Computer Vision and Image Understanding 195 (2020) 102949

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Visual complexity analysis using deep intermediate-layer features

Elham Saraee *, Mona Jalal, Margrit Betke

Boston University, Boston, MA, 02215, United States

ARTICLE INFO

Communicated by Nikos Paragios

Keywords:
Visual complexity
Convolutional layers
Deep neural network
Feature extraction
Convolutional neural network
Activation energy
Memorability
Object classification
Scene classification

ABSTRACT

In this paper, we focus on *visual complexity*, an image attribute that humans can subjectively evaluate based on the level of details in the image. We explore unsupervised information extraction from *intermediate* convolutional layers of deep neural networks to measure visual complexity. We derive an activation energy metric that combines convolutional layer activations to quantify visual complexity. To show the effectiveness of our proposed metric for various applications, we introduce Savoias, a visual complexity dataset that compromises of more than 1,400 images from seven diverse image categories (e.g., advertisement and interior design). We demonstrate high correlations of our deep neural network-based measure of visual complexity with human-curated ground-truth (GT) scores on various widely used network architectures, e.g., VGG16, ResNet-v2-152, and EfficientNet, and in networks trained on two classification tasks (object and scene classification). This result reveals that intermediate convolutional layers of deep neural networks carry information about the complexity of images that is meaningful to people. Furthermore, we show that our method of measuring visual complexity outperforms traditional methods on Savoias and two other state-of-the-art benchmark datasets. Moreover, we perform extensive analysis on the performance difference between our unsupervised method and a supervised method trained on the feature map, and show that by supervision, we can improve the prediction. Finally, we demonstrate that, within the context of a category, visually more complex images are also more memorable to human observers.

1. Introduction

In recent years, deep learning has revolutionized research in computer vision. The use of deep neural networks, in particular, has enabled the design of solutions to many computer vision tasks. Deep neural networks perform millions of computations in their hidden layers that transcend the usefulness of the network beyond the task they were originally designed for. Attempts to extract such information have mainly focused on two approaches, high-level semantic feature representations extracted from deep layers and re-use of pre-trained networks for transfer learning (Wang et al., 2015; Yosinski et al., 2014; Tzeng et al., 2015; Huh et al., 2016).

Feature extraction from intermediate convolutional layers, especially those related to attributes directly mapped to human perception, has been less explored (Zhang et al., 2016; Liu and Han, 2016; Simonyan et al., 2013; Li and Yu, 2015). As opposed to deep features extracted from fully-connected layers of a deep network, the use of features extracted from intermediate convolutional layers has three important advantages. First, they can be extracted straightforwardly, irrespective of the input image size or aspect ratio. Second, they carry spatial information corresponding to receptive fields of particular features in the local regions of an image. And third, specifically in intermediate layers, they are more transferable and less domain or task specific.

In this work, we contribute to the exploration of how information computed by intermediate hidden layers can be utilized. We consider the activations from convolutional layers of deep neural networks as "feature maps" and propose a metric that correlates with a certain visual attributes of images. Here we focus on visual complexity, an image attribute that humans can also subjectively quantify.

Visual complexity is a broad concept with decades of basic and applied research in a variety of areas such as computer vision and other areas of computing, psychophysics and cognitive psychology, product design, and marketing. Various definitions can be found in the literature. One definition relates visual complexity to the level of intricacy and details in an image, or the level of difficulty of a human observer to describe an image (Heaps and Handel, 1999; Snodgrass and Vanderwart, 1980). Another definition of visual complexity is based on the level of visual clutter and the amount of information conveyed in the image, which then relates the study of visual complexity to the fields of image compression and information theory (Rosenholtz et al., 2007).

Analysis of visual complexity is connected to a variety of problems in the field of computer vision. For example, visual clutter is a determining factor in measuring the difficulty of a visual search task (Ionescu et al., 2016). Other important computer vision tasks,

* Corresponding author.
E-mail address: esaraee@bu.edu (E. Saraee).

https://doi.org/10.1016/j.cviu.2020.102949
Received 14 May 2019; Received in revised form 9 January 2020; Accepted 11 March 2020
Available online 2 April 2020

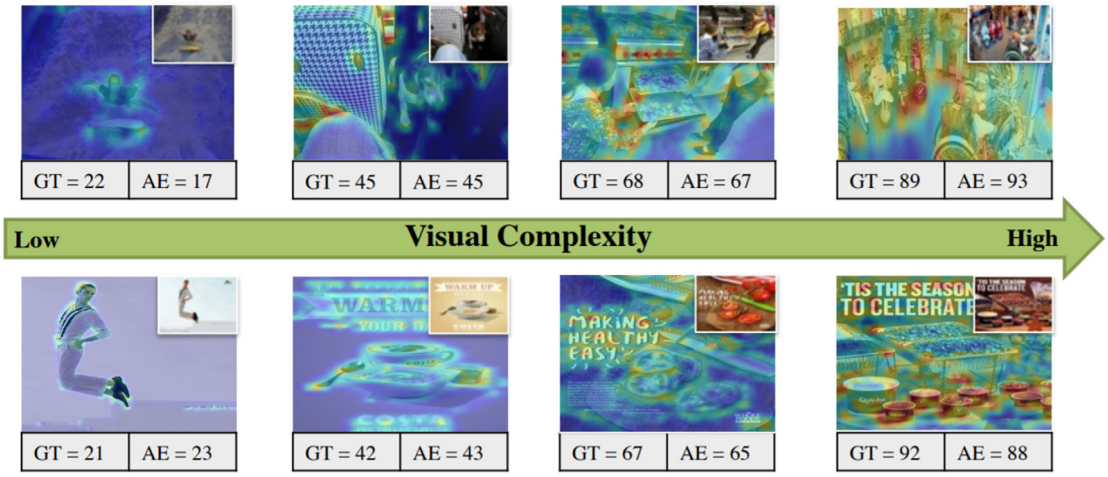


Fig. 1. Energy maps overlaid on sample images of the proposed SAVOIAS dataset along with the ground-truth (GT) visual complexity and our proposed *Unsupervised Activation Energy* method (UAE) scores. The top samples are selected from the "Scene" category, the bottom from the "Advertisement" category. AE, extracted from an intermediate convolutional layer of a deep network trained for visual analysis, correlates highly with GT complexity. This shows that, even at the intermediate layer, such networks carry information beyond the task they were originally designed for, and that this information correlates with what humans perceive as the complexity of an image.

such as automatically creating image captions, detecting objects, or segmenting object outlines, are particularly challenging for images of cluttered scenes with many objects that are partially occluded. Another topic is image memorability, which, as visual complexity, is an image attribute related to human perception. In this paper, we show that memorability can be estimated based on the visual complexity of the image. Furthermore, a visual question answering (VQA) algorithm may benefit from analysis of the complexity of image regions – a visually complex region is likely to need more algorithm-generated questions and answers, and a visual complexity map can guide the VQA towards where to look for questions and answers.

Understanding visual complexity of images is not only relevant to computer vision – it is also beneficial in the context of computer graphics and crowdsourcing. With regards to graphics, for example, the more complex a 3D scene is, the more time it takes for an algorithm to render it (Ramanarayanan et al., 2008a). With regards to crowdsourcing, evaluating the difficulty level of a vision task is essential for assigning the adequate number of internet workers to that task (Sameki et al., 2019). A measure of visual complexity can be used to estimate this difficulty level. Moreover, visual complexity carries significant information that can lead to solutions in other fields, including artwork, marketing, advertising, and web design (Machado et al., 2015; Ramanarayanan et al., 2008b; Reinecke et al., 2013).

For quantifying visual complexity, various factors have been studied in the literature. Image colorfulness, edge density, luminance, patterns, mirror symmetry, and number of objects are some of the examples (Mack and Oliva, 2004; Rosenholtz et al., 2007; Gartus and Leder, 2017). However, depending on the type of image, e.g., whether it contains abstract patterns versus real-world scenes, the contributions of these factors towards representing its complexity are different. Supervision is needed to tune the contribution of each of these factors. This requires access to additional information about the image data and can make a learning algorithm more prune to overfitting.

To overcome the need for tuning, we introduce the Unsupervised Activation Energy (UAE) method, which is based on analyzing the activations in the intermediate convolutional layers of a deep neural network. We show a few example images with their UAE-predicted visual complexity scores as well as human-curated ground-truth scores in Fig. 1. The overlaid energy maps visualize the image regions which contribute to the UAE-predicted visual complexity scores.

To showcase the applicability of our method to different network architectures for predicting visual complexity, we include the VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), Inception (Szegedy et al., 2017), DenseNet (Huang et al., 2017), MobileNet (Howard et al., 2017), and the state-of-the-art EfficientNet (Tan and Le, 2019) network architectures in our experiments. In addition, to demonstrate that the proposed method can predict visual complexity from the layers that are not task or domain specific, we perform experiments for two tasks, object classification, with network pretrained on the ImageNet dataset (Krizhevsky et al., 2012), and scene recognition, with network pretrained on the Places dataset (Zhou et al., 2017). It has recently been shown that the structure of a generator network is sufficient to capture a great deal of low-level image statistics prior to any learning (Ulyanov et al.). To examine whether such a statement holds for our application with classification networks, we further explored the use of untrained network architectures to predict of image visual complexity.

Due to the wide range of applications for the analysis of visual complexity, the study of visual complexity requires adequate amounts of data for different types of images. Although some image datasets with ground-truth labels for visual complexity exist, they are either very small or lack diversity (both in number and type of categories, as well as diversity of image content and appearance within each category). In addition, the methodologies to obtain the ground truth for these datasets are not consistent and in most cases, they are based on a limited-point Likert scale. To overcome these limitations, we here introduce Savoias, a new dataset for the analysis of visual complexity. Savoias covers a variety of topics and provides a sufficient number of images per topic, therefore improving diversity and scale of publicly available datasets. Specifically, Savoias consists of seven diverse categories and a total of 1,420 images. Savoias is an acronym for **S**cenes, **A**dvertisement, **V**isualization and infographics, **O**bjects, **I**nterior design, **A**rt, and **S**uprematism (a category of art).

In order to minimize the potential bias from individual ground-truth contributors and limited rating scales, we obtained the ground-truth labels using a forced-choice pairwise crowdsourcing methodology. Labels were obtained from 1,687 contributors on more than 37,000 pairs of images. The pairwise scores were then converted to absolute scores on a [0,100] rating scale using the Bradley–Terry (Bradley and Terry, 1952) and matrix completion (Candès and Recht, 2009) methods.

Our results show the superiority of the proposed UAE method in quantifying visual complexity compared to existing methods. Moreover, we perform analysis to compare the unsupervised method with a supervised method trained on the same feature maps extracted from convolutional layers of the deep neural network and show that improvement can be made by training a supervised model on the feature

maps. Finally, we investigate the relationship between visual complexity and image memorability and show that within the context of scene categories, image categories that are more visually complex are also more memorable.

To summarize, in this paper, we make the following contributions:

- We investigate the unsupervised extraction of information from convolutional layers of neural networks with regard to visual complexity and devise the *Unsupervised Activation Energy* (UAE) method.
- We introduce SAVOIAS, a publicly-available dataset for the analysis of visual complexity. We obtained detailed ground-truth labels (a [0,100] scale) using a forced-choice crowdsourcing methodology involving more than 37,000 comparisons.
- We show that scores obtained by our UAE method correlate with the qualitative measurements of visual complexity based on human perception. Our UAE method outperforms all previous work about the analysis of visual complexity on SAVOIAS and two other datasets.
- We compare the performance of our unsupervised method with the supervised Activation Energy (SAE) method trained on the same feature maps extracted from convolutional layers of a deep neural network and show that the prediction results can be improved by supervision. To the best of our knowledge, there is no publicly available supervised approach detailed in the literature to which we could compare our supervised method.
- We show that, within the context of a category, visually more complex images are also more memorable to human observers.

