# Color and Shape efficiency for outlier detection from automated to user evaluation

Loann Giovannangeli, Romain Bourqui, Romain Giot, David Auber



Version of Record: https://www.sciencedirect.com/science/article/pii/S2468502X22000146
Manuscript_fcc86b81667b942d28450cd9bfdf3a67

# Color and Shape efficiency for outlier detection from automated to user evaluation

Loann Giovannangeli[a,*], Romain Bourqui[a], Romain Giot[a], David Auber[a]

[a] *Univ. Bordeaux, CNRS, Bordeaux INP, INRIA, LaBRI, UMR 5800, F-33400 Talence, France*

**Abstract**

The design of efficient representations is well established as a fruitful way to explore and analyze complex or large data. In these representations, data are encoded with various visual attributes depending on the needs of the representation itself. To make coherent design choices about visual attributes, the visual search field proposes guidelines based on the human brain perception of features. However, information visualization representations frequently need to depict more data than the amount these guidelines have been validated on. Since, the information visualization community has extended these guidelines to a wider parameter space.

This paper contributes to this theme by extending visual search theories to an information visualization context. We consider a visual search task where subjects are asked to find an unknown outlier in a grid of randomly laid out distractors. Stimuli are defined by color and shape features for the purpose of visually encoding categorical data. The experimental protocol is made of a parameters space reduction step (*i.e.*, sub-sampling) based on a machine learning model, and a user evaluation to validate hypotheses and measure capacity limits. The results show that the major difficulty factor is the number of visual attributes that are used to encode the outlier. When redundantly encoded, the display heterogeneity has no effect on the task. When encoded with one attribute, the difficulty depends on that attribute heterogeneity until its capacity limit (7 for color, 5 for shape) is reached. Finally, when encoded with two attributes simultaneously, performances drop drastically even with minor heterogeneity.

*Keywords:* Visual search, Outlier detection, User evaluation, Deep learning, Automated evaluation

## 1. Introduction

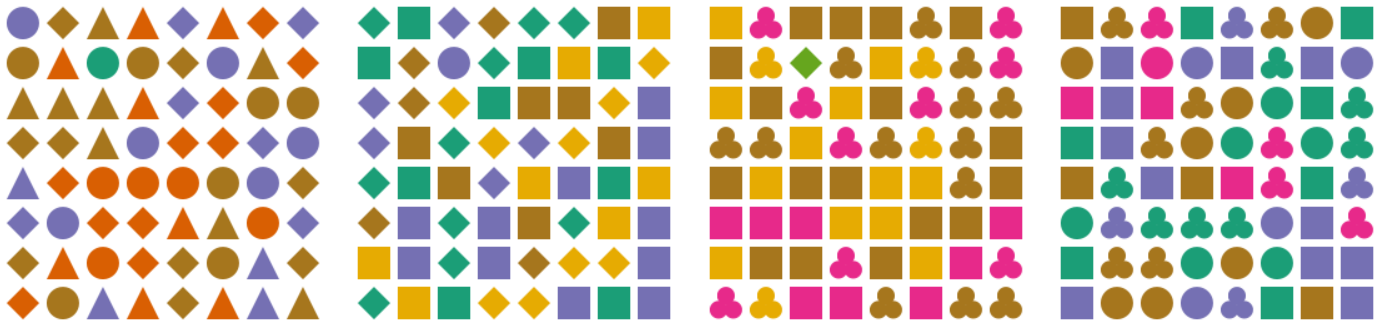
One of the main goals of the information visualization research field is to ease the search for data that is not trivially queryable. It is achieved by designing abstract representations of these data that can easily be explored to enable users/experts to extract knowledge they were not specifically looking for (otherwise, a query in a database would be sufficient).

To design efficient representations, experts must optimize how they encode data (*i.e.*, select visual attributes and features). As shown by Ware [1] or Healey and Enns [2], these visual choices should be driven by visual search and perception guidelines. Since then, various recommendations about visual attribute efficiency have been produced to help experts in their choices when highlighting data in their representations [3, 4, 5, 6, 7].

Color [8, 9, 3] and shape [10, 11] are two widely used and often combined visual attributes for encoding data in representations (*e.g.*, scatter plots[12], geographic maps [13], graphs [14], and parallel coordinates [15]). However, it is often unclear how well these visual attributes remain efficient as visualizations become increasingly complex (*e.g.*,, the number of data items or classes to represent increases). For example, Perception researches [16] shown that color was a *preattentive feature*, meaning that all colors of a representation could be processed in parallel. Yet, heterogeneous representations tend to overwhelm the search process at some *point*, even when data are encoded with color. This *point* is known as the *capacity limits of attention* and varies according to data encoding. For color, the capacity limit is assumed to be around $7 \pm 2$, though we couldn't find any reference that support it. This seems to be a common misinterpretation of the Miller's magic number [17] which measures the number of categories we can *distinguish* but not necessarily *remember*. We think this limit is over optimistic and that the difficulty of a search task might significantly increase with less colors in dense representations, especially when data are encoded with combinations of visual attributes. This work aims at verifying such assumption and more generally measure the capacity limits of attention when data are encoded with color and/or shape in dense representations. It is important to note that we are interested in the maximum number of features a visual attribute features can take in a representation, no matter what the features of this attributes are. Since it is not feasible to test every possible feature of the considered attributes, the results of this work should be observed within the scope defined by its experiment, *i.e.*, its set of selected shapes and colors.

In this paper, we study how difficult it is to find an outlier in representations with various heterogeneity and outlier encoded using *shape* and *color* visual attributes. There is always exactly

*Corresponding author
*Email addresses:* loann.giovannangeli@labri.fr (Loann Giovannangeli), romain.bourqui@labri.fr (Romain Bourqui), romain.giot@labri.fr (Romain Giot), david.auber@labri.fr (David Auber)



(a) *color* type; 3 shapes; 4 colors (b) *shape* type; 3 shapes; 4 colors (c) *red.* type; 3 shapes; 4 colors (d) *conj.* type; 3 shapes; 4 colors

Figure 1: Experimental object examples for the 4 possible *types* of outliers. The outlier is always at position 10 (*i.e.*, second row, third column from the top left corner) in these examples. (a) In *color* type images, the outlier color is unique; (b) in *shape* type images, the outlier shape is unique; (c) in *redundant* type images, both the outlier color and shape are unique; and (d) in *conjunction* type images, the outlier combination of color and shape is unique.

one unknown outlier in each representation to reproduce a standard data exploration where ones would like to identify outlier elements. Outlier encodings (later referred to as *Type*) are the four possible combinations of color and shape dimensions (see Figure 1) and we study the effect of heterogeneity within each *Type* condition separately. Our results show that the level of heterogeneity at which the task becomes significantly hard is lower than one could expect and confirm that it depends on the outlier encoding [18, 19].

As opposed to visual search researches, our work focuses on the study of *what* makes the task becomes harder rather than *why*. In fact, even forty years after the pioneer work of Treisman and Gelade [18] "*Feature-Integration Theory of Attention*" (FIT), the Perception community still proposes several competing templates to explain the brain strategy to process visual search tasks [20, 21]. We are interested in studying how some widely used visual attributes features in information visualization actually make the task more complex for human subjects from a practical point of view.

In the following, we adapt some terminology from the literature to align ourselves with the information visualization community. We call *parameters* what Treisman (*e.g.*, [19, 18]) calls *dimensions*, we use *value* instead of *feature*, *outlier* instead of *target* and *distractor* instead of *nontarget*. A *parameter* therefore refers to either a visual attribute (*e.g.*, color, shape) or any other variable of the representation (*e.g.*, number of shapes), both of which having *values* (*e.g.*, color: red, shape: circle, number of shapes: 4).

In our work, we study the capacity limits of human attention with varying heterogeneity in color and shape attributes, both alone and combined. Among the information visualization field, the closest work to this paper is Haroz and Whitney [3] who studied the *capacity limits* of human attention on representations with varying heterogeneity in terms of *color* or *motion* under grouped or random layouts. However, they did not study the combination of these two visual attributes, and results about motion have limited application since many representations are not dynamic. Even where conditions are the closest between their study and ours (*i.e.*, data encoded with color only, with an unknown outlier in a random grid), our subjects performances are significantly worse than theirs, which indicates that our task is not at the same level of difficulty.

Another originality of our work is the use of a Deep Neural Network (DNN) as a metric to base the sub-sampling of the experiment parameters on. As for many experiment, the combination of all the considered parameters is too high to be exhaustively studied through a user evaluation. In some research fields, there are dedicated metrics to evaluate the readability of a representation (*e.g.*, graph drawing [22, 23]). But every representation technique do not have a set of metrics to measure its quality for solving various tasks. On the other hand, DNNs have proven to be very efficient computer vision techniques and are capable of learning to solve a wide variety of tasks on representations. Here, we leverage DNNs as they can (and *must*) process large datasets to learn a task. In fact, these models can learn to solve any task on any representation, as long as both can be programmatically expressed. Moreover, they get better as they see more data samples, which enable to evaluate a high number of combinations of parameters. The DNN predictions are then statistically analyzed to identify the conditions of the parameters space that make it consistently fail (*resp.* succeed) its predictions. We interpret this as a *difficulty metric* based on the assumption that easy (*resp.* hard) conditions for the DNN tend do be easy (*resp.* hard) for users as well. Finally, this *difficulty metric* is used to lead a parameters space pruning. The assumption of correlations between DNNs and humans on perception tasks was extensively studied in some works [24, 25, 26] and we specifically validated it in our context in a separated study [27]. This assumption is also intrinsic to all aesthetic metrics as no function can efficiently evaluate every task difficulty for humans on every possible representation, which is why metrics results should be interpreted by informed experts.

To summarize, the main contribution of this paper is the study of the capacity limits of *color* and *shape* visual attributes when mixed in representations with tens of stimuli. We also propose the idea of a computable *difficulty metric* relying on Deep Neural Networks to assess representations efficiency and that can be adapted to any task-representation a DNN can learn. Here, the difficulty metric is used to sub-sample a parameters space. Finally, a user evaluation is conducted on the reduced

space to study preliminary hypotheses about the outlier detection task. The validity of using the statistical analysis of a Deep Learning model performances to assess users performances was the object of another paper [27] and is discussed in Section 4.1.

The remainder of this paper is structured as follows. Section 2 presents related work from the visual search literature and its relation to information visualization. Section 3 presents the task parameters space definition and experimental objects dataset generation, while Section 4 describes the sub-sampling of this parameters space. Section 5 presents the experimental evaluation setup and results. Finally, Section 6 discusses these results, and Section 7 draws conclusions and presents future work leads.

## 2. Related Work

The task in this experiment relates to the identification of a target stimulus in a background of nontarget stimuli, and has various application domains. In perception, researchers have used this task to understand how the brain processes displays. In the field of information visualization, studies are focused on optimizing the time required to solve the task. In this section, we present some literature on the two domains. We also present recent works about the use of deep neural networks (DNNs) for evaluating the readability of representations.

### *2.1. Visual Search in Perception*

The seminal theory in the visual search research field is the *Feature-integration theory of attention* (FIT) by Treisman and Gelade [18]. It defines attention as a two-stages system with a *preattentive* step followed by an *attentive* process. Some visual attributes would then be considered *preattentive* if its features could be processed in parallel [16]. The theory also distinguishes *feature search*, where the brain looks for a feature of a single attribute, from *conjunction search*, where combination of different visual attribute features are required to identify a target (*e.g.*, binding of features). However, this template is now contested and the Guided Search of Wolfe [28], which has been regularly updated (today version is 4.0), is now preferred. Nonetheless, the *FIT* was essential and engaged researches on perception and visual search as shown in the tribute to Treisman contributions [20, 21].

