cronin. Quantifying the pathways to life using assembly spaces. *arXiv preprint ArXiv:1907.04649*, 2019.

Stuart M. Marshall, Cole Mathis, Emma Carrick, Graham Keenan, Geoffrey J. T. Cooper, Heather Graham, Matthew Craven, Piotr S. Gromski, Douglas G. Moore, Sara Walker, et al. Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nature Communications*, 12(1):1–9, 2021.

Fintan Nagle and Nilli Lavie. Predicting human complexity perception of real-world scenes. *Royal Society Open Science*, 7(5):191487, 2020.

Ram M. Narayanan, Sudhir K. Ponnappan, and Stephen E. Reichenbach. Effects of noise on the information content of remote sensing images. *Geocarto International*, 18(2):15–26, 2003.

Irina E. Nicolae and Mihai Ivanovici. Preparatory experiments regarding human brain perception and reasoning of image complexity for synthetic color fractal and natural texture images via eeg. *Applied Sciences*, 11(1):164, 2020.

Richard Alan Peters and Robin N. Strickland. Image complexity metrics for automatic target recognizers. In *Proceedings of the Automatic Target Recognizer System and Technology Conference*, pp. 1–17. Citeseer, 1990.

Georg Pichler, Pierre Jean A. Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. A differential entropy estimator for training neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 17691–17715. PMLR, 2022.

Behnood Rasti, Paul Scheunders, Pedram Ghamisi, Giorgio Licciardi, and Jocelyn Chanussot. Noise reduction in hyperspectral imagery: Overview and application. *Remote Sensing*, 10(3):482, 2018.

Christoph Redies, Seyed Ali Amirshahi, Michael Koch, and Joachim Denzler. Phog-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects. In *Proceedings of the European Conference on Computer Vision*, pp. 522–531. Springer, 2012.

Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

Edward W. Schwieterman, Nancy Y. Kiang, Mary N. Parenteau, Chester E. Harman, Shiladitya DasSarma, Theresa M. Fisher, Giada N. Arney, Hilairy E. Hartnett, Christopher T. Reinhard, Stephanie L. Olson, et al. Exoplanet biosignatures: A review of remotely detectable signs of life. *Astrobiology*, 18(6):663–708, 2018.

Bino Sebastian V., A. Unnikrishnan, and Kannan Balakrishnan. Gray level co-occurrence matrices: generalisation and some new features. *arXiv preprint ArXiv:1205.4831*, 2012.

Aleksandra Sherman, So Yum Lim, Marcia Grabowecky, and Satoru Suzuki. Visual-object working memory affects aesthetic judgments. *Journal of Vision*, 13(9):1308–1308, 2013.

Christian Stickel, Martin Ebner, and Andreas Holzinger. The xaos metric–understanding visual complexity as measure of usability. In *Proceedings of the Symposium of the Austrian HCI and Usability Engineering Group*, pp. 278–290. Springer, 2010.

W. Sun, G. Xu, P. Gong, and S. Liang. Fractal analysis of remotely sensed images: A review of methods and applications. *International Journal of Remote Sensing*, 27(22):4963–4990, 2006.

MTCAJ Thomas and A. Thomas Joy. *Elements of Information Theory*. Wiley-Interscience, 2006.

Xiao-Tian Wang, Wan-chao Ma, Kai Zhang, and Jie Yan. Complexity metric of infrared image for automatic target recognition. In *Proceedings of the 2018 3rd International Conference on Computational Intelligence and Applications (ICCIA)*, pp. 175–180. IEEE, 2018.

Xiaomei Yang and Chenghu Zhou. Analysis of the complexity of remote sensing image and its role on image classification. In *Proceedings of the IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No. 00CH37120)*, volume 5, pp. 2179–2181. IEEE, 2000.