The rest of the paper is organized as follows: We review the previous work in Section 2. Our proposed methodology is explained in Section 3. Section 4 describes our proposed dataset. Our experiments are discussed in Section 5. In Section 6, we compare the unsupervised approach (UAE) with our supervised approach (SAE). In Section 7, we investigate the relationship between image visual complexity and image memorability as an application of our method. Finally in Section 8, we discussion our conclusions.

2. Related work

Visual complexity has been studied extensively in various fields such as psychophysics, cognitive science, and more recently in computer vision. While the temporal dimension of complexity is an interesting topic of research (Cardaci et al., 2009; Marin and Leder, 2016; Palumbo et al., 2014), in this work, we focus on the spatial dimension of visual complexity and algorithmic approaches to quantify it.

In this section, we begin with discussing the applications of the visual complexity in other domains. We then review the previous work focused on quantifying visual complexity on different types of images and explain how our work is related to other problems such as saliency and image compression. Lastly, we summarize couple of works that study feature extraction from convolutional layers of deep neural network for different vision tasks.

2.1. Visual complexity in other fields

Visual complexity and aesthetic beauty have been shown to be related (Birkhoff, 1933; Eysenck, 1941; Jacobsen and Höfel, 2001; Reinecke et al., 2013). In one of the earlier works on visual complexity, (Birkhoff, 1933) defines aesthetic measure M of an art object as a function of the ratio between its order and complexity (M=f(O/C)), where O stands for order, often found in the forms of harmony or symmetry and C stands for complexity. Based on his formula, the aesthetic measure (M) of an art object decreases with increase in complexity.

Furthermore, Visual complexity has been found to be a dominant factor in determining the pleasingness of a stimulus and to be related to aesthetic preference for artistic works (Forsythe et al., 2011). It has been shown that the relation between visual complexity and preference follows an inverted U-curve, in which images with intermediate levels of visual complexity are most appealing (Berlyne, 1971). The perception of appeal of a web page design has been shown to have a connection to the visual complexity of the web page (Reinecke et al., 2013), and thus, understanding the visual complexity of a design can lead to a better subjective experience for viewers (Krishen, 2008).

Studies show that in hedonic shopping experiences, shoppers are more satisfied with mall interiors that have a higher perceived complexity, while in utilitarian shopping experiences, shoppers prefer lower visual complexity (Haytko and Baker, 2004). For online shopping, the perceived visual complexity has been shown to negatively influence individuals' satisfaction (Sohn et al., 2017).

Visual impression of advertisement images plays a crucial role in engaging viewers. Two different types of perceptual complexity have been considered for advertisement: feature complexity, which depends on image features, and design complexity, which depends on the creative design of the image. It has been argued that feature complexity hurts attention to the brand, whereas design complexity can improve the consumer's attention (Pieters et al., 2010).

2.2. Algorithms to quantify visual complexity

In one of the early works on visual complexity, Chipman (1977) explained the importance of two factors in the analysis of visual complexity of simple abstract patterns – a quantitative factor (related to amount of elements), which correlates positively with visual complexity, and a structural element (determined by different forms of structural organization, but mostly by symmetry), which correlates negatively with visual complexity (see also, Chipman and Mendelson (1979)). Following a similar approach, recent studies explored the impact of these two factors on the visual complexity of complex abstract patterns (Gartus and Leder, 2013, 2017) and images in four image categories (abstract artistic or non-artistic and representational artistic or non-artistic), which included three complexity levels (low, intermediate, and high) (Nadal et al., 2010).

Fan et al. (2017) introduced three features, color richness, stroke thickness, and white spaces, to evaluate the visual complexity of Chinese ink paintings. Miniukovich and Angeli (2014) proposed the five factors visual clutter, symmetry, contour density, figure-ground contrast, and color variability to encapsulate visual complexity for web design and GUI applications.

Oliva et al. (2004) studied the role of a task constraint on representation of visual complexity. They argued that although the contribution of the perceptual dimensions are affected by task constraints, visual complexity can still be represented by perceptual dimensions such as quantity of objects, clutter, openness, symmetry, organization, and variety of colors. The notion of clutter can be captured by measuring the density of edges in an image (Mack and Oliva, 2004).

Visual complexity is also related to image saliency. One of the common measures to predict visual complexity is feature congestion, which is derived from an image saliency model. This measure can be described by the analogy that it is more difficult to add an attention-grabbing component to a visual interface that is already very busy than to an interface that is not (Rosenholtz et al., 2007). To evaluate the relation between attention and image complexity, Da Silva et al. (2011) studied the correlation between human eye movements as well as different models of computational attention against a ground-truth of image complexity. Their experimental results, confirmed the existence of such correlation between attentional behavior and image visual complexity. Moreover, to predict the visual complexity of paintings, it has been argued that in addition to the global and local features, salient region features should be considered (Guo et al., 2018). The purpose of the global and local features are to capture the characteristics of the first impression of the viewer and the regional information of the painting, respectively; while the salient region features represent the

Table 1
The datasets previously used in visual complexity studies, as well as our proposed dataset. A "1-step" process means that the users were asked to directly rate the visual complexity of a single image; "Shared" means the dataset may be shared with other researchers upon request. *Sample images of these datasets are shown in Figures 2, 8, and 9.

Reference	# Images	Application category	Ground-truth process	Vis. Comp.scale	Opensource
(Gartus and Leder, 2013)*	912	Black and white 8 × 8 abstract patterns	1-step	5-point	Shared
(Nadal et al., 2010)	120	Abstract & representational (artistic & non-artistic)	1-step	3-point	No
(Oliva et al., 2004)	100	Indoor scenes	3-step	8-point	No
(Miniukovich and Angeli, 2014)	142	Webpage	1-step	5-point	Yes
(Fan et al., 2017)	40	Chinese ink painting (abstract and representational)	1-step	7-point	Shared
(Schnur et al., 2018)	9	Web maps	1-step	5-point	No
(Corchs et al., 2016)(RSIVL)*	98	Real-world scenes	1-step	[0-100]	Shared
(Corchs et al., 2016)	122	Textures	1-step	[0-100]	Shared
Ours*	**1,420**	**Scenes, advertisement, visualizations, objects, interior design, art, Suprematism**	**Pairwise comparison**	**[0–100]**	**Yes**

characteristics of the most visually important region of a painting for the viewer.

Visual complexity can be approximated using compression algorithms (Rosenholtz et al., 2007), defining it by the resulting file size when a compression algorithm such as JPEG or ZIP is applied to a given image. One of the common methods to quantify visual complexity from information-theoretic perspective is the subband entropy measure (Rosenholtz et al., 2007). This algorithm is derived based on the notion that clutter is related to the number of bits required for subband (wavelet) image coding. Recently compression algorithms based on deep neural network and recurrent neural networks have shown to provide promising results in terms of subjective quality of compressed images (Liu et al.; Toderici et al., 2017), when evaluated by measures inspired by human visual system (Wang et al., 2003; Gupta et al., 2011). For example, a combination of a perception loss and an adversarial loss is proposed to train a deep image compression model (Liu et al.). Rather than calculating distortions directly in pixel domain, these loss functions measure the similarity in high-level feature domain. As a result, this compression algorithm can better mimic the discriminative characteristics of the human visual system.

A linear combination of multiple features such as edge density, compression ratio, and number of objects has also been studied (Corchs et al., 2016). However, for combining the features, a supervised algorithm is required to set the weights assigned to each of these features. The stimuli used in these experiments were real scene or texture images.

In another study (Schnur et al., 2018), visual complexity of the interfaces of three different online map providers (Google Maps, Bing Maps, and OpenStreetMaps) was explored with the objective to better understand design decisions for Web maps. The study results imply that clutter, measured by feature congestion, is more important in perceived complexity than diversity of symbology.

Here, we discussed image features that contribute to visual complexity. We note that a single feature cannot perform well for all possible types of images. Supervision is required to combine various features in measuring visual complexity and adjust their contributions for a specific image category. Thus, previous multidimensional models are mostly designed to predict the visual complexity of images in a specific image category. To address the problem of generalization, here, we propose an unsupervised method to quantify the visual complexity of images. Since our method inherently encompasses different types of features, no further supervision is required. It can thus quantify the visual complexity of different types of images.

2.3. Datasets

The characteristics of several visual complexity datasets are summarized in Table 1. The table shows differences in scales and image collection methods, as well as our approach to establish ground truth, and reveals that the diversity and number of samples in these datasets are in need of improvement for extensive analysis of visual complexity. It is also worth noting that not all the datasets mentioned above are publicly available or shared among researchers.

Savoias is introduced to address the lack of an appropriate dataset with an adequate number of diverse images and a reliable and fine-grained ground-truth value for visual complexity. Savoias is a new dataset, consisting of more than 1,400 sample images that are organized in seven categories, selected to show a great deal of variety. In order to obtain ground truth values for the visual complexity of these images, instead of asking the participants to rate the visual complexity of a particular image on a continuous scale, we asked them to compare the visual complexity of two images. The pairwise comparison methodology helps avoid rating biases and also provides a more fine-grained range of scores compared to the common 3, 5, or 7-scale ratings used in the aforementioned datasets.

2.4. Feature extraction from convolutional layers of deep neural network

The use of convolutional layers of deep neural networks to provide local image descriptors, or auxiliary features to represent edges or textures has been shown to produce state-of-the-art results in various application such as semantic image segmentation, super-resolution and image retrieval (Cimpoi et al., 2016; Liu et al., 2015; Babenko and Lempitsky, 2015; Razavian et al., 2016; Ng et al., 2015; Yang et al.; Paulin et al., 2017; Uricchio et al.; Gordo et al., 2017).

It has been shown that the intermediate CNN features can be utilized in the semantic image segmentation task to produce task-specific edges in an end-to-end trainable system (Chen et al., 2016). Such systems can be an alternative to the fully-connected conditional random fields to enhance their object localization accuracy, while being less computationally expensive. Additionally, to recover the fine texture details of an image in a super-resolution problem, the use of feature maps from a deep neural network is proposed (Ledig et al., 2017). In this work, in addition to the adversarial loss, the authors introduced a second content loss based on the high-level feature map to represent the perceptual similarity instead of similarity in pixel space. The activations from the convolutional layers of a deep neural network have also been utilized to encode images into compact global signatures for the instance-level image retrieval task (Gordo et al., 2016).

Activated features of deep neural networks can also be incorporated in the context of visual similarity and image quality assessment (Amirshahi et al., 2017; Zhang et al., 2018; Gao et al., 2017; Kim and Lee, 2017). Gao et al. (2017) and Amirshahi et al. (2017) propose techniques involving leveraging internal activations of deep networks for image quality assessment. To evaluate the applicability of deep features in assessing the human perceptual similarity judgments, Zhang et al. (2018) conducted a study across different architectures. They concluded that networks trained to solve visual prediction and modeling tasks are capable of learning a representation of the world that correlates well with perceptual judgments.