Duncan and Humphreys [29] leveraged Treisman theory to propose theirs, based on stimuli similarities and templates. They showed that as the target to nontarget (T-N, *i.e.*, outlier to distractor) similarity increases, the task becomes more difficult. This situation is even worse if the nontarget to nontarget (N-N) similarity increases, except in cases where the T-N similarity remains small. Furthermore, the number of possible nontargets in the representation, which they called *nontarget heterogeneity*, severely affects the task difficulty. Finally, they stated that if a target can be identified by a specific dimension (*i.e.*, a relevant dimension), heterogeneity in other dimensions (*i.e.*, irrelevant dimensions) should only have a minor impact on the search task; this corroborated the results of Treisman [19]. Pashler [30] also studied heterogeneity in irrelevant dimensions and shown that it had no effect even when the target was unknown. In view of these two works, our experiment should enable to observe the harmful effects of *nontarget heterogeneity*, which is attenuated when it occurs in irrelevant dimensions.

Quinlan and Humphreys [31] found that the visual search of a target defined by shape is slightly linearly related to the total number of stimuli, whereas the latter has no impact on a target defined by color. This corroborates that color is a preattentive attribute and is more efficient to represent data than shape. For conjunction search, they found that error rates increase with the number of stimuli and that response time is linearly related with it. Moreover, they showed that in conjunction search, T-N similarities have more impact on subject performances than in single feature search. Finally, they pointed out that the more features the outlier shares with the distractors, the more difficult the task is.

### *2.2. Visual Search in Information Visualization*

If the perception research field is a cornerstone of the information visualization community, their results do not always apply to information visualization. For example, Treisman and Gelade [18] claimed that *"we cannot normally locate an item which differs from a field of distractors without also knowing at least on which dimension (color or shape) that difference exists"*, but their experiment was made of trials for which the time limit was set to 3 seconds. Such design scale is far from meeting the complexity of most of the modern representations. To that extent, Healey and Enns [2] drew a landscape of the visual perception literature that was dedicated to computer graphics applications; and the information visualization community has kept running its own measurements of humans processing of representations efficiency.

Haroz and Whitney [3] studied how *colored groups* and *motion* influence the effectiveness of information visualizations. They conducted several experiments on 5 subjects that had to solve a target-present task on 960 trials each. Some parameters, such as the number of colors or the layout of the color groups, were studied. They found that grouping colors (*i.e.*, classes) significantly eased the task when the target to find was unknown. Moreover, when colors were grouped, it was easier to access overall information, such as the total number of colors/classes. Inspired by [3], Gramazio *et al.* [4] studied how the same task was sensitive to representation size by varying the number of stimuli, their layout, their size and the number of colors in representations. In our work, we study the same task when stimuli are encoded with *color*, *shape* or a combination of both in randomly laid out representations.

Demiralp *et al.* [32] introduced the notion of a *perceptual kernel*, a distance matrix that represents the perceived distances between members of a set of stimuli composed of one or several visual attributes. In their experiment, they estimated the perceptual kernels for the color, shape and size visual attributes, as well as their pairwise combinations. They showed that color and shape have very different kernels. In the shape kernel, we observe several *distant* clusters of *close* shapes, whereas the distances between colors are more evenly distributed. On the other hand, all stimuli are close to many others in the color-shape

kernel. Their experiment considered 4 colors and 4 shapes (*i.e.*, 16 stimuli), but only 4 clusters could be distinguished in the kernel, meaning that all stimuli had high levels of similarity with others. We expect that varying the number of shapes or colors should not have the same effect on the performances in our experiment as their kernels are different; and using conjunction of both attributes should have a significant impact.

According to Mackinlay [7], *position* is the best parameter for visually encoding data in representations. For example, in Western culture, as one reads from left to right and top to bottom, one could assume that cells placed in the top left corner of the grid are processed first. On the other hand, a central fixation bias [33] could favor stimuli in the middle of the grid. In this experiment, the stimuli layout in the representations is fixed ($8 \times 8$ regular grid) as we do not aim at studying the impact of the *outlier position* on the participants performances. To mitigate its impact on the results, outlier positions are uniformly distributed in the dataset during both the deep neural network model learning phase (see Figure 2a) and the user evaluation (see Figure 5a).

### 2.3. Neural Networks for Visualization Evaluation

Behrisch *et al.* [34] conducted a recent survey on quality metrics for information visualization and claimed that deep neural networks (DNNs) were a promising direction for evaluating the quality of a representation. On the same line of research, Haehn *et al.* [35] reproduced the Cleveland and McGill [36] study with different convolutional neural networks (CNNs) to evaluate how these networks performed compared to humans on various elementary graphical perception tasks (*e.g.*, position relative to a scale, angle, or area). They found that CNNs and humans behave differently on these elementary graphical perception elements but were still enthusiastic about evaluating representations with DNNs. Later, Haleem *et al.* [37] trained a CNN to predict various graph node-link representation quality metrics while feeding it with laid out graph images only (*i.e.*, the CNN did not have access to the node coordinates, edges, etc.). Their model reached an accuracy above 85% at a 95% confidence level. These quality metrics were designed to encode some graph readability information for humans (although we already noted that they do not always accurately reflect human perception capabilities). Their study proved that CNNs *can* strongly approach them and thus efficiently estimate human perception capabilities. Finally, Giovannangeli *et al.* [38] partially reproduced two evaluations comparing node-link to adjacency matrix graph representations [39, 40] with CNNs on counting and connectivity tasks. They proposed an automated method to compare visualization techniques and concluded that humans and machine-learning-based computer vision techniques can be correlated on the tasks they considered.

All these studies remained cautious about their results and raised several limitations. The task definition, data generation process, network architecture, hyperparameters, initial weights, etc. can lead to different network strategies and performances. As this research field was recently developed, it is still not well understood how CNNs and humans can be correlated, and we currently know more about their differences than correlations.

Table 1: All parameters values considered in this study. Color values are given as hexadecimal RGB codes. Shape and color values can be used by either the outlier, through the *outlier color* and *outlier shape* parameters, or by the distractor stimuli.

| Visual attribute values | | | Image | | |
|---|---|---|---|---|---|
| **Shape** | **Color** | **Position** | **Type** | **#colors** | **#shapes** |
| ▲ | #1B9E77 | 0 | color | 1 | 1 |
| ● | #D95F02 | ⋮ | shape | 2 | 2 |
| ■ | #7570B3 | 63 | redundant (red.) | 3 | 3 |
| ♣ | #E7298A | | conjunction (conj.) | 4 | 4 |
| ◆ | #66A61E | | | 5 | 5 |
| | #E6AB02 | | | 6 | |
| | #A6761D | | | 7 | |

## 3. Task and Parameters Space

This section details the parameters (and their values) considered in this study.

### 3.1. Task

The chosen task consists of identifying an outlier in an $8 \times 8$ grid of colored shapes drawn in an image of $256 \times 256$ pixels. These image properties enable (i) the consideration of a reasonable number of values for our key parameters (presented immediately after) and (ii) a good trade-off between image readability for a user and the possibility of feeding the image to a standard deep learning model architecture. In such an image, a colored shape (*i.e.*, stimulus) is considered an outlier if there is no other stimulus with the same *color* and *shape* visual attributes. The dimension(s) on which the outlier is made unique varies according to the *type* parameter.

**Type** relates to the dimension(s) that makes the outlier unique. It has 4 possible values: (i) *color*, when the *color* of the outlier is unique in the grid; (ii) *shape*, when its *shape* is unique in the grid; (iii) *redundant*, when *both* its color and shape are unique in the grid, (this refers to redundant encoding [41]); and (iv) *conjunction*, when its color-shape *combination* is unique in the grid. Examples of type values are provided in Figure 1.

Each image contains exactly one outlier and 63 distractors. A colored shape is considered a distractor if it appears at least twice in the grid (otherwise, it is an outlier). There are at most 31 different color-shape combinations for distractors in a grid.

### 3.2. Data Space Definition

The experimental objects of this study are images representing a grid. They are defined by six parameter values (see Table 1).

**Outlier shape** values are chosen among a set of shapes (see Table 1 column 1). A shape can be defined by many sub-features (*e.g.*, lines, orientation, size). In this experiment, every shape appears in a single orientation, and its size is set to the maximum value to fit in a $32 \times 32$ pixels cell using a 3 pixels padding. Five shapes are selected – Triangle ▲, Circle ●, Square ■, Clover ♣ and Diamond ◆ – to mix the use of straight vertical/horizontal, diagonal and curved lines.

**Outlier color** values are chosen among a set of colors (see Table 1 column 2). Some methods already exist for finding an efficient color set to represent targets (*e.g.*, [42], or more recently

Colorgorical [43]). In this experiment, colors are considered as basic features (*i.e.*, we do not study the effects of hues or saturation) and are chosen from the *7 qualitative classes* palette named *Dark2* in the ColorBrewer[1] tool [44], a well-known color palette provider. The palette is *qualitative* as the colors should be as independent as possible (*i.e.*, categorical) in this experiment and *Dark2* was selected because it was one of the proposed sets with highest saturation. From the beginning, we planned to exclude colorblind individuals. It is complicated to find reliable color palettes of this size and for which colors are as distinguishable for non-colorblind individuals as they are for colorblind ones (which can themselves be of different types).

The **outlier position** relates to the position of the outlier in the experimental object. In this study, the position varies between 0 and 63 corresponding to the row-major order of the grid.

The **type** of an image relates to the outlier dimension(s) that makes it unique in that experimental object (see Section 3.1).

The **number of colors** (*#colors*) relates to the total number of distinct colors used in an experimental object. In this experiment, the number of colors varies between 1 and 7. It is noteworthy that if an experimental object type is *color* or *redundant*, the number of colors cannot be set to 1 as a color must be reserved for the outlier.

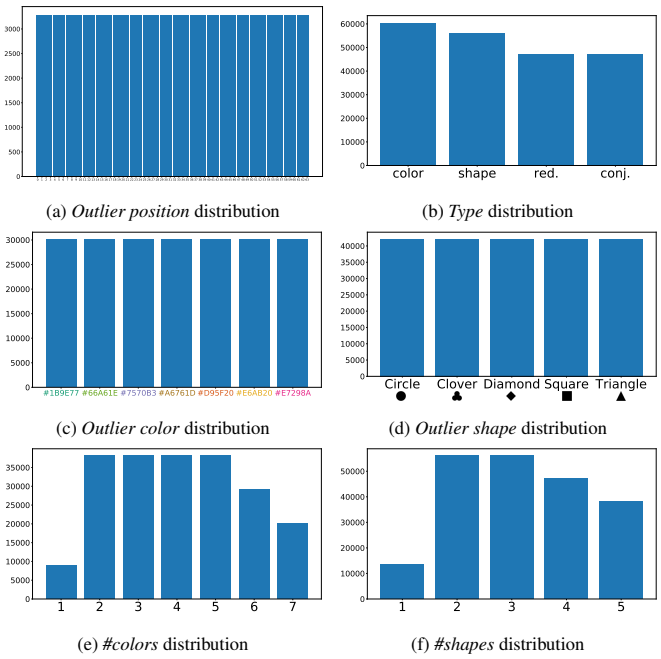
The **number of shapes** (*#shapes*) relates to the total number of distinct shapes used in an experimental object. In this experiment, the number of shapes is between 1 and 5. For an experimental object of type *shape* or *redundant*, the number of shapes cannot be set to 1 as one is reserved for the outlier.