Aligned with the aforementioned works, we focus on the activated features in the deep neural networks. In our work, however, we examine multiple architectures and tasks, and we focus on extracting

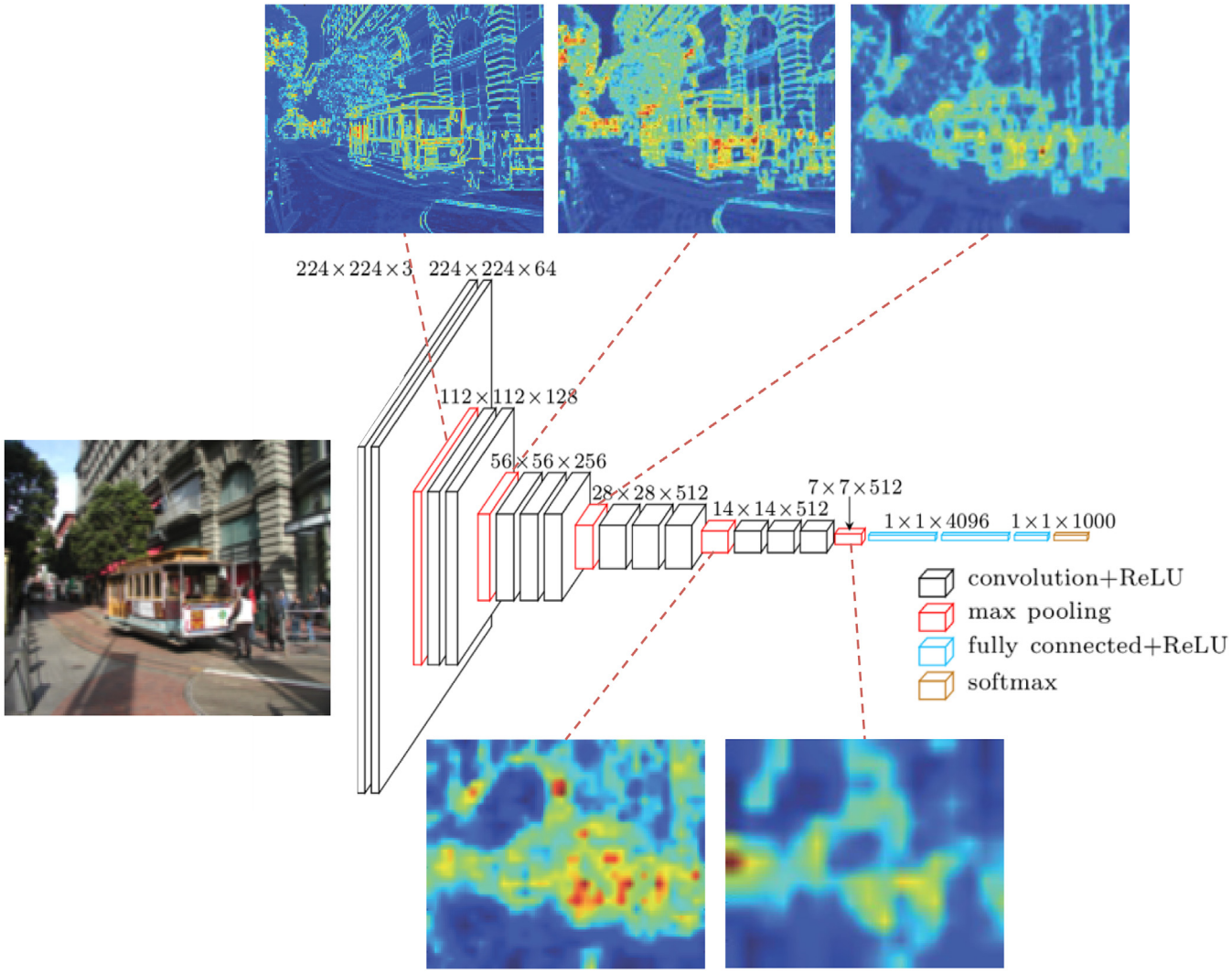


Fig. 2. Visualization of the activations from five max-pooling layers of the VGG architecture (Simonyan and Zisserman, 2014) trained for scene recognition. Each visualization shows what type of features are activated in each layer, by indicating the average of activated neurons for the corresponding spatial coordinate (high activation red, low activation blue). Image edges are best represented by the output of the first pooling layer; high-level features can be extracted from the output of the fifth (last) pooling layer. The image is taken from the RSIVL dataset (RSIVL, 2016).

descriptors as *standalone* features to predict visual complexity, an image attribute that can be directly mapped to a quality of an image that humans can quantify.

3. Approach

Deep neural networks consist of multiple layers, each responsible to activate different types of features. Here we investigate how we can extract visual complexity information from the feature maps of these layers.

3.1. Unsupervised activation energy (UAE) method

In order to quantify visual complexity, we devise a simple metric based on the feature maps of neural network layers. We define our metric for each layer simply as the average over all values of the receptive fields and all the channels in a layer, i.e.,

$$UAE(l) = \frac{1}{h \times w \times d} \sum_{i=1}^{h} \sum_{j=1}^{w} \sum_{k=1}^{d} F_l[i, j, k], \tag{1}$$

where F is the feature map, l is the layer number, and h, w, and d are height, width, and depth (number of channels) of the feature map, respectively. Note that, in the deep networks we consider, the extracted feature maps are outputs of rectified linear units, ensuring that the values in the feature maps are all non-negative, and, thus, *UAE(l)* is also non-negative. We refer to the feature map of a given layer, averaged over the channels and thus preserving the spatial information of the activated features, as its *energy map*:

$$Energy\ map_l[i, j] = \frac{1}{d} \sum_{k=1}^{d} F_l[i, j, k]. \tag{2}$$

Examples of energy maps, computed for different convolutional layers in the VGG architecture, are shown in Fig. 2.

3.2. Utility of different convolutional layers in quantifying visual complexity

We measure the Pearson correlation coefficient (PCC) between feature maps of an image in different convolutional layers of a deep neural network and the human-curated score of visual complexity of the image. Such a correlation can be interpreted as a direct measurement of how much information a specific layer carries with respect to the visual complexity of the input image.

To evaluate the correlation between our unsupervised activation energy method and human-perceived visual complexity, measured by the

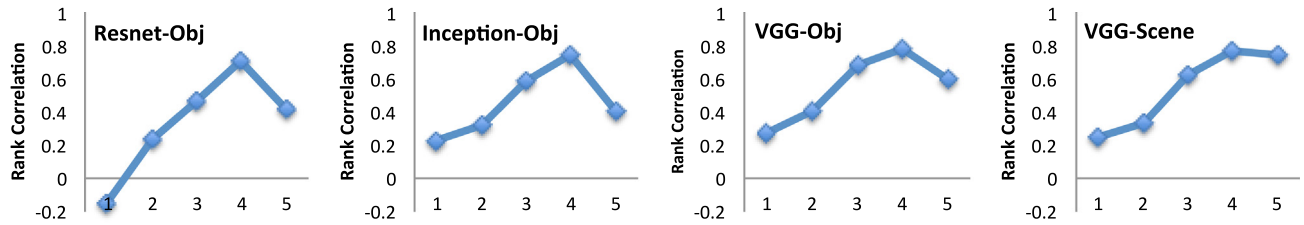


Fig. 3. Pearson correlation between the *Unsupervised Activation Energy* of an output layer and the human-curated visual complexity score as a function of layer number (1,..,5). The correlation for models trained for object classification with the architectures ResNet, Inception, and VGG are shown in the first three panels. The graphs in the last two panels are presented here to enable visual comparison of the results for the same architecture (VGG) for two different classification tasks, namely, object classification and scene recognition. The four graphs have an arch shape. The top of the arch indicates a significant correlation (up to 0.77) between human-curated visual complexity and the unsupervised activation energy of intermediate layers for these examined cases. These results are obtained for the *Scene* category of SAVOIAS dataset. Similar results obtained from other architectures pretrained on Imagenet are not presented for brevity.

Pearson correlation coefficient, we examine multiple widely-used network architectures, namely, VGG-16 (Simonyan and Zisserman, 2014), ResNet-v2-152 (He et al., 2016), Inception-v4 (Szegedy et al., 2017), DenseNet (Huang et al., 2017), MobileNet (Howard et al., 2017), and the state-of-the-art EfficientNet (Tan and Le, 2019). We selected five convolutional layers for each architecture for our study, such that activated neurons from different layers of the deep architectures are evaluated in our analysis. All these networks are pre-trained on the ImageNet dataset.

In addition, to further investigate how different models pretrained for different tasks correlate with visual complexity, we evaluate VGG-16 when pretrained on the Places dataset as well as the untrained network and the results are compared against the ImageNet dataset.

Deep neural networks learn their task by first extracting low-level features (edges, corners) in the early layers. These lines are then convolved to form higher-level features, and the content of the image is extracted in the deeper layers. As known from previous studies, various image factors, including both low-level features (e.g., edge density or patterns) and high-level features (e.g., number of objects or image content), impact visual complexity, and, thus, simply quantifying one type of features is not an adequate method for determining the visual complexity of an image.

Our analysis revealed that the UAE scores of the first two and the last layers have lower correlations with the human-curated scores than the UAE scores of the intermediate layers (Fig. 3). This observation confirms that the first layers, mostly representing edges, corners, etc., lack the kind of higher-level information needed to capture the concept of visual complexity. Similarly, the last layer of models trained for object classification does not yield consistently high correlations either, showing that although the existence of objects in an image affects its visual complexity, that alone is not sufficient information for evaluating visual complexity adequately. The intermediate convolutional layers, however, in all four models, show significantly higher correlation, and, therefore, are suitable representatives of visual complexity. The high correlation values we obtained reaffirm the importance of both lower- and higher-level features (here, output of layers 3 and 4) for evaluation of visual complexity.

Here we argue, in the transition from the low-level features to higher-level features in the convolutional layers of a deep neural network, there are layers that carry information about both types of features, and, thus, we can extract metrics from these layers that correlate with visual complexity. Further analysis will be provided in Section 5 after our dataset is introduced.

4. Dataset description

In this section, we introduce SAVOIAS, our visual complexity dataset. To date, SAVOIAS is the largest and most diverse open-source visual complexity dataset with 1,420 images in seven categories. In this section, we will first describe our image collection process and the different categories that we have used in our dataset. Next, we will explain the data annotation process.

4.1. Image collection

In our dataset, we have used images from seven diverse categories. Examples from each category in the increasing order of visual complexity are shown in Table 2. Despite the connection between the problem of visual complexity and other problems in computer vision, there exists a gap between these topics. To overcome this gap, we have randomly sampled images from commonly-used datasets with the following categories:

Advertisement: 200 images from the Advertisement dataset (Hussain et al., 2017). Visual impression of advertisement plays a crucial role in economical competitions (Pilelienė and Grigaliūnaitė, 2018). This category is selected in order to give ad designers insight into what factors impact the perceived complexity of an advertisement.

Objects: 200 images from the MSCOCO dataset (Lin et al., 2014). The purpose of this category is to understand how a human perceives the visual complexity of various objects and combination of objects. The number of objects is one of the leading factors contributing to the visual complexity of an image. This category can help researchers study the impact of characteristics of objects as well as the number of objects and their interaction with each other on visual complexity.

Scene: 200 images from the Places2 dataset (Zhou et al., 2017). The purpose of this category is to understand how humans perceive visual complexity of various scenes. It may facilitate the study of the roles of the image foreground and background in visual complexity analysis.