### *3.3. Dataset Generation*

The six parameter values were balanced to minimize distribution bias and train the model correctly when generating the experimental dataset. The main concern was to balance outlier shape-color-position occurrences (see Figure 2a) to prevent the deep learning model from learning to find some stimuli or locations more easily than others because they were more common in the dataset.

The generation process also followed some constraints. Obviously, images with 1 color and 1 shape could not be generated, but less evident cases could not be generated either. Images of type *redundant* cannot be generated with either 1 color and 2 shapes or 2 colors and 1 shape. For images of type *conjunction*, the combinations of parameter values using 1 color or 1 shape were not considered, as they would result in images of type *shape* or *color*, respectively. In addition, type *conjunction* images cannot be generated using 7 colors and 5 shapes in an $8 \times 8$ grid. One of the $7 * 5 = 35$ combinations should be reserved for the outlier, and 34 should appear twice (minimum condition to be a distractor), so this would lead to at least 69 stimuli.

The configurations using[4 shapes, 7 colors] and [5 shapes, 6 colors] were removed. From our experience, knowledge of the literature and pilot experiments, we strongly expect that the capacity limits we aim to study will be reached before (*i.e.*, with less heterogeneity) these high-valued configurations.




(a) *Outlier position* distribution (b) *Type* distribution

(c) *Outlier color* distribution (d) *Outlier shape* distribution

(e) *#colors* distribution (f) *#shapes* distribution

Figure 2: Parameter value distributions in the 210 560 images.

These constraints explain why the *type*, *#colors* and *#shapes* values are not fully balanced, as shown in Figures 2b, 2e and 2f.

By generating one image per combination of parameter values (see Table 1) while excluding those described above, we ended up with 210 560 different images. As stated in Section 1, this study was not designed to investigate the effect of the *outlier position* on the task. This parameter was only used to generate several samples with other parameter value combinations and is balanced uniformly to mitigate its consequences on the results. Therefore, the experiment studies 3290 different parameter combinations, repeated 64 times each.

Finally, the dataset was randomly split into 3 subsets for supervised learning purposes (*hold-out* validation [45]): *train* (to learn the model), *validation* (to prevent overfitting during training) and *test* (to evaluate the model on unseen data).

## 4. Parameters Space Reduction

The parameters space presented in Section 3 is too large (3290 conditions $\times$ 64 positions) to directly evaluate the task through a user evaluation. To overcome this, we sub-sample the parameters space based on the analysis of a Deep Neural Network (DNN) performances on the same task-representations. This section presents the design of DNN *as a metric* and its implementation: how we selected our network architecture and trained it to solve the task. Then, it presents the trained model results as well as the statistical study that drove us to refine our hypotheses and sub-sample the parameters space.

[1] `https://colorbrewer2.org/#type=qualitative&scheme=Dark2&n=7`, consulted on October 2021

### 4.1. DNN as a metric

As we have seen in Section 3, the complete parameters space is too large to conduct a user evaluation. A common approach to overcome this issue is to sub-sample it so that we do not need to evaluate the task on all possible conditions, and yet can generalize the experiment results to the whole parameters space. However, such *approximation* only remains correct if the sub-sampling method is representative for the task, which is difficult to know before conducting the experiment.

Two common sub-sampling approaches are usually accepted: (i) arbitrary sampling (*e.g.*, random, systemic) for which conditions are pruned based on pilot experiments, literature and beliefs ; and (ii) metric-based sampling. For reproducibility and objectivity reasons, the second method seems preferable to the first one. However, there rarely already exists a metric designed to evaluate new representations effectiveness to solve given tasks, unless the problematic is very specific (*e.g.*, graph aesthetic metrics for graph drawing [22, 23]).

Inspired by the Giovannangeli *et al.* [38] method and following the recommendations of Haehn *et al.* [35] and Haleem *et al.* [37] (see Section 2.3), we propose a novel approach based on a deep neural network to compute a difficulty metric. It assesses how difficult a task is to solve on a given representation, based on the task and representation parameters themselves. The concept of *DNN as a metric* is quite intuitive. The first step is to generate annotated data for training the DNN model to learn to solve the task (see Section 3.3). Following the recommendations of [38], hundreds of thousands of data samples are generated while trying to keep the parameters distribution uniform. The objective is to ensure (i) that the model truly learns to solve the task and does not learn the ground truth distribution, and (ii) that the model does not perform better with a given parameter value because it has been seen more often in the training dataset. Then, a generic DNN architecture is trained solve the task, and we keep its tuning to a minimum to avoid biasing the model with any a priori belief we could have about the task difficulty. The obtained model is then evaluated to ensure it learned to solve the task so we can analyze its performances. Its performances are aggregated in different ways to statistically study the effect of each parameter on the task difficulty for the model. The outcome of the statistical study is finally used as a difficulty metric. Based on this metric, the parameters space is reduced for the user evaluation (see Section 5).

The main advantages of this metric design are: (i) it fits any task and representation that can be programmatically expressed, and (ii) it does not require any a priori information about the task. The model learns by itself what *areas* and *graphical elements* of the representation are relevant for solving the task.

The major concern of this approach is that it considers the DNN as a *meta-user* and assumes its performances are correlated with human users ones. To study this assumption, we conducted an *a posteriori* study of the correlations between the DNN and human participants performances gathered in our experiment. This correlations study was the object of an other publication [27] and concluded that the DNN and human participants were strongly correlated (up to 0.988 correlation score, a perfect correlation being a score of 1) and gave better insights on how to interpret the model results to assess humans performances. Yet, every aesthetic metric is sort of heuristic of human perception and cannot always successfully model humans perception system, and this approach is no exception in that matter. It is neither more accurate or less correct than other sub-sampling approaches, and should be interpreted by informed experts only. Nevertheless, it does enable to study broad parameters space of any task and representation that can be programmatically expressed.

### 4.2. Model Selection and Training

As mentioned in [38], a generic deep neural network (DNN) should be used rather than an architecture dedicated to the task(s) (or visualization technique(s)) to be studied. To that extent and following the recommendation of [35], we tried several network architectures (*e.g.*, *LeNet* [46] or *VGG-16/19* [47]) and finally selected *ResNet* [48] as it correctly learned to solve the task. The default weights of *ResNet* were set to their pretrained values on ImageNet [49]. He *et al.* [50] showed that such a model pretrained for image recognition already encodes some saliency information, which is expected to speed up the learning process with regard to spatial identification.

The *ResNet* architecture was slightly tuned [48]: its input layer was set to fit the generated image resolution, and two successive dense (*i.e.*, fully connected) layers were added after its output to fit the required number of classes for prediction. We consider the identification of the outlier as a classification problem rather than a regression problem, where there would be a notion of distance between the predictions and their ground truths. The size of the last dense layer was therefore set to predict the outlier position (*i.e.*, to predict 64 classes) and the size of the penultimate dense layer was set to 1024.

While the optimizer and default tuning of the learning phase were not modified, the batch size was set to 64 (instead of 256). We used the *early stopping* function of the Keras library [51] with a *patience* of 15 epochs to end the training process.

### 4.3. Results

At the end of the learning phase, the best epoch accuracy rates on the *validation* and *test* sets reached 74% and 76%, respectively, showing that the model has not overfitted and is able to generalize. A *Matthews correlation coefficient* [52] of 0.754 on the *test* set confirms that the model learned to solve the task. Thus, we can expect that incorrect predictions are not due to hazards but rather combinations of parameter values from the data.

A Kruskal-Wallis ANOVA test [53] was conducted on each parameter value prediction sequence to verify whether they had significant effects on the performances (overall or on a specific *Type* value). For the parameters found to have significant effects (opaque plots in Figure 4), pairwise Wilcoxon rank-sum tests were conducted to check if their values led to significantly different performances. The significance level for the overall studies was set to $\alpha = 0.05$. When splitting the data *by type*, a Bonferroni correction was applied, reducing the significance

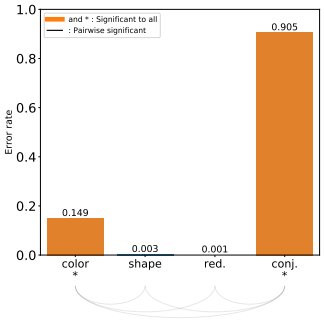



Figure 3: Trained *ResNet* error rates (ER) on the *test* set for the *type* values. An arc between two labels means that the pairwise comparison between the ER of the two parameter values is significant ($p$-*value* $< 0.05$) according to a Wilcoxon rank-sum test. When performances associated to a label are significantly different to all others: the bar is colored orange and the symbol * is added below the label. *Reading example*: the *color* type is significantly easier than the *conjunction* type and significantly harder than the *shape* and *redundant* types. There is no significant difference between the *redundant* and *shape* types.

level to $\alpha = 0.025$. All parameters showed to have a significant effect under at least one condition, except *outlier shape* which will not be studied further in this section.

For the remaining parameters, Figures 3 and 4 present the trained model error rates (ERs) on the *test* set. Next, we describe the main insights we can learn from these results.

**Type:** As we could expect, *type* is a key parameter relative to the task difficulty, as shown by the large differences between the ERs in Figure 3. The *conjunction* type led to significantly more errors than other values. The *color* type is significantly harder than both *shape* and *redundant*; these last two not being significantly different from each other. The significant gaps between the *type* values motivated the study of other parameters for each *type* value separately (see Figure 4). We also see that experimental objects of the *color* type led to a significantly higher ER than those of the *shape* type, which is surprising in view of the visual search literature that considers *shape* as a harder visual attribute than *color*. This is most likely induced by the design of *ResNet* architecture and will be discussed in Section 4.4.1.

**Number of colors:** Overall, the ER almost linearly increases as *#colors* increases, as shown in Figure 4. A significant shift in performance between 1 and 2 *#colors* (and basically, between 1 and any other value) can be observed. This shift was probably induced by a bias in our data generation process, which will be discussed in Section 4.4.1. The increase in difficulty is even stronger as *#colors* raises on experimental objects where color is the only relevant dimension for identifying the outlier (*i.e.*, *color* type). When color is not a relevant dimension (*i.e.*, type *shape*), *#colors* does not have any significant effect on the task difficulty. Finally, for experimental objects of type *conjunction*, the only significant differences in performances are between *exactly 2* and $> 2$ *#colors*, meaning that the task difficulty is *thresholded*. Beyond 2 colors, it seems that the task is already so hard to solve that further increasing the number of colors does not make the task significantly harder.

**Number of shapes:** Figure 4 shows that, overall, there is a significant ER shift between 1 and higher values of *#shapes*, as it was observed with *#colors*. Again, this will be discussed in Section 4.4.1. This value set aside, there remains only one significant difference between the other *#shapes* values. Hence, we can assume that, overall, increasing the number of shapes does not significantly increase the task difficulty. When shape is an irrelevant dimension for identifying the outlier (*i.e.*, type *color*), the ANOVA test reveals that error rate differences between *#shapes* values are significant. This result is counterintuitive as the outlier *cannot* be guessed using the shape dimension in *color* type images. It might reveal some sort of overfitting from the Deep Neural Network, and we must be careful while using the DNN results on that condition. When shape is the only relevant dimension (*i.e.*, *shape* type), the ANOVA test indicates that ER differences are significant between *#shapes* values, but no pairwise significant difference is found with post-hoc test. Since ERs remains under 1%, we could assume that the task is very easy on *shape* type images whatever the number of shapes is (*i.e.*, heterogeneity in a relevant dimension has no effect). However, this is most likely not how human participants would perceive the task difficulty and is probably the source of the uncorrelation between the DNN and the participants that was observed in [27] on *shape* type images; and which we discuss in Section 4.4.1. Finally, the ER of *#shapes* follows the same trend as that of *#colors* for experimental objects of the *conjunction* type: the difficulty is thresholded between *exactly 2* and $> 2$ *#shapes*.