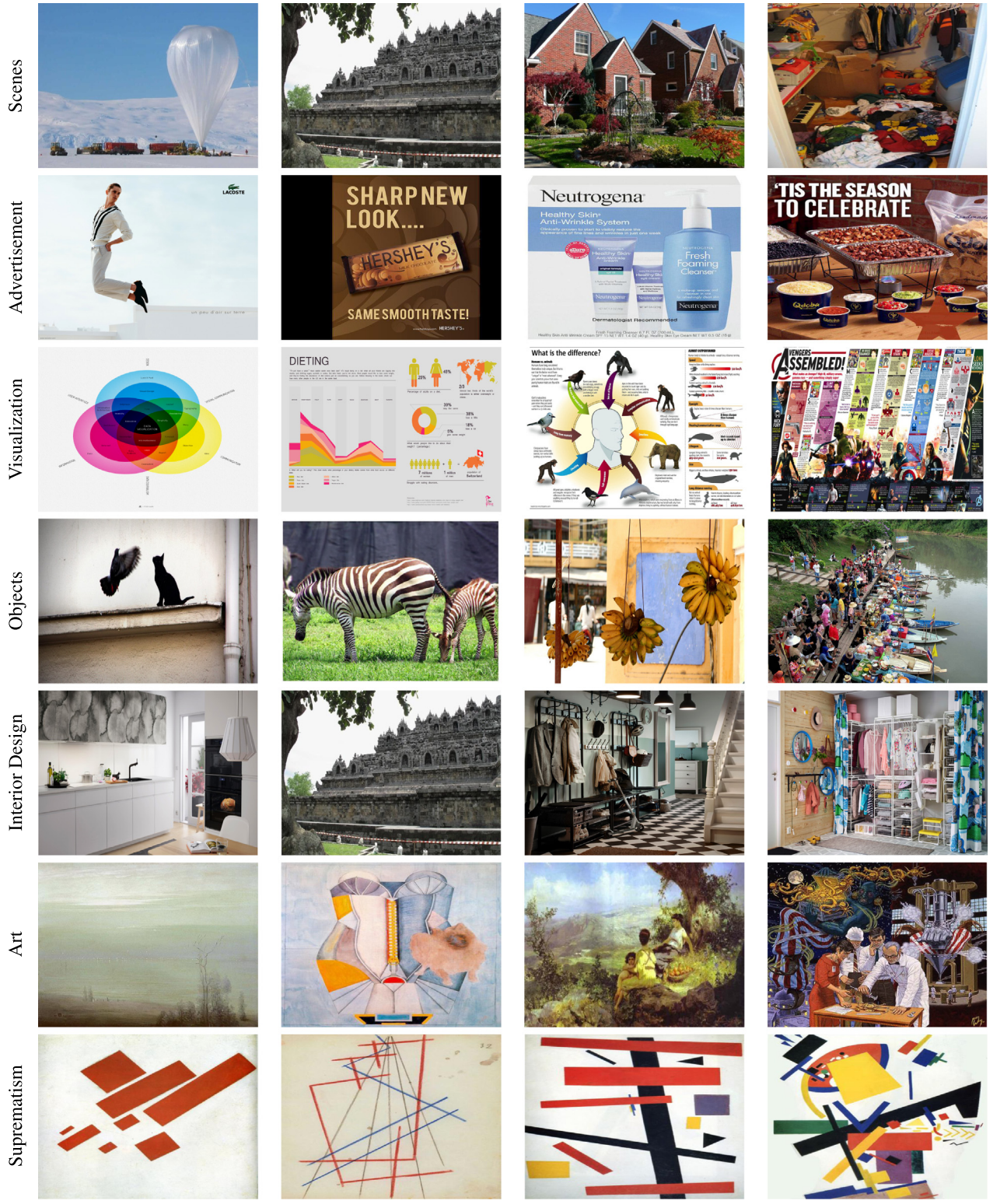
Interior Design: We have collected 100 interior design images from the IKEA website (IKEA, 0000). This category is specifically selected to provide insight into how humans perceive the visual complexity of indoor spaces at home such as bedroom, living room, dining room, kitchen, and bathroom. Interior designers may use visual complexity analysis to understand how to design appealing interior spaces.

Visualization and Infographics: 200 images from the MASSVIS dataset (Borkin et al., 2013). The category consists of charts, graphs, texts, and tables. Understanding the impact of visual elements as well as their composition may lead to an understanding of the cognitive and perceptual processing of a visualization, which can greatly influence the memorability, recognition, and comprehension of these designs.

Art: 420 Artistic images from the PeopleArt dataset (Westlake et al.). This category consists of 10 images from each of the 42 categories of art styles and movements (e.g., Naturalism, Cubism, Socialist Realism, Impressionism, and Suprematism). Since the aesthetic beauty of an artistic image is directly influenced by the level of its visual complexity (Eysenck, 1941; Reinecke et al., 2013), understanding the visual complexity of an artistic image can help artists to create more engaging artworks.

Suprematism: 100 images from the Suprematism category in the PeopleArt dataset for the analysis of geometric abstract art. The Suprematism category conveys various geometric shapes and objects in abstract form and thus presents a different challenge as it contains types of images that we do not commonly see. This category enables studying the impact of various shapes, geometric objects, and composition on the perception of visual complexity.

Table 2
Sample images of the SAVOIAS dataset with increased visual complexity from left to right in each row.



4.2. Data annotation using pairwise comparisons

In order to obtain absolute ranking scores for an attribute of an image, in our case, visual complexity, one approach would be to ask users to assign a score to each image, where the score represents the ranking of the image relative to all other images. However, it has been shown that most people can only evaluate 5 to 9 options at a time. In addition, bias in the rating scale is a common problem in this type of establishing ground truth (Miller, 1956).

The use of pairwise comparison and conversion of the pairwise ranking to global ranking is a better alternative (Arrow, 1950; David, 1963; Kendall and Smith, 1940). Pairwise comparison is a relative measure that helps reduce bias from the rating scale. It is also invariant under monotone transformation of the rating values and depends only

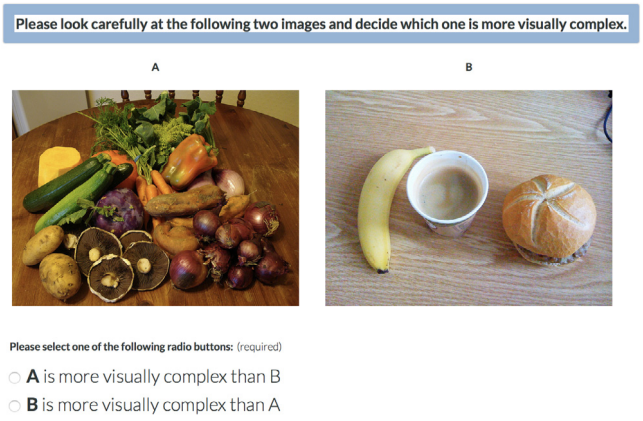


Fig. 4. Screenshot of one of the pairwise comparisons shown to Figure-Eight crowdsourcing platform contributors for the objects category.

on the degree of relative difference between one option over the other in the pair (Gleich and Lim).

Note that, for example, for a set of $n = 200$ images, including all of the pairs would result in $\binom{n}{2} = 19{,}900$ comparisons. However, it is shown that for the pairwise comparisons, not all of the pairs are required in order to obtain the final global ranking, and information about a small fraction of the pairs, $\ell \ll \binom{n}{2}$, is adequate (Chang et al., 2016). In many practical applications with partially observed measurements or budget constraints (e.g. (Kim et al., 2017), it is possible to use matrix completion methods in order to complement the results (Candès and Recht, 2009; Gleich and Lim).

In this work, we follow the pairwise comparison approach. Our algorithm is iterative and selects two images randomly from the set of images in a particular category in each step. Images that have been selected in previous steps are less probable to be chosen in subsequent steps. The algorithm terminates once a target number of comparisons is reached. We decided on different target numbers for different categories, assuming that the visual complexity of images in some categories is easier to evaluate by human judgment than in others. For the categories scenes, advertisement, visualization, and objects we decided to run our algorithm until $\ell = 4{,}000$ pairs are found, which results in 40 comparisons per image, on average, given that these categories have $n = 200$ images each. For the interior design category, we ran the algorithm until $\ell = 2{,}000$ pairs are found, also resulting in 40 comparisons per image, on average. For the art category, we obtain $\ell = 14{,}700$ pairs, which results in 70 comparisons per image, on average. Finally, for the Suprematism category, we used all possible $\ell = \binom{n}{2} = 4{,}950$ pairs, which resulted in 99 comparisons per image. In 4.3.1, we further discuss how we evaluated the accuracy of the absolute visual complexity scores as a function of ℓ, the number of pairwise comparisons between images for the Suprematism category.

4.2.1. Crowdsourcing

To minimize potential bias caused by specific raters, we collected human judgments via crowdsourcing. Our study involved more than 1,687 contributors who were recruited and paid through the Figure-Eight crowdsourcing platform.[1] Each task was distributed to five contributors. Contributors (also known as crowdworkers in Amazon Mechanical Turk) were shown ten pairs of images per page and asked which of the two images in each pair was visually more complex. We explained visual complexity by attributes such as cluttered background, numerosity and variety of objects, people, textures, patterns, and shapes. We used a forced-choice methodology, in which the contributors are supposed to select either image A or image B (Fig. 4). In the case of similar complexity, contributors were requested to select intuitively which image they considered more visually complex.

The study contributors were selected from a pool of "level-3 contributors" who had produced accurate answers in previous work (level 3 is the highest level of expertise on the Figure-Eight crowdsourcing platform). Contributors were not restricted by their locality. Each contributor was shown 10 pairwise comparison tasks per page. For each page, a contributor was paid $0.10. We did not allow any contributor to perform more than 300 tasks, but did not select a lower bound on the number of tasks.

Test questions, geared towards quality control, were distributed to contributors randomly throughout the entire time they performed the comparison tasks. While we paid all comparison tasks, we only kept the comparison labels provided by the contributors who maintained a passing score of 90% or above on test questions.

4.2.2. Conversion of pairwise scores to absolute scores

After we collected the information about which image, among a pair, is considered more complex, we needed to convert this pairwise score into an absolute visual complexity score. In order to convert pairwise ranking of images to global ranking, we applied two separate approaches, namely the Bradley–Terry method (Bradley and Terry, 1952) and matrix completion (Candès and Recht, 2009), as described below.

We denote the pairwise comparison matrix as a count matrix $S = \{s_{i,j}\}$, where $s_{i,j}$ is the ratio of the number of times that the contributors have selected image i as more visually complex compared to image j over the total number of times that images i and j have been compared. Thus, $s_{i,j} + s_{j,i} = 1$. The problem here is to find the absolute score c_i of image i.

The **Bradley–Terry method** (Bradley and Terry, 1952) describes the probability of choosing image I_i over image I_j as a Sigmoid function of the score difference between the two images,

$$P(I_i > I_j) = F(\Delta_{i,j}) = \frac{e^{\Delta_{i,j}}}{1 + e^{\Delta_{i,j}}}, \tag{3}$$

where $\Delta_{i,j} = c_i - c_j$. The score parameter c can be estimated by solving a maximum a posteriori (MAP) problem, i.e., maximizing

$$\log Pr(S|\, c) = \Sigma_{i,j} s_{i,j} F(\Delta_{i,j}), \tag{4}$$

where the prior is a uniform distribution. This optimization problem can be solved using gradient descent (Chang et al., 2016).

The **Matrix Completion method** assumes, if $s_{i,j}$ is greater than 0.5 (image i is more visually complex than image j), and $s_{j,k}$ is greater than 0.5, and the pairwise comparison between image i and k is missing, by using image j as a link, we can infer that image $s_{i,k}$ is also greater than 0.5. Now we can create matrix $\hat{S}$ by filling the missing elements of matrix S:

$$\hat{s}_{i,k} = \begin{cases} s_{i,k} & \text{if } s_{i,k} \in S \\ \frac{1}{m}\Sigma_{j=1}^{m} \frac{s_{i,j}+s_{j,k}}{2} & \text{else if } s_{i,j} \in S,\ s_{j,k} \in S \\ & \text{and } s_{i,j} > 0.5, s_{j,k} > 0.5 \end{cases} \tag{5}$$

where m is the number of existing pairs of $s_{i,j}$ and $s_{j,k}$ in the count matrix S. For matrix completion, note the following points:

- We only consider $s_{i,j}$ and $s_{j,k} \in S$ if they are greater than 0.5. Therefore, if $s_{i,j} > 0.5$ and $s_{j,k} < 0.5$, we will not make any judgments about the missing pair $s_{i,k}$.
- For those pairs for which we have the result in one direction, we can fill the matrix in the other direction by using $\hat{s}_{k,i} = 1 - \hat{s}_{i,k}$.
- For the rare case that a pair is not connected in either directions, we use $\hat{s}_{i,k} = \hat{s}_{k,i} = 0.5$.

When the count matrix $\hat{S}$ is completed, the absolute score for each image is the mean of the pairwise scores for that image:

$$c_i = \frac{1}{n} \sum_{j=1,\ \hat{s} \in \hat{S}}^{n} \hat{s}_{i,j}. \tag{6}$$

[1] https://www.figure-eight.com

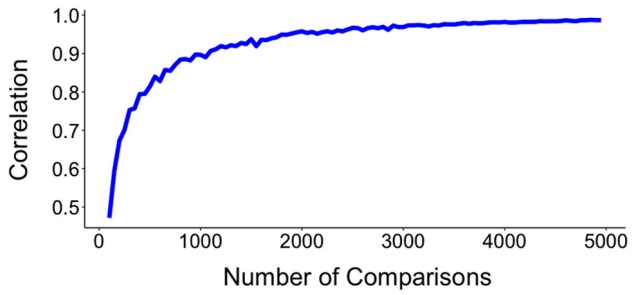


Fig. 5. For the category Suprematism, the correlation between C_ℓ and C_f is shown as a function of ℓ. High correlations can be achieved for some values of $\ell \ll \binom{n}{2}$.

4.3. Verifying the method to obtain ground truth

To confirm the consistency of the two aforementioned methods to convert the pairwise matrix to absolute final scores, we evaluated the correlation between the global ranking scores obtained by the two methods. We obtained correlations higher than 0.98 between the two methods for all seven image categories. It is therefore appropriate to use the ground-truth values of only one method (we use the Bradley–Terry method) in our subsequent analysis below.

4.3.1. Validity of partial matrix versus full matrix comparison

Here, we evaluate the accuracy of the absolute visual complexity scores as a function of ℓ, the number of pairwise comparisons between images. Since we have the full comparison matrix for the Suprematism category, we can perform such an analysis. Recall the notations from Section 4.2.2, where $S = \{s_{i,j}\}$ is the count matrix for the pairwise comparisons and $C = \{c_i\}$ is the list of absolute visual complexity scores, i.e., the output of the Bradley–Terry algorithm. We define S_ℓ and C_ℓ as the count matrix and absolute scores, respectively, where ℓ number of pairs have been selected by crowdsourcing. The full count matrix and the resulting absolute scores are denoted by S_f and C_f, respectively, with $\ell = \binom{n}{2}$.

The correlation between the visual complexity scores based on C_ℓ and C_f for ℓ in the range of $[100 - 4950]$ is shown in Fig. 5, which highlights the trade-off between the accuracy and the number of pairwise comparisons. For example, if only 2,000 pairs had been chosen to define S_ℓ for the Suprematism category, the result would be close to the result of a full comparison, since the correlation between C_l and C_f is 0.96. Given this result for the Suprematism category, we hypothesize that high correlations can also be achieved for the other six categories if ℓ is selected to be much smaller than $\binom{n}{2}$.

4.3.2. Distribution of the ground-truth scores

Initial analysis of the distribution of the absolute scores showed that the absolute scores are mostly distributed around zero. To mitigate this issue, we rescaled the range of pairwise scores, so that they are in the interval of $[0.33, 0.66]$ instead of $[0, 1]$, while still maintaining 0.5 as the score that represents equal visual complexity of an image pair. This adjustment is basically adding a temperature parameter to the sigmoid in Eq. (3).

Visual inspection of Fig. 6, which presents the distribution of scores for the seven categories, shows that the rescaling step was successful — each histogram is well distributed among the range of visual complexity numbers.

5. Experiments on unsupervised model

5.1. Datasets

In order to evaluate the performance of our proposed methodology, we compared our results on SAVOIAS, as well as two other datasets, namely RSIVL (Corchs et al., 2014) and abstract patterns (Gartus and Leder, 2013). A few sample images of these datasets are shown in Figs. 7–9.

5.1.1. RSIVL Dataset

To analyze the performance of our proposed metric on the real world scenes, we used two datasets provided by Corchs et al. (2014): *RS1* contains 49 images of the *RSIVL* dataset (RSIVL, 2016), and *RS2* with 29 images of the LIVE dataset (H. et al., 2006; Wang et al., 2004; Sheikh et al., 2006) and 20 of the IVL dataset (Corchs et al., 2014; IVL, 2014). Images were selected to have a wide variety of low-level and high-level features, i.e., colors, spatial frequencies, faces, buildings, outdoor scenes, animals, close-up or wide-angle shots, and various foreground and background configurations. Psychophysical experiments involving up to 39 participants were conducted by Corchs et al. (2014) to obtain human judgments of the visual complexity of these images. Sample images of the RS1 dataset along with their feature maps and human-curated and network-computed visual complexity scores are presented in Fig. 8.

5.1.2. Abstract patterns

This dataset consists of 912 abstract patterns selected from an extended set of patterns that were used by Gartus and Leder (2013). The patterns are designed by placing 36 to 44 black triangular elements in an 8×8 rectangular grid on a white background according to several criteria like number of objects and symmetry axes and thus are used to study the impact of these criteria with the visual complexity. Some examples of this dataset are shown in Fig. 9.

5.2. Baselines

We compare our results to the results reported by the state-of-the-art unsupervised methods. Note that while the UAE method is not a single feature, no further supervision is applied and the features are simply averaged. Thus, our method is still **unsupervised**. For the SAVOIAS dataset we compare our results with edge density, number of regions, compression ratio, feature congestion, and subband entropy algorithms. For the RSIVL dataset, we compare our results to those of three baseline methods (edge density, number of regions, and compression ratio). For the abstract patterns, we compare our results with mirror symmetry and RMSGIF metrics, previously used for the analysis of visual complexity on abstract patterns (Gartus and Leder, 2017). The RMSGIF metric is computed by first applying a root mean square contrast edge detection on an image and then measuring the compression ratio when the GIF file compression is applied on the edge detected image.

5.3. Results: UAE method

We selected a high-performing intermediate layer of each architecture as the best candidate for quantifying visual complexity and the Pearson correlation coefficient between this layer and the human-curated ground truth is used to evaluate the performance of the mentioned methods.

Through our psychology literature review, we learned that humans can perceive visual complexity of images better in comparison to other images, hence the scores assigned to images are relative to each other. Therefore, in designing our dataset and obtaining the ground truth labels, we performed pairwise comparison between the images. The same is true for our prediction algorithm. Our algorithm provides relative visual complexity scores with respect to the other images. These scores do change when a different depth layer in a network is considered. Note though, if the output numbers are not between [0,100], it does not matter because we use the Pearson correlation coefficient (which is normalized) to compute the correlation between the extracted values and the ground truth.

The correlations between the methods (baseline and ours) and the ground-truth complexity labels for SAVOIAS are shown in Table 3

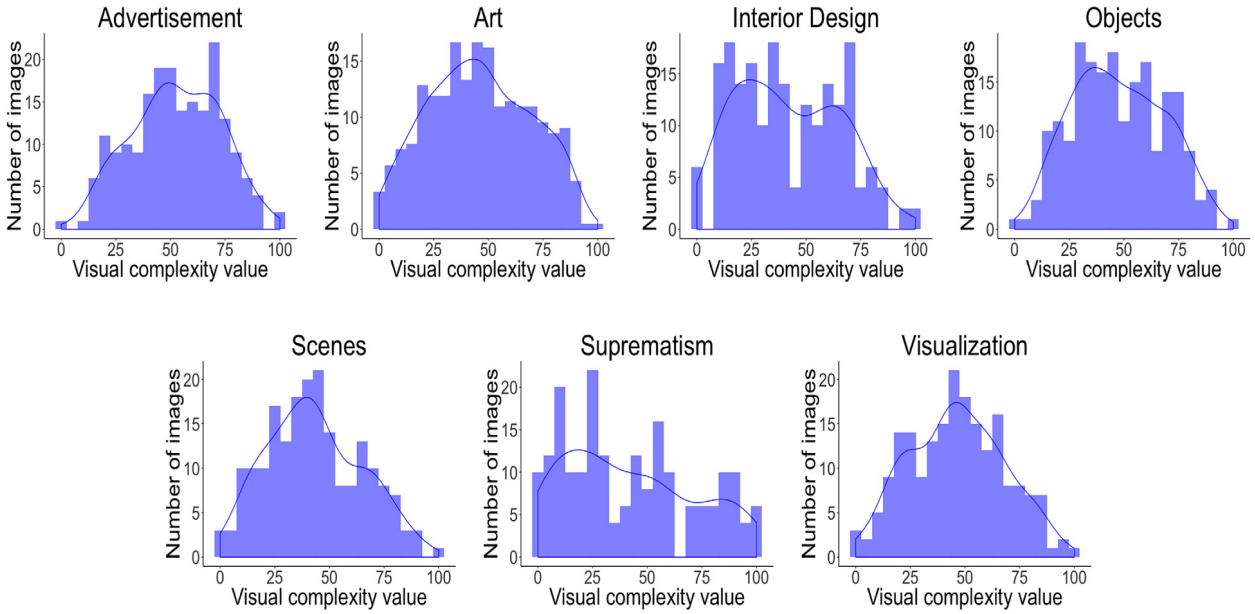


Fig. 6. Distribution of absolute visual complexity scores per category for the SAVOIAS dataset.

Fig. 7. Sample images of the Savoias dataset and their corresponding energy maps from the fourth max-pooling layer in the VGG-16 architecture trained for the scene recognition task. The images from top left to bottom right belong to datasets Advertisement, Places2, MASSVIS (Visualization and Infographics), MSCOCO, IKEA, and art respectively. Note that the last three images belong to the art dataset where the last sample is from the Suprematism category.

Fig. 8. Sample images of the RS1 dataset (top), energy maps (bottom). The energy maps correspond to the activations of the fourth max-pooling layer in the VGG-16 architecture trained for the scene recognition task. The order of image correspond to increasing visual complexity from left to right.

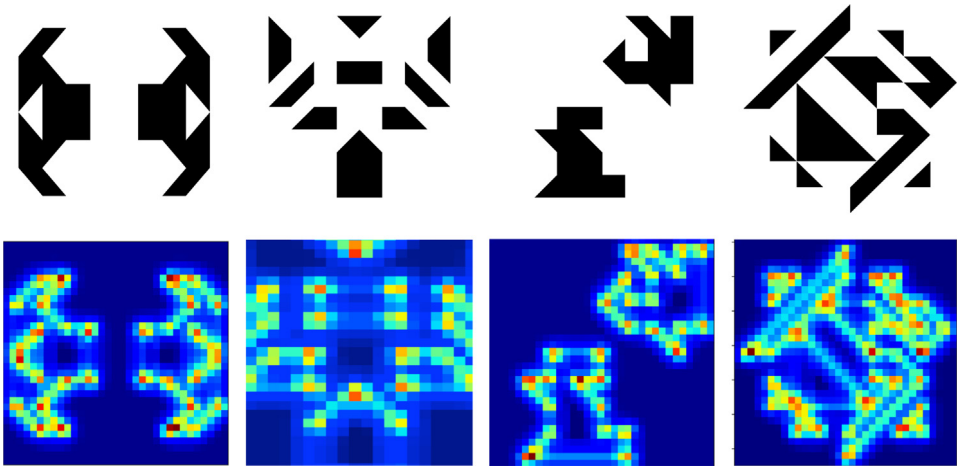


Fig. 9. Sample images of the abstract patterns dataset (top), energy maps (bottom). The energy maps correspond to the activations of the fourth max-pooling layer in the VGG-16 architecture trained for the scene recognition task. The images are in the increasing order of visual complexity from left to right.

Table 3
Results and comparison to prior work: correlation between human-curated and computed visual complexity scores based on UAE for the Savoias dataset. Our results outperforms all other methods by a significant margin and thus can be used for variety of image categories. The categories Ad, Sup and Vis refer to Advertisement, Suprematism and Visualization and Info graphics respectively.

Model	Ad.	Art	Int.	Obj.	Scenes	Sup.	Vis.
Edge Density (Rosenholtz et al., 2007)	0.54	0.48	0.63	0.27	0.16	0.18	0.57
Compression Ratio (Rosenholtz et al., 2007)	0.56	0.51	0.72	0.16	0.30	0.60	0.55
Number of Regions (Comaniciu and Meer, 2002)	0.41	0.65	0.69	0.29	0.57	0.84	0.38
Feature Congestion (FC) (Rosenholtz et al., 2007)	0.56	0.22	0.63	0.30	0.42	0.48	0.52
Subband Entropy (SE) (Rosenholtz et al., 2007)	0.54	0.33	0.31	0.10	0.16	0.39	0.61
VGG16 Scene Recognition [Ours]	**0.73**	0.50	0.82	**0.67**	0.76	0.84	**0.71**
VGG16 Object Classification [Ours]	0.71	0.59	**0.83**	0.64	**0.77**	**0.85**	0.70
Inception-v4 Object Classification [Ours]	0.61	**0.68**	0.81	0.60	0.74	0.84	0.67
ResNet-v2-152 Object Classification [Ours]	0.71	0.41	0.71	0.60	0.71	0.73	0.68
DenseNet Object Classification [Ours]	0.58	0.26	0.75	0.48	0.66	0.71	0.66
MobileNet Object Classification [Ours]	0.47	0.48	0.78	0.53	0.65	0.82	0.64
EfficientNet-B7 Object Classification [Ours]	0.54	0.46	0.69	0.39	0.60	0.77	0.40

for all seven categories. For all the architectures shown in this table, correlations exist between our proposed unsupervised activation energy method and the ground truth. This confirms our hypothesis that the intermediate layers of deep convolutional networks carry information regarding the visual complexity, and that our proposed UAE method outperforms all the previous work.