**Outlier color:** The *outlier color*, unlike *outlier shape*, does impact the task difficulty in some conditions. However, as we can see in Figure 4, ANOVA tests only indicate that *outlier color* variations impact the task Overall and on *color* type images. The Overall data being the aggregation of the 4 *type* values, it mostly means that *outlier color* only had an effect on *color* type images, which makes sense as color is the only relevant dimension on that condition. On the opposite, when color is an irrelevant dimension (*i.e.*, *shape* type), *outlier color* variations has no significant effect. On *redundant* type images, the outlier can be found using either its color or its shape (or both). As presenter earlier, *outlier shape* never had any effect on the DNN performances. Since *outlier color* does not have any significant effect on *redundant* type condition either, it means the task is very easy, no matter what the attributes that define the outlier are. With the same reasoning, we can conclude for *conjunction* type images that the task is very hard, no matter what the attributes of the outlier are.

In this section, we did not focus on parameter value effects on the experimental objects of the *redundant* type. As we can see in Figure 4, no parameter had any effect on the experimental objects of this type. The overall ER of the *redundant* type is 1%, and all *#colors*, *#shapes* and *outlier color* ER values are under 1%. We conclude that there is no univariate condition that affects experimental objects of type *redundant*.

### *4.4. Results Interpretation*

#### *4.4.1. Limitations*

As mentioned in Section 4.1, aesthetic metrics cannot exactly model human perception of a task difficulty on some rep-

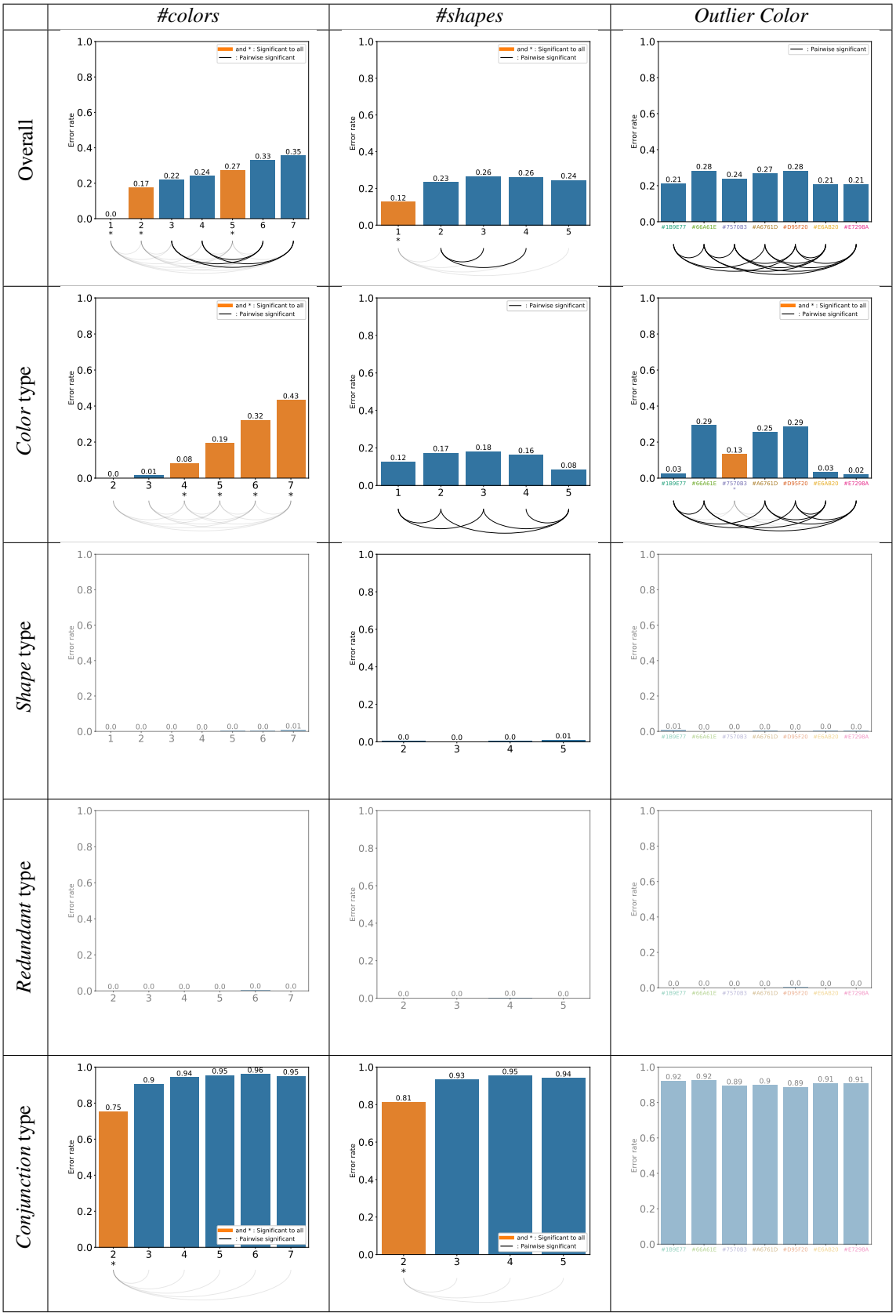



Figure 4: Trained *ResNet* error rates (ER) on the *test* set. The first row shows the overall parameters ERs, while the next rows present the parameters ERs by *type*. A plot is faded if the ANOVA test on its given parameter and type aggregation failed, meaning the parameter values variations did not have any significance effect on the condition; otherwise, it is opaque. An arc between two labels means that the pairwise comparison between the ERs of the two parameter values is significant according to a Wilcoxon rank-sum test. When performances associated to a label are significantly different to all others: the bar is colored orange and the symbol * is added below the label. The significance threshold was $p\text{-value} < 0.05$ for the ANOVA and pairwise tests in the *overall* studies, while it was $p\text{-value} < 0.025$ in the *per type* studies. A reading example is given in the caption of Figure 3.

resentations and there most likely exist conditions for which the metric diverges from what humans would effectively be capable of achieving. In the study of correlations we conducted separately [27], we also found that the DNN Error Rate was not the best correlated metric with the participants performances, though the correlation scores with participants Error Rate and Response Time remained significantly high enough (respectively 0.806 and 0.903) to validate the DNN based metric in this experiment. Since the computation of correlations requires both the model and human participants performances, it could not be done *a priori* and such conclusion was expected.

Another limitation comes from the performances we observe on Type *shape* objects and that do not corroborate with human participants performances, as shown in [27]. The model learned a strategy that does not correlate with human behaviors. We suspect that when the outlier was identifiable by its shape, the model simply estimated the number of colored pixels in each cell (regardless of the color itself) and predicted the only colored pixel counts that did not appear at least twice, which greatly favored Type *shape* and *redundant* objects. This is why we need to monitor correlations with humans performances and take them into consideration when interpreting the metric results.

It can also be noted that the *black-box* effect of CNNs makes us unable to explain why *outlier color* had a significant effect on *color* type images, while *outlier shape* did not significantly affect *shape* type images; especially considering that the model achieved better performances on *shape* type images (see Figure 3). In addition, since we consider colors as basic features in this study (*i.e.*, do not decompose into hue, saturation, etc.), we are not interested in "what" makes an *outlier color* better than another to efficiently solve the task. Moreover, the easier *outlier colors* for the DNN would be based on divergences in RGB space, whereas we know RGB is not a good representation of the human perception of colors.

Finally, as stated in Section 3.3, some parameter configurations could not be generated. The bars corresponding to the error rates for both 1 *#colors* and 1 *#shapes* were only computed from a specific type value (respectively *shape* and *color* type). In the Overall, *#colors*, ER plot in Figure 4, "1" has a lower ER than other values since it is only composed of experimental objects of the *shape* type, whereas other values are computed from experimental objects of all type values.

Thus, we should bear in mind these limitations when interpreting the model performances for assessing the task difficulty.

#### 4.4.2. Hypotheses

In the next, we present the hypotheses studied in the user evaluation with human participants, partly inferred from the model performances. They are built based on knowledge from the literature and include some of the *DNN metric* results to confront them with human participants performances. Again, the study of correlations between the DNN and human participants performances [27] has been done *a posteriori* since it required to gather human participants data, which is why knowledge about these correlations could not be taken into account at the time we built the hypotheses. Nevertheless, as for any metric we expected it to have pitfalls where its results would not corroborate with human perception.

$H_{type}$: **The type difficulty should follow the order (easiest to hardest): *redundant*, *shape* and *color*, *conjunction*.** This refers to the contribution of Quinlan and Humphreys [31], who showed that the more features the target shares with the nontargets, the more difficult the task is.

$H_{conj}$: **The search task on experimental objects of type *conjunction* is the hardest among all types. The task difficulty increases with both *#colors* and *#shapes* and quickly caps** (*i.e.*, the difficulty no longer increases when there are more than 2 shapes and colors, see Section 4.3). The fact that conjunction search is harder than feature search was confirmed by prior work [29, 18].

$H_{red}$: **The search task on experimental objects of type *redundant* is the easiest among all types. The task difficulty is not affected by neither *#colors* nor *#shapes***, as shown by Nothelfer *et al.* [41] and according to the results in Section 4.3.

$H_{color}$: **When color is the only relevant dimension, the task difficulty increases with *#colors*, whereas *#colors* has no effect when color is not a relevant dimension.** These assumptions corroborate both the DNN results and statements from the literature, hence we expect this hypothesis to be accepted.

$H_{shape}$: **When shape is the only relevant dimension, the task difficulty increases with *#shapes*. When it is not a relevant dimension, *#shapes* should have no significant effect.** These assumptions follow statements from the literature and go against the DNN results as we have seen that the DNN behavior on experimental objects of *shape* type could hardly be trusted.

#### 4.4.3. Parameter Space Sub-sampling

As we have seen earlier, the whole point of computing the *DNN difficulty metric* was initially to have some criteria to subsample the parameters space. The sub-sampling process has to verify two main concerns [54]. First, the sub-sampled set of *trials* (*i.e.*, experimental objects on which subjects are asked to solve the task) should be small enough so that the completion time of the evaluation with human participants remains reasonable. Second, the sub-sampled set should be large enough to remain representative of the task parameters space. Eventually, the goal is to find an optimal medium-sized set of trials that makes it *practical* to conduct the evaluation while still enabling to study the research question finely enough. This Section presents the design choices of the parameters space sub-sampling for this experiment, which fairly rely on the metric results.

As opposed to the study of Haroz and Whitney [3] (see Section 2.1), we target a diverse sample of participants for a limited sample of trials. In their experiment, their population was made of 5 subjects which answered 960 trials in each of their three experiments. Their participants response time to each trial was always below 10 seconds. According to the pilot experiments we conducted to assess the mental effort required to solve the task with our representations, we know that even 30 seconds might not be enough to solve the task on some trials. Hence, the task considered in this study is at a completely different level of difficulty. This is why we aim at having more participants that answer less trials to prevent them from becoming tired, which

would bias the results. Based on the pilot experiments feedbacks on the mental efforts required to solve our trials, we planned to give participants 30 seconds to solve each trial, and aim for a total of about 50 trials to keep the evaluation duration reasonable (*i.e.*, about half an hour).

The *type*, *#shapes* and *#colors* values distribution in the sub-sampled parameters space should remain uniform among the selected values since they are the main conditions upon which the hypotheses are built. Each value of each parameter should also occur more than once. The values of other parameters are distributed as uniformly as possible within the selected combinations of *types*, *#shapes*, and *#colors*. But following this condition still leads to too many trials. To further reduce the parameter space, some *#shapes* and *#colors* values are removed from the study. First, the value "1" is removed from both the *#shapes* and *#colors* parameters. As previously seen (Section 4.4.1), the value "1" leads to *type* distribution imbalances. As the numbers of trials would still be too large, we applied a strategy that prunes values so as to minimize the loss of pairwise significances (arcs below bars in the plots) in the complete parameters space while maximizing their conservation in the sub-sampled space. Following this strategy, we filtered out the values 3 and 6 from *#colors*, which loses 9 pairwise significant differences for Overall, 7 on Type *color* and 0 otherwise (count the arcs of pairwise significance in Figure 4 *#colors* column). For *#shapes*, we filtered out the value 4; which removes 2 pairwise significant differences for Overall, 1 on Type *color* and 0 otherwise.

The *#colors* values are thus reduced from {1, 2, 3, 4, 5, 6, 7} to {2, 4, 5, 7}, and the *#shapes* values are reduced from {1, 2, 3, 4, 5} to {2, 3, 5}. Still excluding the combination [7, 5] for (*#colors*; *#shapes*), we end up with $(|\#colors| * |\#shapes| - 1) * |type| = 11 * 4 = 44$ trials for the user evaluation. The parameter value distributions within these 44 trials are shown in Figure 5.

Finally, these choices represent one of the possible uses of the metric results, but we expect that following them when suited definitely helps keeping the reduced space representative of the complete one. Other choices of interpretation could have been done to reduce the parameters space. For example, one could have assumed that as the DNN was almost always correct on *redundant* type experimental objects, keeping the distribution of *Type* parameter values uniform was not necessary as only a few samples of *redundant* objects would suffice to validate its simplicity.

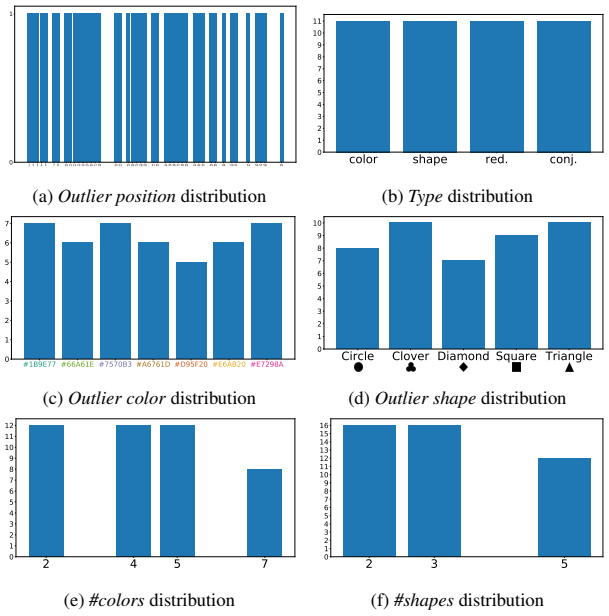
## 5. User Evaluation

This section presents the setup, choices, constraints and results of the user evaluation. The evaluation aims at studying the capacity limits of *color* and *shape* visual attributes as well as the hypotheses defined in Section 4.4.2.

### *5.1. Experimental Setup*

#### *5.1.1. The Task*

As mentioned in Section 3.1, the task consists of identifying an outlier stimulus in an $8 \times 8$ grid of distractor stimuli. A time limit for each trial is included to encourage participants




(a) *Outlier position* distribution (b) *Type* distribution

(c) *Outlier color* distribution (d) *Outlier shape* distribution

(e) *#colors* distribution (f) *#shapes* distribution

Figure 5: Parameter value distributions in the 44 selected trials of the end-user evaluation.

to solve the task as quickly as possible. Based on the pilot experiments, this time limit is set to 30 seconds and leads to a good compromise between the evaluation completion time and the error rates. When participants exceeded the time limit to answer a trial, their non-answer was registered as *Out of time*, later referred to as *OOT*.

#### *5.1.2. Dataset and Order*

The data used in this experiment are randomly extracted from the *test* set defined in Section 3.3 to fit the reduced parameter space defined in Section 4.4.3.

A random order of the trials is set, and every participant runs the trials following this order but with a random shift so that they do not all start with the same trial.

#### *5.1.3. Evaluation Protocol*

The participants are first asked to read and understand the task statements. These statements present all the colors, shapes and types that can occur during the experiment and provide a grid example. Participants are free to ask any question, and we, in return, make sure they understand the task. Then, they have to follow an 8-trials tutorial. The first 4 trials are shown already solved, along with information about their parameters. Each of them represents a different *type* value. In the next 4 tutorial trials, they are asked to solve the task without a time limit and are given feedback about the correctness of their answers. Again, the 4 trials each represent a different *type* value. Once a participant has completed the tutorial, he/she can replay it or start the evaluation.

Following the recommendations of Purchase [54], we designed an additional 8 trials for practice before starting the 44 evaluated trials. Participants are not aware that there are practice trials, and we do not consider them during the results study. This

ensures that all participants are at peak performance when the real evaluation (with monitored trials) starts. The 44 experimental trials are then displayed one after another with a three seconds break between each trial (either validated or skipped due to the time limit). During the three seconds break, the space reserved for the trial images is filled with white (the background color). For each trial, participants response times and answers are recorded. After the $26^{th}$ trial, a one minute long pause is given, with the possibility of resuming the evaluation before the pause period expires. When all trials are completed, they are asked to fill out a questionnaire about what, according to them, made the task easier (or harder) to solve. The whole protocol lasts approximately 20 to 30 minutes for each participant.

#### 5.1.4. Evaluation User Interface

The evaluation tool consists of a website specifically implemented for this study. The website is displayed in a full-screen browser on a $1920 \times 1080$ resolution monitor. Every trial image is displayed with a 1:1 ratio ($256 \times 256$ pixels) in the middle of the screen, with a black border to bring it out of the white background. The task statements and the advancement of the evaluation are succinctly written above the trial image. Below the image, the remaining time for the current trial and a validation button are displayed. To solve the task, participants have to select their answer by clicking directly on the corresponding stimulus on the image, which surrounds it with a black border. An answer can then be validated by clicking the validation button. The validation button is set wide enough so that it does not require any specific focus to be clicked on.

#### 5.1.5. Involved Participants

The participants of this experiment are 18 men and 6 women, all of whom are undergraduate students, research staff or engineers in computer science. All of them are between 21 and 50 years old with an average age of 24.8. They all reported having a perfect or corrected-to-perfect visual acuity, and none reported suffering from colorblindness.

### 5.2. Results

During the evaluation, participants response times (RTs) and answers are recorded. In the next, their performances are studied with regard to their RTs and error rates (ERs). The results are computed for 21 out of the 24 participants; after looking at the participants performances and answers to the questionnaire, we removed 2 participants for whom RTs were lower than average by more than 1.5 times the standard deviation. In addition, these two participants had ERs lower than average and were therefore considered outliers. We also removed 1 participant for which both ER and RT were higher than average plus 1.5 times the standard deviation. We also found evidence in his/her questionnaire answers that the task was either misunderstood or not seriously solved.

#### 5.2.1. Quantitative Results

In the following, we describe the participants results and interpret them in regard of the hypotheses defined in Section 4.4.2.

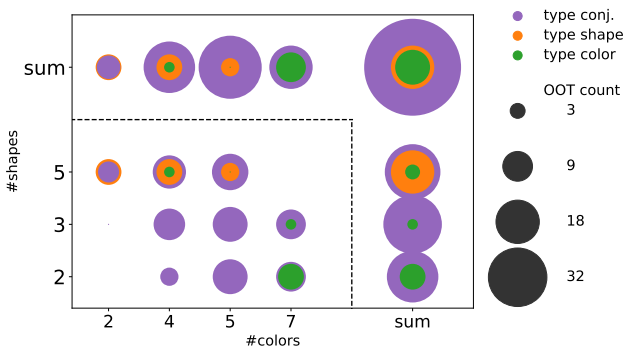
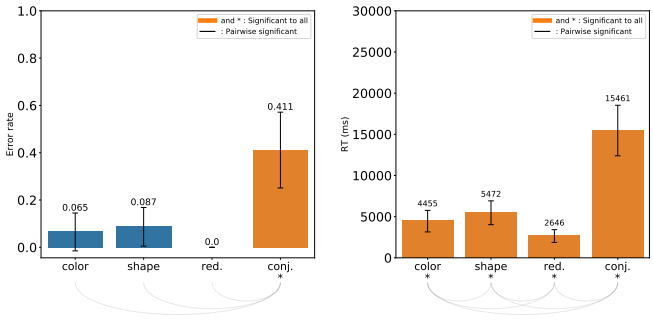



Figure 6: Number of trials for which the participants ran *out of time* (OOT) per *type*, *#colors* and *#shapes*. There are 116 OOT trials in total (5.5 per participant on average), 74% of which are of type *conjunction*, 16% are of type *shape* and 10% are of type *color*.




(a) ER per Type

(b) RT per Type

Figure 7: participants ERs and mean RTs with standard deviation bars for *type* values. ANOVA tests showed that *type* had a significant effect on the ERs and RTs. An arc between two labels means that the pairwise comparison between the corresponding performance values is significant ($p\text{-value} < 0.05$) according to a Wilcoxon rank-sum test. When performances associated to a label are significantly different to all others: the bar is colored orange and the symbol * is added below the label. *Reading example*: the *redundant* type is the condition that is fastest to solve as its RT is significantly lower than other conditions. No pairwise significance test could be run for the *redundant* type ER since no error was ever made on these trials.

Similar to the DNN results analysis, we first run Kruskal-Wallis ANOVA tests[53] on all considered parameters for the ER and RT measures. For the RTs, only validated answers are considered (*i.e.*, participants did not run *out of time*, OOT). On average, OOT trials account for 5.5 trials out of the 44 of the evaluations per participant, 74% of which are of type *conjunction*, and their distribution among *type*, *#colors* and *#shapes* is shown in Figure 6. Again, results were studied overall and per type value. The significance level was the same as those of the DNN results analysis: $\alpha = 0.05$ for the overall studies and $\alpha = 0.025$ for the *per type* studies. These tests showed that the three considered parameters presented significant effects on performance. The results are presented in Figures 7 and 8.

We **accept** $\boldsymbol{H_{type}}$ since the results plainly corroborate its statement. We can see in Figure 7 that the ER and RT performances show the same trend in terms of *type* value difficulty, although there are fewer significant pairwise differences in the ERs than in

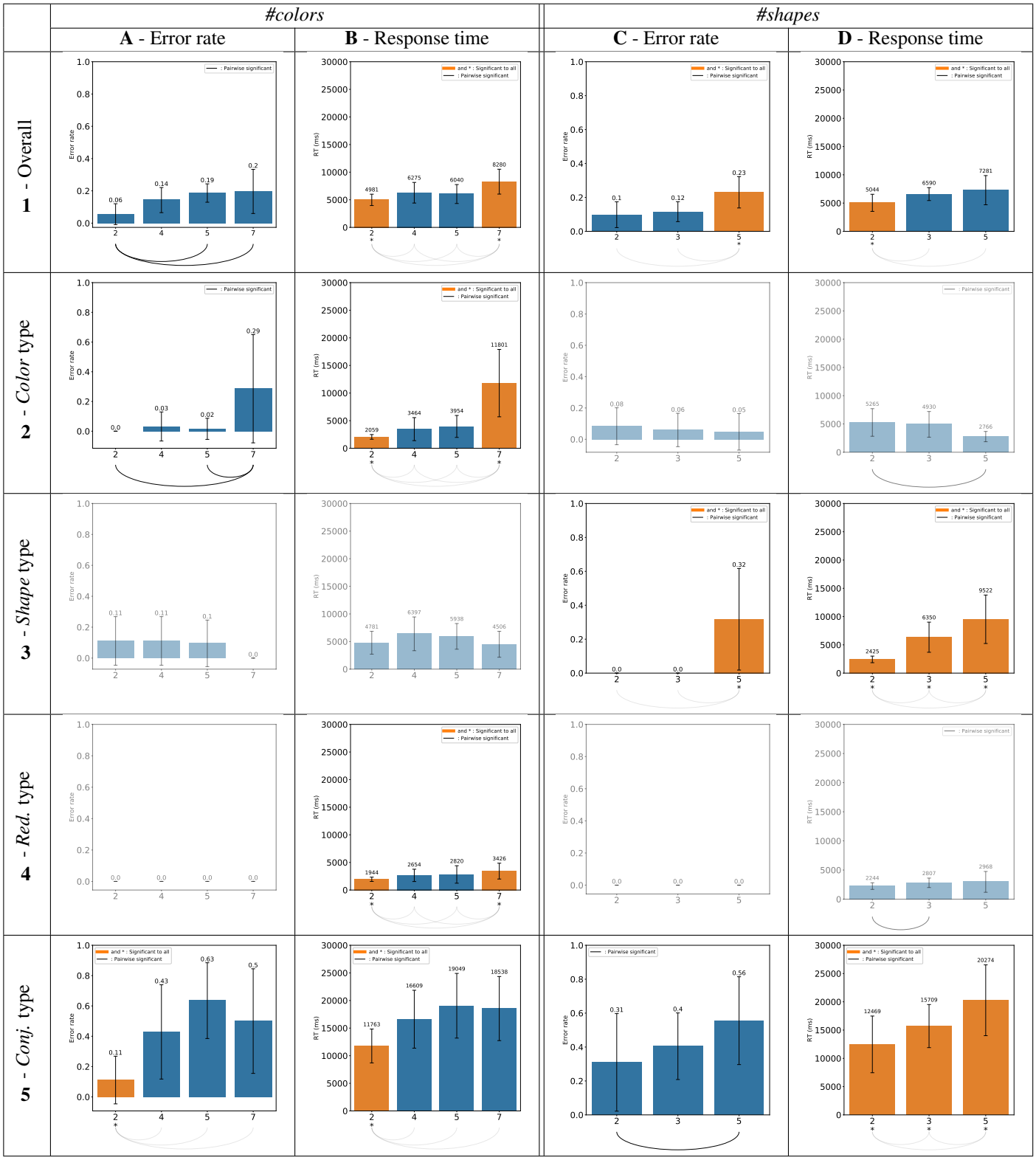



Figure 8: participants ERs and mean RTs with standard deviation bars measured during the evaluation. The first row shows the overall performances in terms of the parameter values, while the next rows present the parameter values achieved by *type*. A plot is faded if the ANOVA test on its given parameter and type aggregation failed; otherwise, it is opaque. An arc between two labels means that the pairwise comparison between the corresponding performance values is significant according to a Wilcoxon rank-sum test. When performances associated to a label are significantly different to all others: the bar is colored orange and the symbol * is added below the label. The significance threshold is $p\text{-}value < 0.05$ for the ANOVA and pairwise tests in *Overall* studies, while it is $p\text{-}value < 0.025$ in the *per type* studies. A reading example is given in the caption of Figure 7. In this table of plots, columns are given letters and rows are given numbers for ease of reference.

the RT results. The *redundant* type led to the best performances, as all participants answered correctly on all type *redundant* trials with a mean RT below 3 seconds. Although the *color* and *shape* type ERs are not significantly different (6.5% and 8.7%, respectively), type *color* (4.5s) trials were significantly faster to solve than *shape* (5.5s) trials, though such a small difference might not be relevant in an information visualization context. Finally, type *conjunction* is hardest, with an ER of 41.1% and an average RT of 15.5s, and it accounts for 74% of the OOT trials (see Figure 6).

We also **accept $H_{conj}$**, although the results are not straightforward to read. In Figure 7, the *conjunction* type is the condition that significantly led to the worst performances in terms of both the ERs and RTs. In this regard, we validate that the *conjunction* type is the hardest type value. In Figure 8 A-5 and B-5, we see that *#colors* has a significant effect on type *conjunction* trials with respect to the ERs and RTs. The effect is only significant between *exactly 2* (11% ER; 11.8 s RT) and *more than 2* (>42% ER; >16.6 s RT) *#colors*. This confirms that the task difficulty of type *conjunction* trials increases with *#colors* but caps very quickly (*i.e.*, the task difficulty no longer significantly increases with more than 2 *#colors*). In Figure 8 C-5 and D-5, no threshold effect on neither the ERs nor RTs can be directly observed as *#shapes* varies. As mentioned above, the RT results on OOT trials were not considered since it would have been incorrect to interpret them as *wrong answers in 30 seconds*. Since 86 type *conjunction* trials were OOT (see Figure 6), 86 out of $11 * 21 = 231$ answers were not considered in the computation of the type *conjunction* RT performances, and these 86 trials would have had RTs above 30 seconds. These missing points make the RT results look less "poor" than they truly are and hide the threshold effect we expected to observe for $H_{conj}$. The same interpretation can be made with the *conjunction* type and *#colors* performances, and strengthens the threshold effect that can already be observed as there are less OOT trials with 2 than with 4, 5 and 7 *#colors*.

We **accept $H_{red}$** with a restriction regarding its context. Figure 7 shows that all answers are correct for type *redundant* trials (*i.e.,* 0% ER) and that they are significantly faster to solve than any other type value. Hence, the *redundant* type is the easiest type value. All answers being correct, we validate that *#colors* and *#shapes* variations do not affect the ERs. For the RT results, *#colors* is shown to have a significant effect on type *redundant* trials (see Figure 8 B-4). The RT variations have small amplitudes, and the results remain between 1.9 and 3.4 seconds. Such RTs are more than acceptable, and their variations do not denote a significant loss of performance with respect to solving an outlier detection task in an information visualization context.

We **accept $H_{color}$** as both its assumptions are verified by the participants performances. When color is the only relevant dimension (*i.e.*, *color* type), we can see in Figure 8 A-2 that the ER remains low until 7 *#colors*, where it dramatically increases. The same behavior can be observed for the RTs (Figure 8 B-2). When color is not a relevant dimension (*i.e.*, *shape* type, Figure 8 A-3 and B-3), it has no significant effect on performance.

We **accept $H_{shape}$** as both its assumptions are verified as well. When shape is a relevant dimension (*i.e.*, *shape* type), we can see in Figure 8 C-3 and D-3 that participants performances are significantly affected by *#shapes* variations, though ER and RT do not follow the same trend. In fact, participants never failed to solve the task when shape was the only relevant dimension and there were less than 5 *#shapes*. On the other hand, the RT performances seem to increase linearly with *#shapes*. However, Figure 6 shows that participants ran out of time 18 times on trials of the *shape* type, always with 5 *#shapes*. With the same reasoning as that of $H_{conj}$, we know that the 5 *#shapes* column in Figure 8 D-3 is missing 18 (out of $(|\#colors| - 1) * 21 = 63$) data points that would have taken more than 30 seconds, and this is a significant number since it represents 28.6% of the data points in that column. When shape is not a relevant dimension (*i.e.*, *color* type, Figure 8 C-2 and D-2), it has no significant effect on the task difficulty.

#### 5.2.2. *Qualitative Results*

The qualitative results of this evaluation are built upon the participants answers to the questionnaire.

The first question was to order, from easiest to hardest, the different *type* values. For this question, almost all participants (20 out of 21) ranked the *redundant* type as the easiest and *conjunction* as the hardest. More than half of the participants (14 out of 21) ranked the *color* type as easier than *shape*. This result corroborates their performances (see Figure 7), as the *conjunction* type has the highest ER and RT; the *redundant* type is significantly faster to answer than *color*, which is also significantly faster to answer than *shape*.

The second question was to report whether an *outlier color* was easier to find than others. Ten participants clearly identified **#E7298A** as easier than others, and few participants found that **#66A61E** and **#1B9E77** made the task more difficult to solve when both were present in a trial. Only a few participants reported **#D95F02** to be an easy color, and neither **#A6761D** nor **#E6AB02** were cited as hard-to-find colors. Finally, some participants reported that the outlier color did not matter as long as the contrast between stimuli colors remained high enough. Although our experiment did not aim to measure the impact of outlier saturation on the task, it would be an interesting extension of the works by Camgöz *et al.* [55, 56] about the effects of saturation on attention.

Then, participants were asked the same question about shapes. The answers were more varied than those for the color question, and only the Square ■ and Circle ● were commonly reported to be easier than average. On the other hand, the Diamond ◆ was never reported as an easy shape, which suggests that participants found it harder to find. More than half of the participants reported that the difficulty of finding the outlier using its shape dimension was dependent on the trial distractors. Many answers were similar to the following example: "triangle among circles is pretty easy to find, whereas triangle among diamonds is hard".

We then asked the participants to provide estimations of the number of colors and shapes from which they found the task to become hard. The answers were spread out between 3 and 6 colors and between 2 and 5 shapes, meaning that the perception of difficulty truly varied from one participant to another. A majority of answers reported a *#colors* value right above that

of *#shapes* (*e.g.*, (3-2), (4-3) or even (5-4)), which allows us to think that the capacity limit of color is higher than that of shape. It is important to note participants were not told the *#colors* and *#shapes* values of the trials and did not know either that we preemptively pruned some parameter values (see Section 4.4.3). Hence, they sometimes answered with values they never saw but which represent their feelings.

The penultimate question asked the participants to report their strategy for solving the task. The main reported strategy was: first, observe if an outlier pops out of the image. If not, identify the colors and shapes in the trial; then, for each color, browse all the shapes of that color to find if there are two occurrences of the same stimulus. Finally, repeat the process until the outlier is found. This strategy is a typical behavior in visual search tasks, where the representation is first processed preattentively, then by sections (*i.e.*, texture segregation [57]), and if no "match" has still been found, the representation is eventually processed serially.

Ultimately, participants were asked what, overall, made a trial hard to solve. Two main factors came out of their answers. The first is the similarity between all the distractors in a grid. This corroborates the Duncan and Humphreys [29] theory about target-nontarget and nontarget-nontarget similarity (see Section 2). The second reason is the direct neighborhood of the outlier in a trial. This may refer to the feeling that illusory conjunctions arose when the outlier was visually *close* to its distractor neighbors.

#### *5.2.3. Capacity Limits of the Color and Shape Dimensions*

As a reminder, what we call the *capacity limit* of color (or shape) is the maximum number of different features of that dimension that can be present in a representation before a visual search task for an outlier becomes too arduous. Hence, we want to answer the following question: "how many colors can one use in a representation before it becomes too complex to find an outlier?" when univariate or bivariate data are encoded with the color and shape dimensions.

We observed with $H_{color}$ and $H_{shape}$ that *#colors* and *#shapes* do not have significant effects on performance when their respective dimensions are not relevant. Next, a dimension capacity limit will only be considered when the dimension is relevant with regard to identifying the outlier. With $H_{red}$, we saw that type *redundant* trials were not affected by *#shapes* and that *#colors* variations led to small RT fluctuations that are not significant in an information visualization context. We assume that either the maximum values of these parameters in this experiment (7 *#colors* or 5 *#shapes*) are too small to observe a loss of performance or that the visual search process does not suffer from the noise induced by distractor heterogeneity when data are redundantly encoded. We believe that the latter assumption is correct since redundant encoding has been shown to make visual search tasks significantly easier [41].

**Color:** When color is the only relevant dimension for identifying the outlier (*color* type), we observe (see Figure 8 A-2 and B-2) that all *#colors* values led to ERs of less than 3% and RTs of less than 4 seconds on average, except for the value 7, which reached a 29% ER in over 11 seconds. We believe this kind of shift in performance is the consequence of exceeding the capacity limit of the color dimension in this experiment. Hence, when color is the only relevant dimension for representing data, its capacity limit is **strictly less than 7**. In conjunction search (type *conjunction*), the capacity limit is **strictly less than 4** since we can see that the performances do not worsen as *#colors* increases from a value of 4.

**Shape:** When shape is the only relevant dimension (*shape* type), the participants did not make any errors when there were less than 5 shapes. At 5 shapes, the ER reached 32%, which is close to the ER obtained for 7 *#colors* when color was a relevant dimension. As observed in Section 5.2.1, the RT measures do show significant variations between every pair of values, though these differences remain linear and no threshold effect can directly be observed. However, we also observed in Figure 6 that there were several *OOT* trials when there were 5 shapes in the type *shape* trials, and these represent 28.6% of the total trials of this variety. We conclude that the shape capacity limit is **strictly less than 5** when it is a relevant dimension. For conjunction search (type *conjunction*), no threshold effect can be observed as *#shapes* varies. As the number of OOT trials on type *conjunction* trials is even among the *#shapes* values (see Figure 6), there is no "hidden" threshold effect due to missing data points either. Finally, no information about the capacity limit of shape can be determined from the results for the type *conjunction* trials. Either 2 *#shapes* is already over the capacity limit, 5 *#shapes* is still under the limit, or the search task for an outlier in this case is actually linearly related to *#shapes*.

Therefore, the capacity limit of color is found to be higher than that of shape. This assumption reflects what participants reported in the questionnaire (see Section 5.2.2). Moreover, we can see in Figure 8 that 5 *#colors* in type *color* (*1.5% ER; 4s RT*) trials led to significantly better performances than those obtained with 5 *#shapes* in type *shape* trials (*31% ER; 9.5s RT*). Such a difference confirms that the shape dimension is more sensitive to heterogeneity than color when each is the relevant dimension for identifying an outlier.

### *5.3. Results Sensitivity*

Similar to Demiralp *et al.* [32], we tested the robustness of our results to participant removal. We expect that more robust results in this sense have a greater ability to be generalized. To this end, we extracted random subsets of participants of different sizes (from 10% to 95% of the full dataset) and computed the ER and RT performances from these subsets with each parameter. That is, for each parameter, we computed its mean ER and RT values on the subsets of participants. Then, we used the Spearman correlation coefficient [58] to quantify the correlation between the performances of a subset of participants with the performances of all the participants. To reduce sampling-related uncertainties, each subset size was sampled 10 times so that the correlation coefficients reported in Figure 9 were averaged over 10 computations. Spearman correlation coefficients range between -1 and 1, where 1 is a positive correlation, -1 is a negative correlation and 0 means no correlation. Figure 9 shows that the correlation coefficients are over 0.5 with 25% of the participant

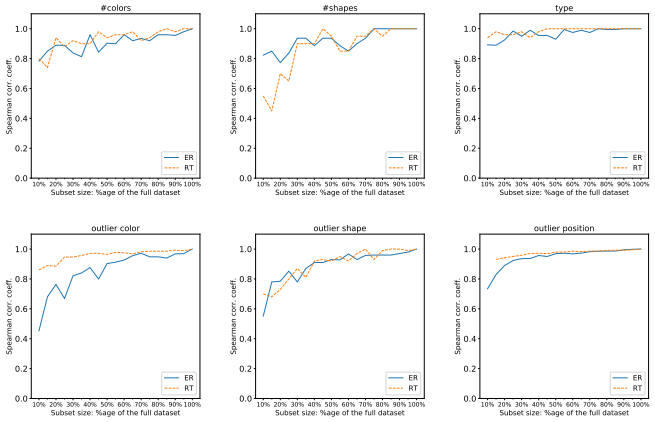



Figure 9: Spearman correlation coefficients [58] between random subsets of participants (of different sizes) and all participant performances for each parameter. The X-axis corresponds to the size (percentage of the full dataset, from 10% to 95% with a 5% step) of the subset of participants for which their performances are compared to those of all participants (100%). Coefficients are averaged over 10 samplings for each subset size to minimize sampling-related uncertainties.

data, indicating that some information is preserved. With 40% to 50% of the data, the correlation coefficients are all between 0.8 and 1, meaning that most of the information is preserved with only 50% of the participants. Such high correlation coefficients with only half the participants show that the results of these experiments are not very sensitive to participant removal. Hence, they are not biased by some participant-specific behaviors.

## 6. Discussion

This evaluation enabled us to study some hypotheses as well as the capacity limits of color and shape. Nevertheless, its results can only be generalized in view of the evaluation design, alongside its limitations. This section presents these limitations as well as some interesting other outcomes of the experiment.

### 6.1. Limitations

As for any experiment, its results have to be considered within the limitations induced by its protocol.

First, our participants are taken from a noticeably educated population in computer science. Hence, they are used to computer devices and are familiar with visualizations (as users or experts). The results of this experiment must be used with caution if one is to generalize them to a more generic population. However, most of the population is now used to computers and visualizations, and we expect that computer science background of the participants have helped them understand faster what the concept of the task was, but did not made them better at actually solving the task.

In the experimental design, we selected the color set from a color palette provider (see Section 3.3) so that all colors could be as distinguishable as possible. Since the experiment aimed to study the capacity limit of color (and not to find which color set is best for representing categorical data), we mitigated the effect of the color set on the performances. However, some participants found that the differences in saturation between some colors had impacts on their performances. Interestingly the participants did not mention any similar observation concerning the shapes used in the experiment. Again, the experiment results are coupled with the selected colors and shapes set. For instance, the results may have been different using another color palette. Recent tools such as Colorgorical [43] are more dedicated to alleviate the impact of a specific color palette, but we used ColorBrewer in this experiment as it is well-established in the community. Anyway, results should be interpreted in view of the scope defined by the experiment. Nevertheless, the selected shapes are remain very common and we think the results about color are generalizable to other categorical palettes which are widely used to represent data classes.

Another limitation is that the observed *#colors* performance variations are limited by an aspect of the experimental design. When generating an image with 5 colors, these are not taken from a designed set of 5 categorical colors but are picked from the 7 color set defined in Section 3.2. It is then important to note that the performances observed on trials with less than 7 colors were not measured from trials with an optimal color set fitted to their *#colors*. The same limitation could be stated for shapes, although the notion of distance between shapes is more complex. Nevertheless, it was necessary to define a static set of colors and shapes so that we could keep track of the *outlier color* distribution and be able to aggregate them to study our results. Having a different and optimal set for each *#colors* value would also have made the participants able to infer properties of the display from the color set of a trial, and this could bias the ways in which they built their strategies.

Finally, as stated in Section 3.2, we balanced our shape sub-features (*i.e.*, straight horizontal and vertical lines, curves, tilted lines) to try not to favor any particular kind. However, such a definition leads to imbalances between shape areas. The Triangle ▲ has a smaller area than the Square ■, so the colors they are filled with are not represented by the same number of pixels. Thus, when color is a relevant dimension for identifying the outlier, it may be harder to solve the task if the outlier shape is a Triangle ▲ than if it is a Square ■. Although our study does not enable us to validate this assumption, it would be interesting to experiment whether, when considering a large number of colored shapes, the shape areas are (or become) more important than the shapes themselves.

### 6.2. Feedback on DNN as a metric

The motivation to design novel sub-sampling approaches emerged from the need to improve the reproducibility of evaluations parameters space sub-sampling since the newcoming representations aim at visualizing data of increasing complexity. Being able to evaluate these complex representation techniques become a challenge as the consideration of all the combinations of their parameters leads to a combinatorial explosion of the number of experimental objects to test. Having objective ways to sub-sample such broad parameters space while preserving representativity of the complete space is clearly a concern.

In this experiment, we interpreted the performances of a DNN as those of a *meta-user* to assess the task difficulty, al-

though we have no reason to think that the DNN strategy to solve the task would corroborate the human perception system. The intuition is that making a representation more complex will affect both the DNN and humans strategies to solve the task, even though their strategies would be different. With knowledge from the literature, we can assess when the DNN is (or not) a good model of human perception. This is typically what we have done in Section 4.4.2, when we sometimes used the DNN results to refine hypotheses, and sometimes discarded them since we had several indicators that they were not a good model of human perception. The *a posteriori* study of correlations [27] between the DNN and humans showed that our assumption was correct. Such a study of correlations is not necessary nor common to evaluate the sampling technique used in an experiment. Yet, since the approach we propose is novel, we conducted it to emphasize the strong correlations between the DNN and human participants and anticipate necessary and constructive criticisms inherent to new methods, as there was no objective reason to think they would be correlated (other than belief).

### *6.3. Common Assumption about Color Capacity Limit*

It is commonly assumed that the maximum number of colors in a representation that humans can efficiently handle is $10 \pm 2$. This common assumption seems to be used by many color palette providers, as many palettes are built for up to 12 classes. For example, the ColorBrewer tool [44] recommends using between 5 and 7 classes for choropleth maps, while isoline maps can safely use more; the online tool provides color palettes with up to 12 classes. To the best of our knowledge, no study has verified this assumption. As our experiment showed the capacity limit of color to be strictly less than 7 for a feature search, it shows that the commonly assumed limit of $10 \pm 2$ colors is overoptimistic. It could be even worse in some contexts as more and more modern representation techniques depict more than 64 stimuli and do not organize them into a regular grid, which further increase the complexity of representations. Color remains an efficient visual attribute for encoding data, but its limit is lower than assumed in many representation tools. Some strategies, such as color grouping [3], can be used to increase this limit but are not suited for all representation designs.

## 7. Conclusion

This paper has presented a study of the capacity limits of attention on representations with varying heterogeneity where data are encoded with color and/or shape visual attributes.

We proposed an approach that leverages Deep Learning techniques to drive the sub-sampling of the experiment parameters space by interpreting its results *as a metric*. As many metric designed to assess human perception system, we have seen that the *DNN metric* also has the common pitfalls that should be carefully addressed when using its results to assess human behaviors. Nevertheless, this automated method is reproducible and enables to evaluate any task–representation that can be expressed programmatically. This experiment and the study of correlation between the DNN and human performances [27], showed that a DNN based approach is a promising means to refine user evaluation designs when parameters variations lead to a combinatorial explosion in the number of configurations.

Then, the study of capacity limits was conducted on the reduced parameters space through a user evaluation on 21 participants. The task consisted in identifying an outlier defined by its color and/or shape in a regular grid of randomly laid out distractors stimuli. The results of the experiment confirm that the task difficulty is highly dependant on the way the outlier is encoded (*i.e.*, easiest with redundant encoding, hardest with conjunction encoding). Results also show that color is more efficient than shape to encode data in simple feature search and confirm that variations in irrelevant dimensions have no effect on the task difficulty. Finally, we have seen that the capacity limits of color and shape are not as high as we could expect, especially for the color dimension for which we found the limit to be significantly lower than what is assumed in many tools.

Many future work ideas have emerged throughout this study. Regarding the experiment itself, a first line of future work could be to search for a set of outlier stimuli instead of a single outlier stimulus. That is, find an outlier cluster in a grid of random stimuli (or random clusters). This problem is also common in information visualization and involves other visual search strategies, such as texture segregation [30, 57]. Moreover, researches about other commonly used visual attributes (*e.g.*, position, size) could be conducted. One could study the different effects of mixing these dimensions to represent data and measure which dimensions are least harmful to representation readability when joined together. Although conjunction search has been shown to make visual search tasks much harder, visualization designers cannot always afford not to mix their visual attributes when representing tens of data classes. However, our experiment showed that mixing colors and shapes quickly made representations arduous to read. Hence, finding conditions that optimize conjunction search in complex displays would be valuable.

Regarding the Deep Learning based difficulty metric, it would be interesting to see how well we could trust the model performances to assess humans performances on unseen conditions (*e.g.*, new colors or new shapes). One could also study how bio-inspired models behave compared to standard CNNs and whether they are closer to human behaviors on visual search tasks. To improve correlations between the Deep Learning model and human behavior, we could also try to inject human saliency information in the model training [59]. Finally, considering an ensemble of networks instead of relying on a single one could help minimizing the bias induced by specific DNN architectures.

## References


[1] C. Ware, Information Visualization: Perception for Design, 3rd Edition, Morgan Kaufmann, 2012.

[2] C. Healey, J. Enns, Attention and Visual Memory in Visualization and Computer Graphics, IEEE Transactions on Visualization and Computer Graphics 18 (7) (2012) 1170–1188.

[3] S. Haroz, D. Whitney, How capacity limits of attention influence information visualization effectiveness, IEEE Transactions on Visualization and Computer Graphics 18 (12) (2012) 2402–2410.

[4] C. Gramazio, K. Schloss, D. Laidlaw, The relation between visualization size, grouping, and user performance, IEEE transactions on visualization and computer graphics 20 (12) (2014) 1953–1962.

[5] D. Huber, C. Healey, Visualizing data with motion, in: VIS 05. IEEE Visualization, 2005., IEEE, 2005, pp. 527–534.

[6] T. Itoh, Y. Yamaguchi, Y. Ikehata, Y. Kajinaga, Hierarchical data visualization using a fast rectangle-packing algorithm, IEEE Transactions on Visualization and Computer Graphics 10 (3) (2004) 302–313.

[7] J. Mackinlay, Automating the design of graphical presentations of relational information, ACM Trans. Graph. 5 (2) (1986) 110—141.

[8] C. Ware, J. Beatty, Using color dimensions to display data dimensions, Human factors 30 (2) (1988) 127–142.

[9] C. Healey, Choosing effective colours for data visualization, in: Proceedings of Seventh Annual IEEE Visualization'96, IEEE, 1996, pp. 263–270.

[10] H. Chernoff, The use of faces to represent points in k-dimensional space graphically, Journal of the American statistical Association 68 (342) (1973) 361–368.

[11] F. Post, T. van Walsum, F. Post, D. Silver, Iconic techniques for feature visualization, in: Proceedings Visualization'95, IEEE, 1995, pp. 288–295.

[12] M. Gleicher, M. Correll, C. Nothelfer, S. Franconeri, Perception of average value in multiclass scatterplots, IEEE transactions on visualization and computer graphics 19 (12) (2013) 2316–2325.

[13] J. Bertin, Semiology of Graphics: Diagrams, Networks, Maps, University of Wisconsin Press, 1983.

[14] D. Altunbay, C. Cigir, C. Sokmensuer, C. Gunduz-Demir, Color graphs for automated cancer diagnosis and grading, IEEE Transactions on Biomedical Engineering 57 (3) (2009) 665–674.

[15] H. Zhou, X. Yuan, H. Qu, W. Cui, B. Chen, Visual clustering in parallel coordinates, in: Computer Graphics Forum, Vol. 27, Wiley Online Library, 2008, pp. 1047–1054.

[16] Wolfe, Jeremy M and Horowitz, Todd S, Five factors that guide attention in visual search, Nature Human Behaviour 1 (3) (2017) 1–8.

[17] Miller, George A, The magical number seven, plus or minus two: Some limits on our capacity for processing information., Psychological review 63 (2) (1956) 81.

[18] A. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive Psychology 12 (1) (1980) 97–136.

[19] A. Treisman, Focused attention in the perception and retrieval of multidimensional stimuli, Perception & Psychophysics 22 (1) (1977) 1–11.

[20] Wolfe, Jeremy M, Forty years after feature integration theory: An introduction to the special issue in honor of the contributions of Anne Treisman, Attention, Perception, & Psychophysics 82 (1) (2020) 1–6.

[21] Wolfe, Jeremy M, Major issues in the study of visual search: Part 2 of "40 Years of Feature Integration: Special Issue in Memory of Anne Treisman", Attention, Perception, & Psychophysics 82 (2) (2020) 383–393.

[22] Purchase, Helen C and Cohen, Robert F and James, Murray, Validating graph drawing aesthetics, in: International Symposium on Graph Drawing, Springer, 1995, pp. 435–446.

[23] Purchase, Helen, Which aesthetic has the greatest effect on human understanding?, in: International Symposium on Graph Drawing, Springer, 1997, pp. 248–261.

[24] Horikawa, Tomoyasu and Aoki, Shuntaro C and Tsukamoto, Mitsuaki and Kamitani, Yukiyasu, Characterization of deep neural network features by decodability from human brain activity, Scientific data 6 (1) (2019) 1–12.

[25] Jacob, Georgin and Pramod, RT and Katti, Harish and Arun, SP, Qualitative similarities and differences in visual object representations between brains and deep networks, Nature communications 12 (1) (2021) 1–14.

[26] Kheradpisheh, Saeed R and Ghodrati, Masoud and Ganjtabesh, Mohammad and Masquelier, Timothée, Humans and deep networks largely agree on which kinds of variation make object recognition harder, Frontiers in computational neuroscience 10 (2016) 92.

[27] Giovannangeli, Loann and Giot, Romain and Auber, David and Benois-Pineau, Jenny and Bourqui, Romain, Analysis of Deep Neural Networks Correlations with Human Subjects on a Perception Task, in: 2021 25th International Conference Information Visualisation (IV), 2021, pp. 129–136. doi:{10.1109/IV53921.2021.00029}.

[28] Wolfe, Jeremy M and Gray, W, Guided Search 4.0, Integrated models of cognitive systems (2007) 99–119.

[29] J. Duncan, G. W. Humphreys, Visual search and stimulus similarity, Psychological Review 96 (3) (1989) 433–458.

[30] H. Pashler, Cross-dimensional interaction and texture segregation, Perception & Psychophysics 43 (4) (1988) 307–318.

[31] P. T. Quinlan, G. W. Humphreys, Visual search for targets defined by combinations of color, shape, and size: An examination of the task constraints on feature and conjunction searches, Perception & Psychophysics 41 (5) (1987) 455–472.

[32] Demiralp, Çağatay and Bernstein, Michael S and Heer, Jeffrey, Learning perceptual kernels for visualization design, IEEE transactions on visualization and computer graphics 20 (12) (2014) 1933–1942.

[33] Tatler, Benjamin W, The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions, Journal of vision 7 (14) (2007) 4–4.

[34] Behrisch, Michael and Blumenschein, Michael and Kim, Nam Wook and Shao, Lin and El-Assady, Mennatallah and Fuchs, Johannes and Seebacher, Daniel and Diehl, Alexandra and Brandes, Ulrik and Pfister, Hanspeter and others, Quality metrics for information visualization, in: Computer Graphics Forum, Vol. 37, Wiley Online Library, 2018, pp. 625–662.

[35] Haehn, Daniel and Tompkin, James and Pfister, Hanspeter, Evaluating 'graphical perception'with CNNs, IEEE transactions on visualization and computer graphics 25 (1) (2018) 641–650.

[36] Cleveland, William S and McGill, Robert, Graphical perception: Theory, experimentation, and application to the development of graphical methods, Journal of the American statistical association 79 (387) (1984) 531–554.

[37] Haleem, Hammad and Wang, Yong and Puri, Abishek and Wadhwa, Sahil and Qu, Huamin, Evaluating the readability of force directed graph layouts: A deep learning approach, IEEE computer graphics and applications 39 (4) (2019) 40–53.

[38] Giovannangeli, Loann and Bourqui, Romain and Giot, Romain and Auber, David, Toward automatic comparison of visualization techniques: Application to graph visualization, Visual Informatics (2020).

[39] Ghoniem, Mohammad and Fekete, Jean-Daniel and Castagliola, Philippe, On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis, Information Visualization 4 (2) (2005) 114–135.

[40] Okoe, Mershack and Jianu, Radu and Kobourov, Stephen G, Node-link or Adjacency Matrices: Old Question, New Insights, IEEE Transactions on Visualization and Computer Graphics (2018).

[41] Nothelfer, Christine and Gleicher, Michael and Franconeri, Steven, Redundant encoding strengthens segmentation and grouping in visual displays of data., Journal of Experimental Psychology: Human Perception and Performance 43 (9) (2017) 1667.

[42] B. Bauer, P. Jolicoeur, W. Cowan, Visual search for colour targets that are or are not linearly separable from distractors, Vision research 36 (10) (1996) 1439–1466.

[43] Gramazio, Connor C and Laidlaw, David H and Schloss, Karen B, Colorgorical: Creating discriminable and preferable color palettes for information visualization, IEEE transactions on visualization and computer graphics 23 (1) (2016) 521–530.

[44] Harrower, Mark and Brewer, Cynthia A, ColorBrewer. org: an online tool for selecting colour schemes for maps, The Cartographic Journal 40 (1) (2003) 27–37.

[45] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Statistics surveys 4 (2010) 40–79.

[46] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[47] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[48] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[50] S. He, H. R. Tavakoli, A. Borji, Y. Mi, N. Pugeault, Understanding and visualizing deep visual saliency models, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10198–10207.

[51] F. Chollet, et al., Keras, https://github.com/fchollet/keras (2015).

[52] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, H. Nielsen, Assessing the

accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (5) (2000) 412–424.
[53] W. Kruskal, W. Wallis, Use of ranks in one-criterion variance analysis, Journal of the American statistical Association 47 (260) (1952) 583–621.
[54] H. Purchase, Experimental Human-Computer Interaction: A Practical Guide with Visual Examples, Cambridge University Press, 2012.
[55] N. Camgöz, C. Yener, D. Güvenç, Effects of hue, saturation, and brightness on preference, Color Research and Application 27 (3) (2002) 199–207.
[56] Camgöz, Nilgün and Yener, Cengiz and Güvenç, Dilek, Effects of hue, saturation, and brightness: Part 2: Attention, Color Research & Application 29 (1) (2004) 20–28.
[57] T. Callaghan, M. Lasaga, W. Garner, Visual texture segregation based on orientation and hue, Perception & Psychophysics 39 (1) (1986) 32–38.
[58] Zwillinger, Daniel and Kokoska, Stephen, CRC standard probability and statistics tables and formulae, Crc Press, 1999.
[59] de San Roman, Philippe Pérez and Benois-Pineau, Jenny and Domenger, Jean-Philippe and Paclet, Florent and Cataert, Daniel and De Rugy, Aymar, Saliency Driven Object recognition in egocentric videos with deep CNN: toward application in assistance to Neuroprostheses, Computer Vision and Image Understanding 164 (2017) 82–91.