The VGG architecture is known to be a useful architecture for vision tasks that are closer to human visual perception such as super-resolution Bruna et al. (2015), Johnson et al. (2016) and Ledig et al. (2017). This architecture seem to be able to extract features that are related to human perception of images. Our results confirms such a quality in the VGG architecture, since in most cases, the highest correlation between our metric and human-curated ground truth is obtained from the VGG architecture.

Table 4
Correlation between human-curated and computed visual complexity scores based on UAE for RSIVL dataset.

Model	RS1	RS2
Edge Density (Rosenholtz et al., 2007)	0.65	0.66
Compression Ratio (Rosenholtz et al., 2007)	0.67	0.67
Number of Regions (Comaniciu and Meer, 2002)	0.74	0.69
VGG-16 Scene Recognition [Ours]	**0.79**	0.72
VGG-16 Object Classification [Ours]	0.76	0.70
Inception Object Classification [Ours]	0.76	**0.73**
ResNet Object Classification [Ours]	0.75	0.70
DenseNet Object Classification [Ours]	0.54	0.62
MobileNet Object Classification [Ours]	0.62	**0.73**
EfficientNet-B7 Object Classification [Ours]	0.43	0.60

Table 5
Correlation between human-curated and computed visual complexity scores based on UAE for the abstract patterns dataset.

Model	
RMSGIF (Gartus and Leder, 2017)	0.63
Mirror Symmetry (MS) (Bauerly and Liu, 2006)	0.58
VGG-16 Scene recognition [Ours]	0.58
VGG-16 Object Classification [Ours]	0.60
Inception Object Classification [Ours]	**0.65**
ResNet Object Classification [Ours]	0.50
DenseNet Object Classification [Ours]	0.49
MobileNet Object Classification [Ours]	0.50
EfficientNet-B7 Object Classification [Ours]	0.56

In case of the object category (from MSCOCO), there exists a margin of 0.37 between our method and the second best performing method; in which our model (VGG-16 architecture trained for the scene recognition task) has a correlation of 0.67 between our method and the ground truth, while the second best performing method is feature congestion and results in a correlation of 0.30. Although our neural networks are trained based on object or scene classification, the UAE method applied to selected layer can successfully predict the visual complexity of the images in all other categories as well.

For both *RS1* and *RS2*, the average correlation values we measured for our method are higher than those reported by prior work for unsupervised algorithms (0.79 and 0.73, respectively)(see Table 4). In order to evaluate the performance of our proposed methodology on a different task than what the neural network has been trained on, we performed a separate experiment with the abstract patterns. The images in the abstract patterns dataset are designed to evaluate the impact of number of shapes and symmetry in quantifying the visual complexity. Although our model has not been trained specifically to focus on the number of shapes or symmetry, it can successfully quantify the visual complexity of the images and outperforms the methods discussed in the work of Gartus and Leder (2017) as shown in Table 5.

It has been recently proposed that the structure of a generator network can capture a great deal of low-level image statistics prior to any learning (Ulyanov et al.). To investigate whether any untrained classification architecture also carries such information, we compared our results from the pretrained models for two different tasks of object and scene classifications with the untrained model. We noticed that there exists a small but consistent correlation between all of the untrained models and the ground truth of visual complexity. This result can suggest that these architectures may capture such low-level features. For prediction of visual complexity, however, extraction of both high-level and low-level features is required. Since the untrained models fall short on extracting semantic features, they cannot perform well as the trained models do.

5.4. Discussion

The results of our experiments and the statistically significant superior performance of our unsupervised activation energy method confirm the applicability and advantage of our method compared to previous work. The results also highlight that our method can be generalized to various types of images with no supervision.

Table 6
Crowdplatform contributor feedback.

	Scenes	Ad.	Vis.	Objects	Interior	Art	Sup.	Avg.
Overall satisfaction	4.2	4.3	4.1	4.5	4.1	4.3	**3.8**	4.2
Ease of job	4.1	4	4	4.3	3.9	4.1	**3.6**	4

- **Model**: Linear Regression
- **Model Name**: Supervised Activation Energy
- **Input**: Feature map of the fourth max pooling layer
- **Output**: Visual Complexity Predictions
- **Trainable Variables**: weights for the neurons in the Feature map

Fig. 10. Description of the Supervised Activation Energy (SAE) Method.

Comparing the correlations between visual complexity scores based on crowdsourcing platform contributors and the state-of-the-art algorithms, we observe that 'Suprematism,' which is the most challenging category for the contributors (Table 6), had the highest correlation with the number of regions method. On the other hand, contributors found the 'Object' category to be the least challenging category, while all the previous work performed poorly on this category (the highest correlation is only 0.3). Based on this observation, we postulate that the previous work is more capable of making decisions based on features such as geometric shapes, textures, and patterns, found in the Suprematism category, than image features such as objects and people, which are easier for human contributors to judge.

Our results show that our proposed unsupervised activation energy method is not only capable of quantifying the low-level features and simple shapes, but can also successfully focus on the features of an image that human observers use when they evaluate the visual complexity of an image.

6. Supervised versus unsupervised methods

In this section, we are interested in exploring whether we can improve the interpretation of the feature maps in the fourth max-pooling layer by training a supervised model and learning the contribution of the activated neurons. In addition, we examine how generalized these supervised models are in predicting the visual complexity of images from a different image category.

In the Unsupervised Activation Energy (UAE) method, we take average of all the activated neurons on the fourth max-pooling layer. Now we define the **Supervised Activation Energy (SAE) method** where models are trained to learn the weights of the activated neurons instead of taking the average (see Fig. 10).

6.1. Approach

We use linear regression to train our supervised model. We first resize the feature maps to size $14 \times 14 \times 512$. Thus the total number of features for each image would be 100,352. Since the number of features to train the regression model are high compared to the number of images (100–400 images depending on the category), we need to make sure that the regression models do not overfit. We regularize the models using the Ridge regression model. In Ridge regression, the least square loss function is augmented by a second term as shown in Eq. (7). While the least squares loss minimizes the sum of squared residuals, the second term penalizes the size of parameter estimates, also called the

Table 7
Results of the supervised models (SAE) in terms of the correlation between the supervised model and the human-curated ground truth, trained on all the activated features of the feature map, averaged across the channels (spatial features), averaged across image dimension (depth features), and compared against the Unsupervised Activation Energy (UAE) method, used here as the baseline.

Model	Ad.	Art	Int.	Obj.	Scenes	Sup.	Vis.
All features	0.74	0.84	0.82	0.77	0.84	0.90	0.74
Spatial features	0.69	0.63	0.83	0.63	0.78	0.83	0.67
Depth features	0.75	0.86	0.86	0.80	0.85	0.90	0.72
Unsupervised activation energy	0.70	0.60	0.79	0.64	0.76	0.81	0.66

Table 8
Comparison of the performance of the SAE models when trained on the same category as being tested versus when trained on the art category. The results are also compared with a baseline (UAE model). For the advertisement and visualization categories, the result of the SAE model trained on the art category is lower than the UAE method which suggests that the type of images in the advertisement and visualization categories are different from the images in the art category. This may be attributed to the fact that there exists text in these two categories, but not in the art category.

Model	Ad.	Art	Int.	Obj.	Scenes	Sup.	Vis.
Same Category as Test	0.74	-	0.82	0.77	0.84	0.90	0.74
Art Category	0.68	-	0.80	0.70	0.79	0.87	0.62
Activation Energy (Unsupervised)	0.70	-	0.79	0.64	0.76	0.81	0.66

vector of the coefficients, in order to shrink them towards zero. The loss function is

$$L_{ridge}(\hat{\beta}) = \|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_2^2, \tag{7}$$

where $\hat{\beta}$ is the vector of the coefficients assigned to each feature that needs to be estimated, X is the input features, and y is the ground truth labels. The first term of the loss function is the least squares loss while the second one is the penalty term for high values in the squared magnitude of the $\hat{\beta}$ vector.

In order to analyze the importance of spatial information in the energy maps and the information activated in different channels of the convolutional neural network, we have performed further analysis by taking averages spatially as well as across the channels. The following features are used separately to train the regression models:

$$F_{all}[i,j,k] = F_4[i,j,k] \tag{8}$$

$$F_{spatial}[i,j] = \frac{1}{d}\sum_{k=1}^{d} F_4[i,j,k] \tag{9}$$

$$F_{depth}[k] = \frac{1}{m \times n}\sum_{i,j=1}^{m,n} F_4[i,j,k], \tag{10}$$

where F_4 represents the feature maps of the fourth max-pooling layer of the deep neural network. If the shape of F_4 is indicated as $m \times n \times d$, then $m = 14$, $n = 14$ and $d = 512$; thus the size of the spatial features and the depth features are $14 \times 14 = 196$ and 512, respectively. Note that the formulas in Eqs. (1) and (9) are the same with $F_l = F_4$, i.e., the feature maps of the fourth max-pooling layer. The features for training the models are flattened into a one-dimensional vector.

6.2. Results: UAE versus SAE

Here we present the results of our SAE models. We report the correlation between model output and the ground truth in Table 7. The first three rows are the results of regression models trained on the features as described in Eqs. (8), (9), and (10), while the last one is the UAE method, used here as the baseline.

The value of the regularization parameter used for the Ridge regression is $\lambda = 0.01$, chosen by cross validation. The feature maps are obtained from the VGG architecture, pre-trained for the object classification task. Similar results are obtained by using the other network architectures and tasks (omitted for the sake of brevity).

As the results in Table 7 suggest, we can obtain higher correlations by training a supervised model on the feature maps. The models trained on all features and the depth features show higher improvement compared to the spatial features. The negligible improvement of the regression model trained on the spatial features compared to the energy metric with equal weights for all the spatial features suggests that the spatial location of clutter does not significantly impact the judgment of the human observers. On the other hand, our results show that the contribution of various channels of the feature maps can be different, and setting weights for each channel using supervised methods can improve the results.

6.3. Discussion

As shown in Table 7, the best results are mostly obtained from the depth features and not all of the features. This suggests that due to a high number of features, when all of them are being used for training, the model may not be able to learn the contribution of different channels.

As stated earlier, the results are obtained by setting the value of the regularization parameter to 0.01, which allows the coefficients of the regression model to have a relatively high value if needed. Setting the regularization parameter to 1 or, in other words, limit the distribution of the coefficients to be in a constrained range, the correlations obtained from the depth features will be closer to the results of the baseline.

Note that the results of the UAE method reported here may be slightly different from the ones reported in Table 3. The reason is that for the analysis presented here, all the images are resized to $14 \times 14 \times 512$ before the calculations for this section. However, no resizing has been performed for results of Table 3.

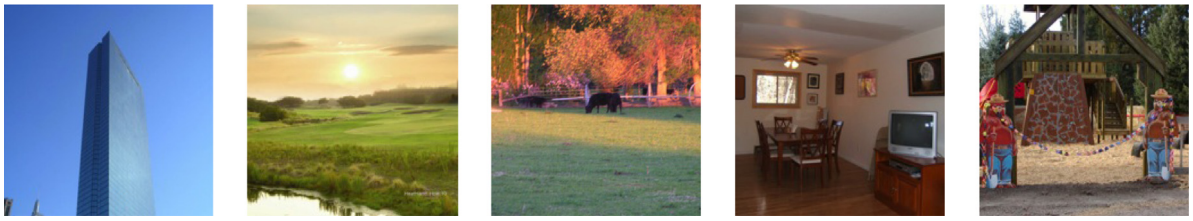
6.3.1. How does a model trained on one category perform on other categories?

In order to evaluate the generalizability of our supervised models, we have also evaluated the performance of a regression model when it was trained for one category and tested on the rest of the categories.

Three different cases are compared in Table 8. The first row represents the results when the models used for training and testing are from the same category, e.g., trained on advertisement and tested on advertisement. The second row represents the results of the regression model trained on the **art** category, while tested on all other categories. These two cases are then compared to the case in the third row, which is the unsupervised model, i.e., the model has no information about any of the categories.

For the categories of Interior design, Objects, Scenes and Suprematism, the correlations reported on the second row (trained on art category) are less than the first row (same category as test), but they are

Table 9
Some examples of the SUN database (Xiao et al.) used in the FIGRIM dataset (Bylinskii et al., 2015) for evaluation of memorability in images when target images were among other images of the same context. From left to right, the ground truth memorability score obtained in the FIGRIM dataset increases.



Skyscraper	Golf Course	Pasture	Kitchen	Playground
$HR_{c1} = 0.33$	$HR_{c1} = 0.47$	$HR_{c1} = 0.55$	$HR_{c1} = 0.76$	$HR_{c1} = 0.82$

higher than the results of the third row (unsupervised energy metric). This states that although the model trained on the art category may not be as accurate the models trained on the same category as being tested, it has learnt a set of coefficients that are applicable in other categories. For the two categories of Advertisement and Visualizations, however, the correlations reported on the third row are even lower then the unsupervised energy metric. This lower performance can be attributed to the difference in the nature of these two categories. We hypothesize that since the Advertisement and Visualizations contain text, the model trained on the art category cannot predict the visual complexity of the images in these two categories very accurately.

7. Application: Visual complexity affects image memorability

Memorability of an image can be defined by metrics corresponding to the probability of an image being remembered by a person. Various efforts have been made to quantify how memorable an image is and what the contributing factors are (Bylinskii et al., 2015, 2017; Perera et al., 2019). Memorability has been previously studied for different visual stimulus sets including faces (Bainbridge et al., 2013; Khosla et al., 2013), scenes (Isola et al., 2011a; Khosla et al., 2015; Bylinskii et al., 2015; Khosla et al., 2012), and graphs and visualizations (Borkin et al., 2016, 2013; Bylinskii et al., 2017) to understand the features driving higher memorability.

7.1. Approach

It has been demonstrated before that some scene categories are intrinsically more memorable (Bylinskii et al., 2015). In this work, we examine whether there exists a correlation between image complexity and image memorability. We choose the feature maps of Pool 4 (the fourth max-pooling layer) in the VGG-16 architecture trained for the scene recognition task as a means to quantify visual complexity of images.

There are numerous factors that contribute to the memorability score of an image. Thus, in order to find the underlying correlation between the visual complexity and memorability, we need to minimize the impact of other factors on the memorability score. In other words, we want to distinguish between context-specific factors versus factors that are specific to one particular image and not the context. We propose to do so by focusing on intra-class and inter-class analyses of image categories, as such analyses mitigate the effect of undesirable factors.

To illustrate our approach, we provide the following example. An image of a simple garden with a lawn and fence is not as memorable as an image of a garden with lawn and fence that also includes roses, perennials, trees, etc. However, the simple garden would become more memorable with an out-of-place purple teddy bear lying on its lawn than the second unaltered garden. To analyze the connection between visual complexity and memorability in *typical* garden images (to separate the effect of a purple teddy bear), we do an inter-class analysis.

In order to perform inter-class analysis, we need to first investigate whether some scene categories are inherently more visually complex, and if such a condition holds, we can examine whether there exists a correlation between image complexity and image memorability. We start by studying how visual complexity varies for different scene categories and then present how visual complexity and memorability are related.

7.2. Dataset for memorability analysis

For our experiments, we used the FIGRIM dataset introduced by Bylinskii et al. (2015). The FIGRIM dataset consists of 21 different indoor and outdoor scene categories (from the SUN database (Xiao et al.) and thus satisfies the requirement for the inter-class analysis.

In order to collect memorability scores, Bylinskii et al. followed the protocol of Isola et al. (2011b), by setting up memory games on Amazon Mechanical Turk (AMT) and showing target images twice among filler images. The number of hits, i.e., participants recognized an image was repeated, misses, false alarms, and false rejections were then reported. In each scene category, a quarter of the images (a minimum of 56 and maximum of 157 images from a scene category) are randomly selected as target images and the rest are used as filler images. In our experiments, we use the metric *Hit Rate (HR)* for an image *I*, as defined by Bylinskii et al. for target images shown among image fillers within the image context (AMT1) and across the image context (AMT2):

$$HR(I) = \frac{hits(I)}{hits(I) + misses(I)} \times 100\%. \tag{11}$$

Example images and their memorability score are shown in Table Table 9.

7.3. Results: Correlation between visual complexity predictions and image memorability

In order to evaluate the correlation between image complexity and its memorability, we first need to demonstrate that the visual complexity of images within the same category is consistent and verify that visual complexity values within a scene category do not fluctuate widely. In order to evaluate such consistency within scene categories, we followed the approach suggested by Bylinskii et al. (2015). The set of images of each scene category are randomly split into two subsets and the average of the energy for each subset is computed. The Pearson correlation coefficient between the two subsets is then computed representing the consistency of visual complexity among scene categories. The same experiment was run multiple times and similar results were obtained. On average, the correlation between the two subsets was 0.97, demonstrating significant consistency among different members of the same scene category.

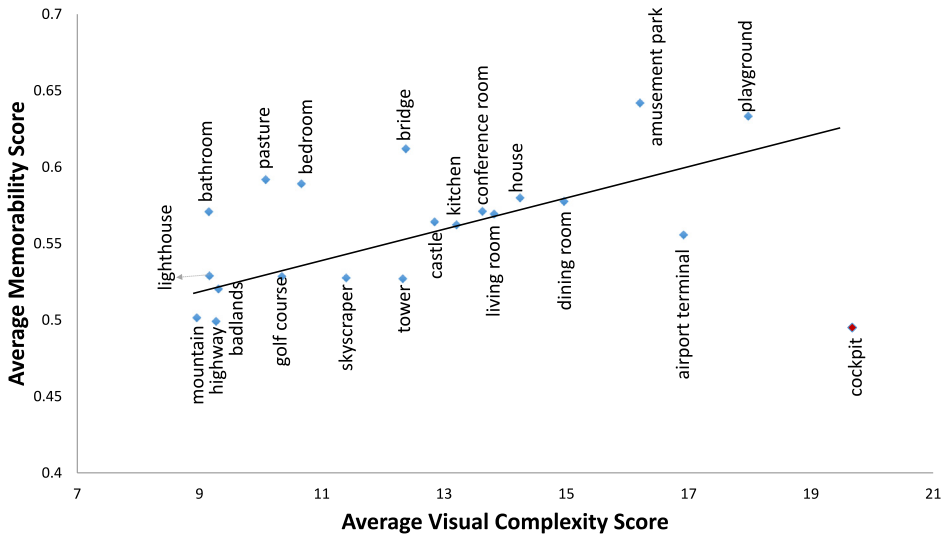


Fig. 11. Scatter plot of the visual complexity and memorability scores averaged for each of 21 scene categories in the AMT1 dataset. The visual complexity scores are computer using the UAE method. The Pearson correlation coefficient excluding one outlier is 0.65, which confirms that classes with higher visual complexity are more memorable.

Establishing the consistency of visual complexity values within classes enables us to evaluate the average of each class instead of the individual images. Next, we explore whether there exists a correlation between the predicted visual complexity values and the memorability scores (HRs). To evaluate our hypothesis for each of 21 categories, we first compute the average of the visual complexity score and the average of the memorability score over images in the same category. Then, using these 21 pairs, we calculate the correlation between the visual complexity and the memorability vectors and perform inter-class analysis.

A scatter plot of the average visual complexity score and average memorability score for each category is shown in Fig. 11. The correlation between the two attributes is 0.35. Excluding the class "cockpit" as an outlier due to the high similarity of images in this class, the correlation jumps to 0.65. The positive correlation between visual complexity and memorability suggests that, in reference to AMT1, images that are more visually complex are more memorable and therefore visual complexity can be used as a factor in memorability analysis.

8. Conclusion

In this paper, we proposed the unsupervised use of information activated by filters in the intermediate convolutional layers of a deep neural network. We showed that the activation energy of these layers can successfully quantify the visual complexity of an image. We also provided evidence that visual complexity information can be extracted for various network architectures and tasks (scene recognition and object classification). In addition, to investigate the applicability of our method to various image understanding tasks, we introduced SAVOIAS, a new dataset for the analysis of visual complexity in images. SAVOIAS compromises of more than 1,400 images, which belong to seven diverse categories. The ground-truth complexity values were obtained by processing the judgments of 1,687 crowdplatform contributors who compared the visual complexity of more than 37,000 pairs of images. In our experiment, we showed that our Unsupervised Activation Energy method can outperform all the previous unsupervised methods in quantifying visual complexity. Furthermore, we show that the performance can be improved by learning the weights of the activated neurons.

Our work may leverage research in other areas of computer vision, such as segmentation, visual search, image captioning, and visual question answering, for example by determining the visually complex regions of the image or estimating the difficulty level of the task based on visual complexity of the image. Furthermore, our proposed dataset facilitates research in the field of psychophysics and cognitive science to find the underlying factors in the stimulus that affect the perception of visual complexity in humans. Lastly, our proposed method enables artists, Web and graphic designers, interior designers, and advertisers to estimate the level of visual complexity of their work in order to maximize the quality of their design and the impact of their work on their audience.

Declaration of competing interest

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

CRediT authorship contribution statement

Elham Saraee: Methodology, Software, Writing - original draft. **Mona Jalal:** Dataset Creation, Writing - original draft, Writing - review & editing. **Margrit Betke:** Supervision, Writing - review & editing.

Acknowledgments

We would like to thank Yifu Hu and Yi Zheng for preparing the images for the interior design category of our dataset. This paper is based on work that was supported in part by the National Science Foundation (grants 1421943 and 1838193).

References

Amirshahi, S.A., Pedersen, M., Yu, S.X., 2017. Image quality assessment by comparing cnn features between images. Electron. Imaging 2017, 42–51.
Arrow, K.J., 1950. A difficulty in the concept of social welfare. J. Polit. Econ. 58, 328–346.
Babenko, A., Lempitsky, V., 2015. Aggregating local deep features for image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1269–1277.
Bainbridge, W.A., Isola, P., Oliva, A., 2013. The intrinsic memorability of face photographs. J. Exp. Psychol. [Gen.] 142, 1323.
Bauerly, M., Liu, Y., 2006. Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. Int. J. Hum.-Comput. Stud. 64, 670–682.
Berlyne, D.E., 1971. Aesthetics and Psychobiology. Appleton-Century-Crofts, New York.
Birkhoff, G.D., 1933. Aesthetic Measure, vol. 38. Harvard University Press Cambridge.
Borkin, M.A., Bylinskii, Z., Kim, N.W., Bainbridge, C.M., Yeh, C.S., Borkin, D., Pfister, H., Oliva, A., 2016. Beyond memorability: Visualization recognition and recall. IEEE Trans. Vis. Comput. Graphics 22, 519–528.

Borkin, M.A., Vo, A.A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., Pfister, H., 2013. What makes a visualization memorable?. IEEE Trans. Vis. Comput. Graphics 19, 2306–2315.

Bradley, R.A., Terry, M.E., 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39, 324–345.

Bruna, J., Sprechmann, P., LeCun, Y., 2015. Super-resolution with deep convolutional sufficient statistics. arXiv preprint arXiv:1511.05666.

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., Oliva, A., 2015. Intrinsic and extrinsic effects on image memorability. Vis. Res. 116, 165–178.

Bylinskii, Z., Kim, N.W., O'Donovan, P., Alsheikh, S., Madan, S., Pfister, H., Durand, F., Russell, B., Hertzmann, A., 2017. Learning visual importance for graphic designs and data visualizations. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. ACM, pp. 57–69.

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. Found. Comput. Math. 9, 717–772.

Cardaci, M., D. Gesu, V., Petrou, M., Tabacchi, M.E., 2009. Attentional vs computational complexity measures in observing paintings. Spatial Vis. 22, 195–209.

Chang, H., Yu, F., Wang, J., Ashley, D., Finkelstein, A., 2016. Automatic triage for a photo series. ACM Trans. Graph. 35, 148:1–148:10.

Chen, L.C., Barron, J.T., Papandreou, G., Murphy, K., Yuille, A.L., 2016. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4545–4554.

Chipman, S.F., 1977. Complexity and structure in visual patterns. J. Exp. Psychol. [Gen.] 106, 269–301.

Chipman, S.F., Mendelson, M.J., 1979. Influence of six types of visual structure on complexity judgments in children and aults. J. Exp. Psychol.: Hum. Percept. Perform. 5, 365–378.

Cimpoi, M., Maji, S., Kokkinos, I., Vedaldi, A., 2016. Deep filter banks for texture recognition, description, and segmentation. Int. J. Comput. Vis. 118, 65–94.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24, 603–619.

Corchs, S.E., Ciocca, G., Bricolo, E., Gasparini, F., 2016. Predicting complexity perception of real world images. PLoS One 11, e0157986.

Corchs, S., Gasparini, F., Schettini, R., 2014. No reference image quality classification for jpeg-distorted images. Digit. Signal Process. 30, 86–100.

Da Silva, M.P., Courboulay, V., Estraillier, P., 2011. Image complexity measure based on visual attention. In: Proceedings of the 18th IEEE International Conference on Image Processing (ICIP), pp. 3281–3284.

David, H.A., 1963. The Method of Paired Comparisons. vol. 12. London.

Eysenck, H.J., 1941. The empirical determination of an aesthetic formula. Psychol. Rev. 48, 83.

Fan, Z.B., Li, Y., Yu, J., Zhang, K., 2017. Visual complexity of Chinese ink paintings. In: Proceedings of the ACM Symposium on Applied Perception, pp. 9:1–9:8.

Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C.J., Sawey, M., 2011. Predicting beauty: fractal dimension and visual complexity in art. Br. J. Psychol. 102, 49–70.

Gao, F., Wang, Y., Li, P., Tan, M., Yu, J., Zhu, Y., 2017. Deepsim: Deep similarity for image quality assessment. Neurocomputing 257, 104–114.

Gartus, A., Leder, H., 2013. The small step toward asymmetry: aesthetic judgment of broken symmetries. i-Perception 4, 361–364.

Gartus, A., Leder, H., 2017. Predicting perceived visual complexity of abstract patterns using computational measures: The influence of mirror symmetry on complexity perception. PLoS One 12, e0185276.

Gleich, D.F., Lim, L.h., 2011. Rank aggregation via nuclear norm minimization. In: Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining, pp. 60–68.

Gordo, A., Almazan, J., Revaud, J., Larlus, D., 2017. End-to-end learning of deep visual representations for image retrieval. Int. J. Comput. Vis. 124, 237–254.

Gordo, A., Almazán, J., Larlus, D., 2016. Deep image retrieval: Learning global representations for image search. In: European Conference on Computer Vision. Springer, pp. 241–257.

Guo, X., Qian, Y., Li, L., Asano, A., 2018. Assessment model for perceived visual complexity of painting images. Knowl.-Based Syst. 159, 110–119.

Gupta, P., Srivastava, P., Bhardwaj, S., Bhateja, V., 2011. A modified psnr metric based on hvs for quality assessment of color images. In: 2011 International Conference on Communication and Industrial Application. IEEE, pp. 1–4.

H., S., Z., W., L., C., A., B., 2006. LIVE Image quality assessment database release 2. http://live.ece.utexas.edu/research/quality.

Haytko, D.L., Baker, J., 2004. It's all at the mall: exploring adolescent girls' experiences. J. Retail. 80, 67–83.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heaps, C., Handel, S., 1999. Similarity and features of natural textures. J. Exp. Psychol.: Hum. Percept. Perform. 25, 299.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.

Huh, M., Agrawal, P., Efros, A.A., 2016. What makes imagenet good for transfer learning?. arXiv preprint arXiv:1608.08614.

Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., Kovashka, A., 2017. Automatic understanding of image and video advertisements. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1100–1110.

IKEA, 0000. https://www.ikea.com.

Ionescu, R.T., Alexe, B., Leordeanu, M., Popescu, M., Papadopoulos, D.P., Ferrari, V., 2016. How hard can it be? Estimating the difficulty of visual search in an image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2157–2166.

Isola, P., Parikh, D., Torralba, A., Oliva, A., 2011a. Understanding the intrinsic memorability of images. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 2429–2437.

Isola, P., Xiao, J., Torralba, A., Oliva, A., 2011b. What makes an image memorable?. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 145–152.

IVL, 2014. Imaging and Vision Laboratory. Department of Informatics, Systems and Communication, University of Milano-Bicocca, http://www.ivl.disco.unimib.it/activities/image-quality.

Jacobsen, T., Höfel, L., 2001. Aesthetics electrified: An analysis of descriptive symmetry and evaluative aesthetic judgment processes using event-related brain potentials. Empir. Stud. Arts 19, 177–190.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. Springer, pp. 694–711.

Kendall, M.G., Smith, B.B., 1940. On the method of paired comparisons. Biometrika 31, 324–345.

Khosla, A., Bainbridge, W.A., Torralba, A., Olivia, A., 2013. Modifying the memorability of face photographs. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3200–3207.

Khosla, A., Raju, A.S., Torralba, A., Oliva, A., 2015. Understanding and predicting image memorability at a large scale. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2390–2398.

Khosla, A., Xiao, J., Torralba, A., Oliva, A., 2012. Memorability of image regions. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 296–304.

Kim, W.H., Jalal, M., Hwang, S.J., Johnson, S.C., Singh, V., 2017. Online graph completion: Multivariate signal recovery in computer vision. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 5019–5027.

Kim, J., Lee, S., 2017. Deep learning of human visual sensitivity in image quality assessment framework. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1676–1684.

Krishen, A., 2008. Perceived versus actual complexity for websites: Their relationship to consumer satisfaction. J. Consum. Satisf. Dissatisfaction Complain. Behav. 21 (104).

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1097–1105.

Ledig, C., Theis, L., Huszár, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690.

Li, G., Yu, Y., 2015. Visual saliency based on multiscale deep features. arXiv preprint arXiv:1503.08663.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in COntext. In: European Conference on Computer Vision. Springer, pp. 740–755.

Liu, H., Chen, T., Shen, Q., Yue, T., Ma, Z., 2018. Deep image compression via end-to-end learning. In: CVPR Workshops, pp. 2575–2578.

Liu, N., Han, J., 2016. Dhsnet: Deep hierarchical saliency network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 678–686.

Liu, L., Shen, C., van den Hengel, A., 2015. The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4749–4757.

Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J., Carballal, A., 2015. Computerized measures of visual complexity. Acta Psychol. 160, 43–57.

Mack, M.L., Oliva, A., 2004. Computational estimation of visual complexity. In: The 12th Annual Object, Perception, Attention, and Memory Conference, Minneapolis, Minnesota.

Marin, M.M., Leder, H., 2016. Effects of presentation duration on measures of complexity in affective environmental scenes and representational paintings. Acta Psychol. 163, 38–58.

Miller, G.A., 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev. 63, 81.

Miniukovich, A., Angeli, A.D., 2014. Quantification of interface visual complexity. In: Proceedings of International Working Conference on Advanced Visual Interfaces, AVI, pp. 153–160.

Nadal, M., Munar, E., Marty, G., Cela-Conde, C.J., 2010. Visual complexity and beauty appreciation: Explaining the divergence of results. Empir. Stud. Arts 28, 173–191.

Ng, J.Y.H., Yang, F., Davis, L.S., 2015. Exploiting local features from deep networks for image retrieval. arXiv preprint arXiv:1504.05133.

Oliva, A., Mack, M.L., Shrestha, M., Peeper, A., 2004. Identifying the perceptual dimensions of visual complexity of scenes. In: Proceedings of the Annual Meeting of the Cognitive Science Society.

Palumbo, L., Ogden, R., Makin, A.D., Bertamini, M., 2014. Examining visual complexity and its influence on perceived duration. J. Vis. 14, 3–3.

Paulin, M., Mairal, J., Douze, M., Harchaoui, Z., Perronnin, F., Schmid, C., 2017. Convolutional patch representations for image retrieval: an unsupervised approach. Int. J. Comput. Vis. 121, 149–168.

Perera, S., Tal, A., Zelnik-Manor, L., 2019. Is image memorability prediction solved?. arXiv preprint arXiv:1901.11420.

Pieters, R., Wedel, M., Batra, R., 2010. The stopping power of advertising: Measures and effects of visual complexity. J. Market. 74, 48–60.

Pilelienė, L., Grigaliūnaitė, V., 2018. Effect of visual advertising complexity on consumers? attention. Economics 3, 489–501.

Ramanarayanan, G., Bala, K., Ferwerda, J.A., Walter, B., 2008a. Dimensionality of visual complexity in computer graphics scenes. In: Proceedings of Human Vision and Electronic Imaging XIII Conference, p. 68060E.

Ramanarayanan, G., Bala, K., Ferwerda, J.A., Walter, B., 2008b. Dimensionality of visual complexity in computer graphics scenes. In: Proceedings of the Human Vision and Electronic Imaging XIII, p. 68060E.

Razavian, A.S., Sullivan, J., Carlsson, S., Maki, A., 2016. Visual instance retrieval with deep convolutional networks. ITE Transa. Media Technol. Appl. 4, 251–258.

Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., Gajos, K.Z., 2013. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 2049–2058.

Rosenholtz, R., Li, Y., Nakano, L., 2007. Measuring visual clutter. J. Vis. 7 (17).

RSIVL, 2016. Imaging and Vision Laboratory, Department of Informatics, Systems and Communication. university of milano-bicocca, http://www.ivl.disco.unimib.it/activities/complexity-perception-in-images; 2016..

Sameki, M., Lai, S., Mays, K.K., Guo, L., Ishwar, P., Betke, M., 2019. BUOCA: budget-optimized crowd worker allocation. Comput. Res. Repos. abs/1901.06237.

Schnur, S., Bektaş, K., Çöltekin, A., 2018. Measured and perceived visual complexity: A comparative study among three online map providers. Cartogr. Geogr. Inf. Sci. 45, 238–254.

Sheikh, H.R., Sabir, M.F., Bovik, A.C., 2006. A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Process. 15, 3440–3451.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Snodgrass, J.G., Vanderwart, M., 1980. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. J. Exp. Psychol.: Hum. Learn. Memory 6, 174.

Sohn, S., Seegebarth, B., Moritz, M., 2017. The impact of perceived visual complexity of mobile online shops on user's satisfaction. Psychol. Mark. 34, 195–214.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence, p. 12.

Tan, M., Le, Q.V., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946.

Toderici, G., Vincent, D., Johnston, N., Ji. Hwang, S., Minnen, D., Shor, J., Covell, M., 2017. Full resolution image compression with recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5306–5314.

Tzeng, E., Hoffman, J., Darrell, T., Saenko, K., 2015. Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4068–4076.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2018. Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9446–9454.

Uricchio, T., Bertini, M., Seidenari, L., Bimbo, A., 2015. Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 9–15.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612.

Wang, N., Li, S., Gupta, A., Yeung, D.Y., 2015. Transferring rich feature hierarchies for robust visual tracking. arXiv:1501.04587.

Wang, Z., Simoncelli, E.P., Bovik, A.C., 2003. Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Ieee. pp. 1398–1402.

Westlake, N., Cai, H., Hall, P., 2016. Detecting people in artwork with cnns. In: European Conference on Computer Vision, pp. 825–841.

Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A., 2010. Sun database: Large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3485–3492.

Yang, F., Choi, W., Lin, Y., 2016. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2137.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks?. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 3320–3328.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595.

Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R., 2016. Unconstrained salient object detection via proposal subset optimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5733–5742.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell..