Seeing Eye to AI? Applying Deep-Feature-Based Similarity Metrics to Information Visualization

Sheng Long
shenglong@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Angelos Chatzimparmpas
a.chatzimparmpas@uu.nl
Utrecht University
Utrecht, Netherlands

Emma Alexander
ealexander@northwestern.edu
Northwestern University
Evanston, Illinois, USA

Matthew Kay
mjskay@northwestern.edu
Northwestern University
Evanston, Illinois, USA

Jessica Hullman
jhullman@northwestern.edu
Northwestern University
Evanston, Illinois, USA

Abstract

Judging the similarity of visualizations is crucial to various applications, such as visualization-based search and visualization recommendation systems. Recent studies show deep-feature-based similarity metrics correlate well with perceptual judgments of image similarity and serve as effective loss functions for tasks like image super-resolution and style transfer. We explore the application of such metrics to judgments of visualization similarity. We extend a similarity metric using five ML architectures and three pre-trained weight sets. We replicate results from previous crowdsourced studies on scatterplot and visual channel similarity perception. Notably, our metric using pre-trained ImageNet weights outperformed gradient-descent tuned MS-SSIM, a multi-scale similarity metric based on luminance, contrast, and structure. Our work contributes to understanding how deep-feature-based metrics can enhance similarity assessments in visualization, potentially improving visual analysis tools and techniques. Supplementary materials are available at https://osf.io/dj2ms/.

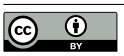
CCS Concepts

• **Human-centered computing** → **Heuristic evaluations**; **Visualization design and evaluation methods**; • **Computing methodologies** → **Computer vision tasks**; **Visual inspection**.

Keywords

evaluation, similarity perception, replication studies, deep-feature-based similarity metrics

ACM Reference Format:
Sheng Long, Angelos Chatzimparmpas, Emma Alexander, Matthew Kay, and Jessica Hullman. 2025. Seeing Eye to AI? Applying Deep-Feature-Based Similarity Metrics to Information Visualization. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3706598.3713955



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713955

1 Introduction

Similarity is a fundamental construct in human cognition, underlying numerous mental processes, including categorization, reasoning, and decision-making [52]. Human-perceived similarity has been studied extensively in psychology and cognitive science [45, 62] and has been leveraged in applications such as image retrieval [28] and human-in-the-loop categorization [63, 75] in recent years [61]. Recent studies in computer vision have shown that deep-feature-based similarity metrics correlate well with perceptual judgments of image similarity [87] and serve as effective loss functions for tasks like image super-resolution [39] and style transfer [21]. The success of these applications in computer vision raises an intriguing question: *Can similar approaches be effectively applied to the domain of information visualization?*

Judging visualization similarity is crucial for various applications, ranging from visual search [83] and visualization recommendation [86] to automated design and sequencing [6, 13]. Knowing whether a visualization is *robustly discriminable* across a range of datasets, i.e., whether it consistently maintains clear visual distinctions across a range of datasets, is crucial for effective communication and design. The idea that perceived visual structure should correspond with structure in the underlying data is recurring in multiple works, such as in the principle of visual-data correspondence by Kindlmann and Scheidegger [44], in the principle of congruence by Tversky et al. [71], in similar principles proposed in multi-view visualization [59, 60], and in the visual embedding model by Demiralp et al. [14]. However, such knowledge is challenging to obtain at design time and costly to evaluate empirically [74].

This paper investigates the use of deep-feature-based similarity metrics to approximate similarity perception in information visualization. We offer two primary research contributions.

(1) While deep neural networks are widely used in visualization for various tasks [76, 84], our work is the first to implement a domain-independent transfer learning technique from computer vision [49, 87] to an information visualization setting. We extended prior work on deep-feature-based similarity metrics using weights trained on *Stylized ImageNet* [23], a modified ImageNet-1K dataset where images are artistically stylized while preserving their original content and labels. Given Stylized ImageNets' increased object detection performance, we hypothesized it would benefit information visualization tasks. Our results suggest that transfer learning

from computer vision can reduce the time and resources needed to develop effective models when applied with caution to information visualization, potentially opening up new avenues for cross-domain research and applications. However, we did not observe any significant performance differences between models trained on Stylized ImageNet versus ImageNet-1K.

(2) We assess the benefits and limitations of applying deep-feature-based similarity metrics in information visualization by conceptually replicating and comparing our results to two crowd-sourced studies that collected similarity judgments [13, 74]. These studies are well-suited for exploring visualization similarity as they offer diverse datasets and methodologies, allowing us to systematically examine how deep-feature-based similarity measures perform across different visualization contexts and analytical approaches.
   (a) Our first replication (Section 5) reveals that when using certain deep learning (DL) networks, *which have not been trained on scatterplots*, deep-feature-based similarity metrics achieve better clustering alignment with human judgments of scatterplot similarity than traditional computer vision metrics whose parameters are optimized on the set of scatterplots.
   (b) Our second replication (Section 6) reveals that for visual channels like color and shape, deep-feature-based similarity metrics struggle to capture what humans perceive to be similar. However, these metrics perform well when assessing the visual channel of size.

We speculate that the strong performance of deep-feature-based similarity metrics in capturing scatterplot similarity (Section 5) is due to their ability to capture patterns in spatial distributions of visual elements. This suggests potential applications to other visualization types that also encode data through spatial arrangements. These metrics struggle to capture judgments of color and shape similarity (Section 6), which may be due to humans making these judgments based on factors beyond spatial features, such as cultural association. The metrics' better performance with size judgments likely stems from its more direct relationship to spatial distribution. These results highlight both opportunities and limitations in applying deep-feature-based similarity metrics trained on natural scenes to information visualization.

2 Background

2.1 Perceptual Similarity Metrics

Computer vision has long studied the problem of measuring the quality of an image for downstream tasks such as lossless compression. Due to the costly nature of running experiments to collect human data, significant research effort has been devoted to deriving an *objective* metric that can *approximate human perception*. For example, if two images appear *similar* to the human eye, then their "distance", measured by this metric, should be close to 0.

Traditional metrics, such as Manhattan $\mathcal{L}_1$, Euclidean $\mathcal{L}_2$, Mean Squared Error (MSE), and Peak-Signal-to-Noise-Ratio (PSNR), rely on *point-wise* pixel differences, which offers computational ease but poorly matches perceived visual quality [18, 24, 77]. Subsequently developed patch-based metrics rely on the hypothesis that the human visual system is highly adapted for extracting *structural* information. One such metric is Similarity Index Measure (SSIM) [77], which compares "local patterns of pixel intensities that have been normalized for luminance and contrast". One popular extension of SSIM is *Multi-Scale SSIM* (MS-SSIM) [78], which extends SSIM by evaluating image similarity at *multiple scales*, capturing both local and global structural information. Nevertheless, SSIM and MS-SSIM often do not capture the nuances of human vision when more structural ambiguity is present [64] and fall short for more complex downstream tasks, such as image generation and image synthesis.

More recently, researchers have leveraged deep neural networks (DNN) for tasks beyond classification to develop metrics such as LPIPS [87], PieAPP [57] and DISTS [16] that utilize *deep features*, i.e., features obtained through deep-learning architectures trained on the ImageNet dataset [15]. One of the primary motivations for utilizing these deep-feature-based metrics is that they can transform pixel representations to a space that is more *perceptually uniform* [16]. They have also demonstrated excellent empirical performance for downstream tasks such as image super-resolution [39] and style transfer [21]. However, their application to the perceptual similarity of information visualization remains unexplored, perhaps due to the uncertainty about whether features learned from natural scenes can effectively transfer to the often abstract and stylistic domain of information visualization — a gap this work addresses.

2.1.1 Collecting Human Similarity Judgments. Regardless of the underlying mechanics, the usefulness of similarity metrics is ultimately determined by how well they align with human judgment. Traditional Image Quality Assessment (IQA) databases, such as LIVE [66], CSIQ [50], and TID2013 [56], collect *Mean Opinion Scores* by asking participants to provide subjective ratings for pairwise images on a specified numerical scale (e.g., Likert scales). More recent IQA databases, such as BAPPS [87], use two-alternative forced choice (2AFC) experiments to collect similarity judgment on a dataset of images applied with low-level distortions. Beyond low-level perceptual similarity, the THINGS database collected behavioral odd-one-out similarity judgments [31, 32] on everyday objects to generate interpretable object dimensions predictive of behavior and similarity, and the NIGHTS dataset [20] collected 2AFC judgment on diffusion-synthesized images perturbed along various dimensions.

Aside from the computer vision/machine learning databases mentioned above, information visualization studies have also collected data on similarity judgments (albeit on a much smaller scale). Demiralp et al. [13] collected similarity judgment via five different tasks on different visual channels, such as color, size, and shape. Pandey et al. [53] asked participants to group scatterplot thumbnails and compared their results to scatterplot diagnostics (scagnostics). Ma et al. [51] collected similarity judgment on scatterplots and trained a deep neural network, `ScatterNet`, to learn features that capture the human perception of scatterplot similarity. Beyond scatterplots, Gogolou et al. [25] collected similarity perception of time-series plots by asking participants to select among four options which time-series chart appears most similar to the reference chart. Others have collected judgments of visualization similarity in the form of transition costs when moving from one visualization to the next

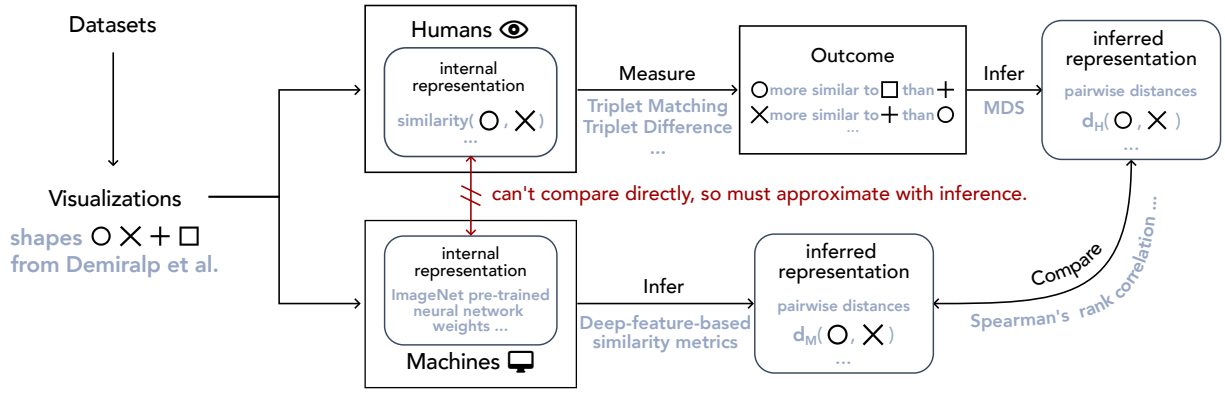


Figure 1: A general framework for comparing two information-processing systems (e.g., humans vs. machines), inspired by the framework proposed by Sucholutsky et al. [68]. Framework uses Demiralp et al.'s experiment on the perception of shape similarity as one concrete example [13].

in a sequence [35, 43]. In this paper, we conduct conceptual replications of the experiments by Demiralp et al. [13] and Veras and Collins [74] (that uses Pandey et al.'s data [53]).

2.2 Comparing Human and Machine Behavior

Given the remarkable performance of DNNs at image classification, research across different fields — computer vision, cognitive science, neural science, to name a few — has sought to answer **whether DNNs are good models of the human visual system**. To answer this question requires *comparing* DNNs to the human visual system. Existing research has done this both at the *mechanistic* level (i.e., comparing internal representations) and *functional/behavioral* level (i.e., comparing outcomes). See Sucholutsky et al. [68] for an overview of comparing and aligning representations across different information processing systems.

2.2.1 Functional Similarity. At the **outcome** level (Fig. 1), much work has focused on comparing and modeling classification performance [61]. Researchers have found that small perturbations that are imperceptible to humans can dramatically affect model classification decisions [26], and that texture and local image features drive classifiers [2, 23], whereas humans are more strongly influenced by Gestalt shape [1]. Beyond comparing classification accuracies, other work [22] has looked at additional outcome measures such as trial-by-trial error consistency.

In information visualization, Haehn et al. [29] trained four different convolutional neural network (CNN) architectures on five different visualization tasks to reproduce Cleveland and McGill's graphical perception experiments [10]. They found that when compared to human performance baselines, CNNs are not "currently a good model for human graphical perception" — while CNNs performed better than humans in elementary perceptual tasks (e.g., estimating quantities from visual marks), humans significantly out-performed CNNs in visual relation tasks (e.g., comparing bar lengths). Yang et al. [85] trained DNNs on *human correlation judgments* in scatterplots across three studies and found that a subset of their trained DL architectures (e.g., VGG-19) has comparable accuracy to the best-performing regression analyses in prior research. The main difference between our work and Haehn et al. [29] is that we are not training network parameters on any domain-specific dataset but instead rely on pre-trained ImageNet and Stylized ImageNet weights. The main difference between our work and Yang et al. [85] is that we utilize DL architectures that are trained on natural images for classification and not on human judgment. Our metrics are therefore *information-visualization-domain-independent* and can be applied directly to any pair of visual stimuli,[1] whereas the trained weights and architectures of Haehn et al. [29] and Yang et al. [29] may face challenges in generalizing to contexts beyond their specific training domains.

2.2.2 Representational Similarity. At the **internal representation** level (Fig. 1), numerous techniques exist, and the most popular one is *Representational Similarity Analysis* (RSA), a multivariate technique introduced by Kriegeskorte et al. [45]. RSA compares different representations of the same set of stimuli, such as neural activity patterns, computational model outputs, and behavioral data, by computing similarity matrices for each representation and then comparing these matrices to assess how well different representations align with each other. Using RSA, Khaligh-Razavi and Kriegeskorte [41] compared the representational geometry of object recognition in the human brain with that of various computer vision models, and they found that a deep convolutional network performed best in both categorization and explaining inferior temporal cortex representations. For more on calculating similarity in psychological space, see Roads and Love [62] for an overview.

2.2.3 Texture vs. Shape. Most DNN models that perform well on Brain-Score [65] and other prediction metrics *do not* rely on global shape when classifying objects, contrary to the fundamental conclusion from psychological research that humans largely rely on shape when identifying objects [3]. DNNs (such as the CORnet-S model [48], one of the best models of human vision) largely classify objects based on *texture* and *local shape features* [79].

Prior work [23] showed that ML architectures can learn shape-based representations when trained on *Stylized ImageNet*, a version

[1] ... so long as their spatial dimensions exceed 64 × 64. Details in Section 3.

of ImageNet where images have their visual appearances transformed by applying artistic styles. ML architectures trained on the Stylized ImageNet have better object detection performance and robustness towards a wide range of image distortions. To the best of our knowledge, existing deep-feature-based similarity metrics have not examined the performance of deep features extracted from networks trained on stylized images. We investigate whether the robust performance of ML architectures trained on Stylized ImageNet can transfer to the domain of information visualization, where humans tend to make judgments based on shape by extending the existing implementation of deep-feature-based similarity metrics, with details in Appendix B.

3 Research Goals and Experiment Overview

Our ultimate goal is to assess the benefits and limitations of deep-feature-based perceptual similarity metrics for information visualization. For this goal, we conceptually replicated two experiments:

(1) Veras and Collins [74] (using data from Pandey et al. [53])
(2) Demiralp et al. [13]

Both studies provide high-quality, crowd-sourced human similarity judgments gathered on stimuli with varying visualization complexity. Together, these studies cover both holistic similarity judgments of data visualizations and fine-grained perceptual comparisons of visual encodings.

Our replication experiments follow a consistent procedure across both studies: assign pairwise distances between stimuli from the original studies using deep-feature-based similarity metrics, analyze the results using the same procedures as the original papers, and compare our results to the analyzed human judgment results in the original studies (Fig. 1). Due to the inability of deep-feature-based similarity metrics (specifically `AlexNet`) to process images with dimensions smaller than 63×63, we generated new stimuli of size 224×224 when replicating Demiralp et al. [13] in Section 6.

4 Implementing Deep-Feature-Based Perceptual Similarity Metrics

We implement deep-feature-based perceptual similarity metrics that use ML architectures and three pre-trained weights. Our implementation primarily follows the implementations of Zhang et al. [87] and Kumar et al. [49]. We focus our attention on supervised models trained on the ImageNet-1k dataset [15], which contains 1, 281,167 training images, 50,000 validation images, and 100,000 test images.

Given two images $x, y \in \mathbb{R}^{H \times W \times C}$, where H, W, C correspond to **h**eight, **w**idth, and **c**hannel, and a network $\mathcal{F}$, the *"perceptual" distance* between x and y is the weighted sum of squared differences between feature activations of x and y across multiple layers and spatial positions. Formally, it is defined as

$$d(x, y) = \sum_{l \in \mathcal{L}} \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{x}^l_{h,w} - \hat{y}^l_{h,w} \right) \right\|_2^2 \quad (1)$$

where $\mathcal{L}$ is the set of layers in network $\mathcal{F}$ from which features are extracted[2], $\hat{x}^l, \hat{y}^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ are the unit-normalized[3] deep feature maps extracted by $\mathcal{F}$ at layer l, and vector $w_l \in \mathbb{R}^{C_l}$ is a channel-wise *scaling vector* for the difference between unit-normalized feature maps $\hat{x}^l$ and $\hat{y}^l$ at spatial location (h, w).

Using out-of-the-box pre-trained weights from ImageNet or Stylized ImageNet corresponds to $w_l = \mathbf{1}_{C_l}\ \forall l \in \mathcal{L}$, where $\mathbf{1}_{C_l}$ is a vector of ones with length C_l, the number of channels at layer l. In other words, we use the extracted deep features *without any scaling or linear calibration*. Using LPIPS (version 0.1) [4] corresponds to using ImageNet pre-trained weights *scaled by* parameters w_l learned from BAPPS [87], a dataset of 16K patches derived by applying exclusively low-level distortions to the MIT-Adobe 5k dataset [5] for training and the RAISE1k data [12] for validation. In other words, $w^l, l \in \mathcal{L}$ constitute a "perceptual calibration" of a few parameters in an existing feature space [87]. Zhang et al. [87] showed that LPIPS achieve better performance than traditional metrics (e.g., L_2, SSIM [78]) and correlates well with human similarity judgments.

Our selection of DNN architectures is motivated by both theoretical considerations and empirical evidence from computer vision and prior work on training and predicting human similarity judgments [85]. We build upon Zhang et al.'s [87] finding that deep features from CNNs trained on natural images correlate strongly with human perceptual judgments. Our architectural choices span from SqueezeNet [36] and AlexNet [47] to more complex Residual Nets [30] and Efficient Nets [69], selected to systematically investigate how different levels of hierarchical feature extraction affect visualization similarity judgments. This range of architectures, covering ImageNet-1K classification accuracies from 56.522% (AlexNet) to 83.444% (EfficientNet B5), allows us to test whether improved natural image classification correlates with better visualization similarity assessment.

While newer architectures like Vision Transformers [17] exist, we prioritized CNNs for their architectural flexibility and relevance to our context: (1) CNNs allow processing of different input sizes, unlike Vision Transformers which require fixed input dimensions; (2) CNNs provide scale-invariant feature detection through their hierarchical structure; and (3) CNNs enable direct comparability with validated perceptual similarity metrics like LPIPS [87]. This flexibility is particularly important as our implementation builds on LPIPS, which has been rigorously validated against human perceptual judgments using the Berkeley-Adobe Perceptual Patch Similarity (BAPPS) dataset at 64×64 resolution. Importantly, our approach is architecture-agnostic and can be readily extended to any CNN architecture that provides hierarchical feature maps, enabling future comparisons to additional architectures.

We use `pytorch` [54], `torchvision` weights[5], and Zhang et al.'s implementation of `LPIPS` [87] at https://github.com/richzhang/PerceptualSimilarity. We use the following `R` packages for computing and presenting our results: `reticulate` [72], `tidyverse` [81], `ggplot2` [80], `aricode` [8], `MatrixCorrelation` [37], `ggtext` [82], and `ggpubr` [40].

[2] the set of extraction layers $\mathcal{L}$ is architecture-specific; details of $\mathcal{L}$ in Appendix A
[3] at the channel dimension
[4] i.e., "`lin`" in Zhang et al. [87]
[5] https://pytorch.org/vision/stable/models.html

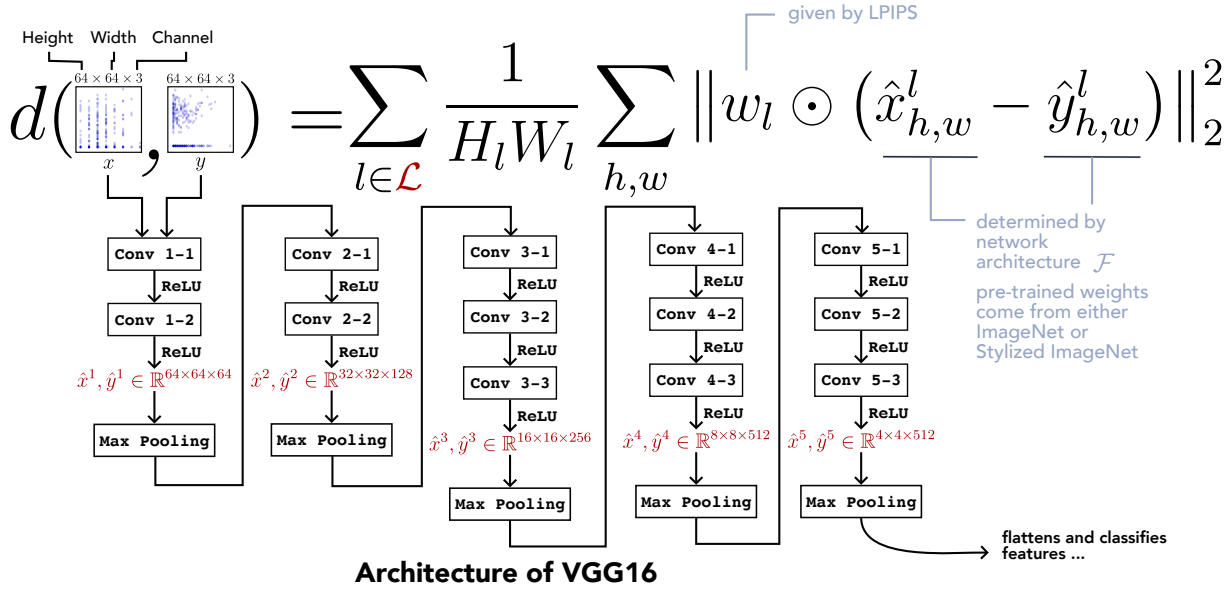


Figure 2: Diagram showing how the deep-feature-based perceptual similarity metric calculates "perceptual distance", using `VGG16` as the DL network $\mathcal{F}$.

Table 1: Network architectures and weights

Network Architecture	Weight
AlexNet [47]	ImageNet, Stylized ImageNet, LPIPS
VGG16 [67]	ImageNet, Stylized ImageNet, LPIPS
SqueezeNet [36]	ImageNet, LPIPS
ResNet18 [30]	ImageNet
ResNet50 [30]	ImageNet, Stylized ImageNet
EfficientNetB0 [69]	ImageNet
EfficientNetB5 [69]	ImageNet

4.1 Pre-processing

We follow the same image pre-processing steps of Zhang et al. [87].[6] This approach differs from the standard $224 \times 224 \times 3$ resolution that ImageNet-1K is typically trained on. For completeness, we have also run the same experiments using the standard resolution with details provided in Appendix C. While the overall trend remains similar, using the standard resolution leads to slightly worse performance than with a lower resolution of $64 \times 64 \times 3$. This finding suggests that the lower resolution may be more suitable for our specific task. By choosing a smaller size, we focus on the low-level aspects of perceptual similarity to mitigate the effect of "differing respects of similarity" [52] that may be influenced by high-level semantics [87]. Furthermore, the lower resolution yields significant improvements in computational efficiency, most notably reducing the computing time for `VGG16` from approximately three hours to merely 30 minutes.[7]

Prior research has shown that pre-processing images by transforming them into different color spaces yielded different image classification results [27]. We verify that all images in ImageNet-1K and BAPPS are in sRGB space, and we transform all input images into sRGB format before processing.

5 Replicating Veras and Collins [74]

5.1 Overview of Original Paper

Veras and Collins [74] investigated the application of a well-established computer vision metric, the Multi-Scale Structural Similarity Index (MS-SSIM) [78], to assess the *discriminability* of a data visualization across a variety of datasets. In this context, discriminability refers to "the average perceptual distance between the corresponding visualizations" for a collection of datasets [74]. Their research proposed using perceptual similarity metrics to evaluate and rank competing visual encodings based on their discriminability, thereby informing visualization selection for specific data distributions.

Veras and Collins first used gradient descent to optimize MS-SSIM parameters using stimuli from a study of scatterplot similarity judgments by Pandey et al. [53]. With this tuned MS-SSIM, they calculated pairwise distances between scatterplots and derived grouping labels using hierarchical clustering. To validate their approach, they compared their group labels against the consensus group labels from human similarity judgments in Pandey et al.'s original study [53].

[6] Namely, we convert the input stimuli into sRGB color space, resize them to $64 \times 64 \times 3$, and normalize them by the same mean and standard deviation coefficients used by Zhang et al. [87], which are the same mean and standard deviation for ImageNet-1K except adjusted for inputs ranging from [-1, 1] instead of [0, 1].

[7] Experiments were conducted on a Dell XPS 15 (2019) with Intel Core i7-8750H @ 2.20GHz, 16GB RAM, NVIDIA GeForce GTX 1050 Ti with Max-Q Design, running Microsoft Windows 11 Pro.

Figure 3: Screenshot obtained from the supplementary materials of Pandey et al. [53] showing Participant 1's final screen after completing the naming phase, with group descriptions and their corresponding easiness and confidence scores (separated by pipe characters).

5.1.1 Data, Visual Stimuli, Task. Veras and Collins [74] based their research on data from Pandey et al.'s study [53], which investigated the subjective perception of similarity in scatterplots. Pandey et al. [53] conducted a study with 18 participants from scientific backgrounds, asking them to group similar scatterplot thumbnails and provide confidence and easiness ratings for their groupings (see Fig. 3). The participants were asked to group a total of 247 scatterplots. These human-generated groupings served as a benchmark for Veras and Collins [74] to validate the performance of their tuned MS-SSIM. Data from Pandey et al. is provided at https://github.com/nyuvis/scatter-plot-similarity.

5.1.2 Similarity Measure. Given that different participants might group the same pair of scatterplots into two different groups, Pandey et al. [53] calculated the *consensus distance* between plots i and j as follows:

$$d_{i,j} = \frac{1}{N}\sum_{k=1}^{N}\left(1 - \frac{c_{i,j}}{\min\{c_i, c_j\}}\right)_k \tag{2}$$

where N is the number of participants, $c_{i,j}$ is the number of clusters that contain both plots i and j, and c_i and c_j are the number of clusters that contain the plots i and j respectively. The authors allowed participants to assign the same plot to multiple groups. These consensus distances formed a *consensus perceptual distance matrix*, which was later clustered using hierarchical clustering to form groupings.

Veras and Collins [74] conceptually replicated Pandey et al.'s experiment [53] by using an alternative similarity measure to approximate the perceptual distance between scatterplots. They used the Multiscale-Structural Similarity Index (MS-SSIM) [78], an extension of the Structural Similarity Index (SSIM) [77] where the contrast and structural similarities are computed at K image scales. For image pair X, Y, MS-SSIM is defined as

$$\text{MS-SSIM}(X, Y) = l(x, y)^{\alpha}\prod_{i=1}^{K} c(x_i, y_i)^{\beta_i} s(x_i, y_i)^{\gamma_i} \tag{3}$$

where $l(\cdot)$ is the luminance similarity function, $c(\cdot)$ is the contrast similarity function, and $s(\cdot)$ is the structural similarity function. Veras and Collins set $\alpha = 1, \beta_i = \gamma_i = w_i$ for $i \in [K]$, where $K = 5$ is the number of scales. The weights w_i can be interpreted as the relative importance of each image at scale i in determining similarity. Veras and Collins [74] employed an iterative process to adjust the scale weights $w_1, w_2, ..., w_5$ in order to minimize the discrepancy between the similarity scores calculated by MS-SSIM and a set of empirical human judgments on scatterplots.

5.1.3 Performance Evaluation. Veras and Collins used pairwise distances calculated by fine-tuned MS-SSIM to construct clustering labels for the scatterplots and compared these clustering labels against empirical judgments using four established *cluster quality measures*[8] — adjusted mutual information (AMI), normalized mutual information (NMI), Rand index (RI), and adjusted Rand index (ARI). Based on information theory, NMI and AMI quantify the shared information between two clusterings, with NMI scaling mutual information to $[0, 1]$, while AMI correcting for chance. They are robust to differences in cluster numbers and sizes. RI measures the percentage of correct decisions made by the clustering algorithm, considering all possible pairs of points, while ARI adjusts RI for chance, providing a more reliable comparison between clusterings. All measures except RI range from 0 (random clustering) to 1 (perfect clustering relative to the ground truth).

5.2 Implementation

We extend Veras and Collins' approach [74] by replacing MS-SSIM with deep-feature-based similarity metrics. One key difference between our approach and theirs is that while we exclusively rely on pre-trained ImageNet and Stylized ImageNet weights, Veras and Collins learned scaling weights for MS-SSIM using the scatterplots specific to their study. As such, their learned metric might not transfer well to different visualization types or domains.

For each ML network (Table 1), we calculate $\binom{247}{2} = 30,381$ pairwise distances between all scatterplots and turn these pairwise distances into 247×247 distance matrices. For a baseline comparison, we also calculate pairwise distances using Mean Squared Error (MSE), a popular pixel-based metric. We then apply hierarchical clustering with Ward's agglomeration strategy to compute the clusters for each distance matrix, using even-height tree cuts to yield 20 clusters — matching the number used by Pandey et al. [53]. To evaluate performance, we calculate clustering quality measures between our labels and the labels obtained by Pandey et al. [53]. We also compare our results to the results obtained by Veras and Collins [74] (represented by ◆ in Figure 4) and the baseline MSE results (represented by × in Figure 4).

5.3 Our Results

We find that our top-performing architectures (● `EfficientNet B0`, ● ● ● `AlexNet`, and ● ● `SqueezeNet`) achieve comparable (if not superior) results to those achieved by gradient-descent-tuned MS-SSIM (◆), and far superior results to those achieved by MSE, the point-wise pixel difference metric. This outcome demonstrates

[8] Cluster quality measures are traditionally used to qualify the agreement between two independent label assignments on the same dataset.

robust performance across different CNN architectures and is particularly noteworthy as it relies purely on transfer learning with no training on the existing set of scatterplots.

Comparing between architectures, we find that ● SqueezeNet (LPIPS) generally outperforms all the other networks.[9] Except for RI, the best-performing ML networks, ● EfficientNet B0, ● AlexNet, and ● SqueezeNet (LPIPS), outperform the gradient-descent-tuned MS-SSIM results (◆) obtained in Veras and Collins [74].

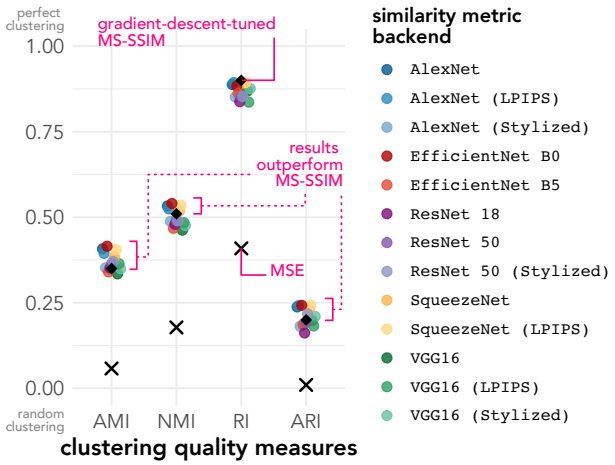


Figure 4: Black diamonds represent Veras and Collins' [74] results using gradient-descent-tuned MS-SSIM. Crosses represent results using Mean Squared Error (MSE). Points are offset for clarity.

Put differently, the clustering labels using only ImageNet pre-trained weights match human labels better than Veras and Collins' approach, which used parameters optimized for these scatterplots. Contrary to initial expectations, we find that incorporating stylized ImageNet weights does not consistently lead to performance improvements. For details, see the full set of clustered scatterplots, their labels (both from Pandey et al. [53] and from using ● SqueezeNet (LPIPS)), and the table of all clustering performance measures (Table 3) in Appendix D.

5.4 Ablation Study

To systematically analyze the factors contributing to the effectiveness of deep-feature-based similarity metrics, we conducted a series of ablation studies. The studies examine (1) the relative importance of network architecture versus learned weights, (2) the impact of different feature extraction layers on clustering quality, and (3) the role of supervised learning and dataset complexity in feature learning, comparing ImageNet-1K against simpler datasets.

5.4.1 Robustness Check — Random Weights. We initialize the same architectures with *random weights* to test whether our results stem from ImageNet pre-trained weights and assess if the resulting distance matrices and clustering labels still match empirical judgments.

[9] ● SqueezeNet (LPIPS) appear most frequently in the top-3 performing similarity metric backends for all clustering quality measures.

We repeat the process ten times and calculate the mean and 95% non-parametric bootstrap confidence interval (CI) for each architecture (Fig. 5).

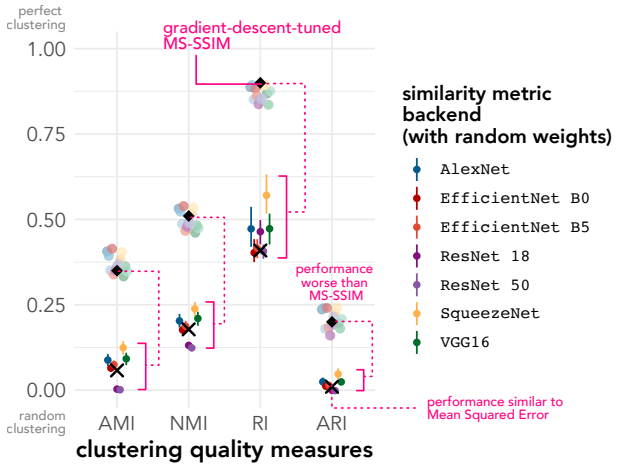


Figure 5: Black diamonds represent Veras and Collins' [74] results using gradient-descent-tuned MS-SSIM. Crosses represent results using MSE. Dots show means of 10 trials. Lines indicate 95% confidence intervals (CIs) from non-parametric bootstrap. Points and CIs are slightly offset for clarity.

We find that when using randomly initialized weights, all network architectures consistently *under-perform* the results of the same networks that use ImageNet pre-trained weights (Fig. 4). In fact, the performance across network architectures is clustered around the performance of MSE, the pixel-based metric. These findings suggest that the clustering performance of deep-feature-based similarity metrics is largely attributable to weights trained on ImageNet.

5.4.2 Impact of Each Feature Extraction Layer. To identify which feature extraction layers contributed most to the clustering quality measures, we analyze the "perceptual distance" extracted from *each individual layer* $l \in \mathcal{L}$, rather than summing across layers as in Equation (1). We also examine the effect of each layer by computing distances using *all layers except layer* $l \in \mathcal{L}$. We focus our analysis on the top-performing network architectures for each clustering quality measure.

Our analysis shows that across all examined neural network backbones, performance around layers 3 or 4 reaches performance comparable to those reported by Veras and Collins (Fig. 6). The best per-layer performance is usually the last or second-to-last layer.

From Figure 7, the performance appears relatively stable, with no specific layer significantly decreasing or increasing performance. The LPIPS version of AlexNet and SqueezeNet (i.e., ImageNet pre-trained weights + linear calibration, ● AlexNet (L) and ● SqueezeNet (L)) in general does not improve performance on top of the existing performance of ● AlexNet and ● SqueezeNet. These findings suggest that: (1) the linear calibration added by LPIPS provides limited value for scatterplot similarity; (2) while no single layer appears to dominate performance, excluding the first layer slightly

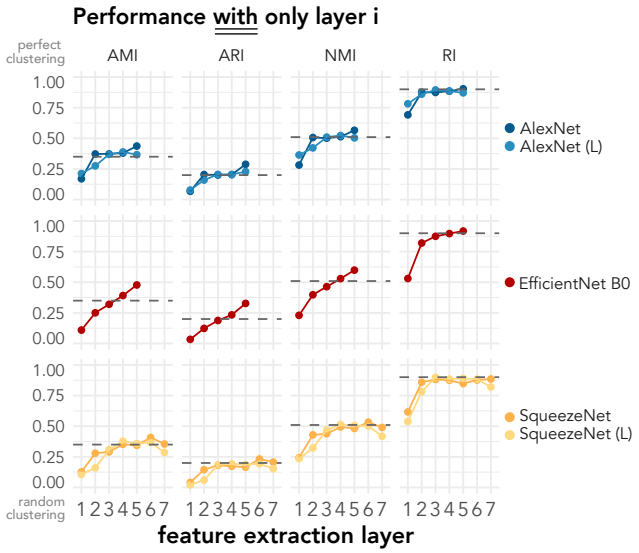


Figure 6: Clustering quality measures for ● AlexNet, ● AlexNet (LPIPS), ● EfficientNet B0, ● SqueezeNet, and ● SqueezeNet (LPIPS) calculated using perceptual distance at each layer $l \in \mathcal{L}$. Dashed lines (- - -) represent Veras and Collins' results using gradient-descent-tuned MS-SSIM.

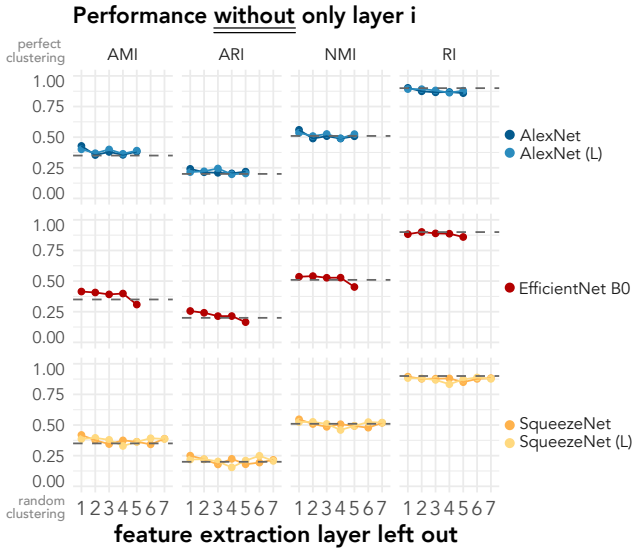


Figure 7: Clustering quality measures for ● AlexNet, ● AlexNet (LPIPS), ● EfficientNet B0, ● SqueezeNet, and ● SqueezeNet (LPIPS) calculated using all layers except layer $l \in \mathcal{L}$. Dashed lines (- - -) represent Veras and Collins' results using gradient-descent-tuned MS-SSIM.

improves performance; and (3) using only the last layer might be sufficient and could offer computational efficiency gains. These results align with prior observations [16, 21] and show that later layers encode shape and structure information that better matches human similarity judgments.

5.4.3 Impact of Training Objective and Training Dataset. We compared models pre-trained using different frameworks: supervised learning (ImageNet-1K and CIFAR-10 [46]) and self-supervised learning (SimCLR on STL-10 [11]). Pre-trained weights were sourced from established repositories (see Appendix B).

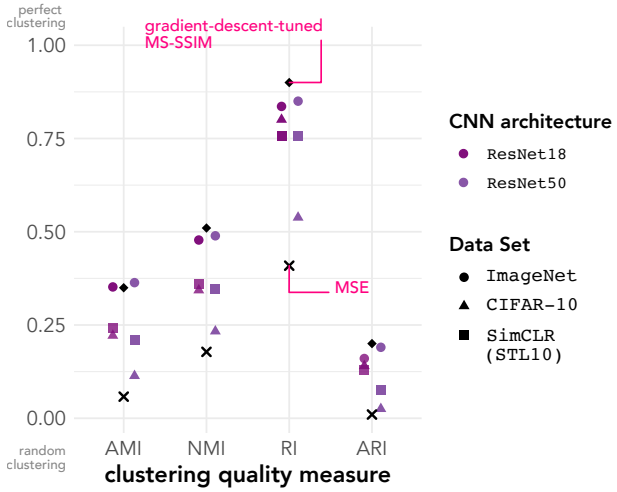


Figure 8: ● Resnet18 and ● ResNet50 performance. Black diamonds represent Veras and Collins' [74] results using gradient-descent-tuned MS-SSIM. Crosses represent results using MSE. Points are dodged for clarity.

Figure 8 reveals a clear performance hierarchy: ImageNet-trained models performed best across all clustering metrics (● ResNet50 slightly outperforming ● ResNet18), followed by ▲ CIFAR10-trained models, while SimCLR achieved similar results to ▲ CIFAR-10 but underperformed compared to ImageNet. This gap could be due to differences in both the learning framework and dataset (STL-10 vs. ImageNet-1K).

We also observe an interesting pattern where ● ResNet50 trained on ▲ CIFAR10 performs similarly to the MSE baseline. This poor performance likely stems from the mismatch between ● ResNet50's deep architecture — designed for complex, ImageNet-like datasets — and CIFAR10's low-resolution (32×32) images. With limited input complexity propagating through 50 layers, the network might default to learning shallow, pixel-level features rather than developing rich semantic representations. This demonstrates how dataset complexity and the learning framework impact representation quality and, in turn, downstream performance.

5.5 Discussion of Veras and Collins

Our replication of Veras and Collins' study [74] yields compelling results: without any training or fine-tuning on the stimuli, the best-performing deep-feature-based similarity metrics outperform

gradient-descent-tuned MS-SSIM, a traditional computer vision metric, on three out of four clustering quality measures by an average of 14.04% in aligning clustering labels to human labels of scatterplot. This is particularly noteworthy given that MS-SSIM's performance was specifically tuned on the scatterplots, while our deep learning models relied solely on pre-trained ImageNet weights. The ablation studies (Section 5.4) suggest that (1) perceptual losses extracted at later stages are generally better (Fig. 6) and (2) high-quality visual representations require both an appropriate learning framework and a sufficiently rich training dataset. The lower performance of CIFAR10-trained models (Section 5.4.3), despite using supervised learning, emphasizes that dataset characteristics (resolution, class diversity, sample complexity) may be as important as the choice of learning framework. These results provide evidence for effective transfer learning from large-scale natural image datasets to data visualizations that utilize spatial encodings, suggesting that the features learned by neural networks on datasets such as ImageNet-1K may capture some fundamental aspects of visual perception that extend beyond the domain of natural images.

6 Replicating Demiralp et al. [13]

6.1 Overview of Original Paper

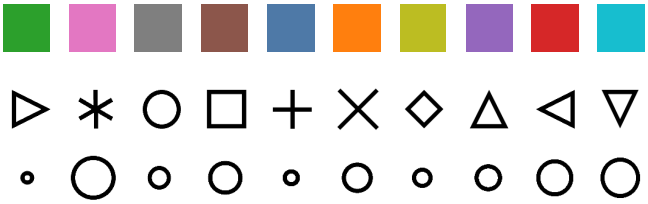


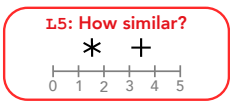
Figure 9: Palettes that maximize perceptual discriminability for color, shape, and size generated by Demiralp et al. [13], using the perceptual kernels from the triplet matching task.

Demiralp et al. [13] introduced the concept of *perceptual kernels*, which are distance matrices derived from aggregate perceptual judgments. Perceptual kernels encode perceptual differences in a reusable form that is directly applicable to visualization evaluation and automated design. For example, one can construct palettes that maximize perceptual discriminability, as shown in Figure 9.

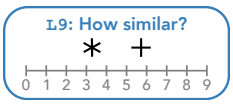
The authors conducted two sets of crowd-sourced experiments, one to estimate the perceptual kernels for *color*, *shape*, and *size*, and one to estimate perceptual kernels for the pairwise combinations (i.e., *color-shape*, *color-size*, *shape-size*). Each experiment used five different judgment types. They found that the triplet matching task, the task that asks participants to select between a pair of options that is most similar to the reference image, exhibits the least inter-subject variance, produces results that are less sensitive to subject count, and enables the most accurate prediction of bivariate kernels from univariate inputs [13].

6.1.1 Data and Visual Stimuli. In their first experiment, Demiralp et al. [13] used the default color and shape palettes from Tableau, each of which contained ten distinct values. For size, they used a palette consisting of ten circles with linearly increasing area. In the second experiment, Demiralp et al. [13] selected four values from each palette and performed a cross product, resulting in $4 \times 4 = 16$ distinct values for each pairwise combination of color, shape, and size. Their source code, experiment interface, and data are at https://github.com/uwdata/perceptual-kernels.

6.1.2 Tasks. Demiralp et al. elicited similarity judgments using five different tasks:

(1) 

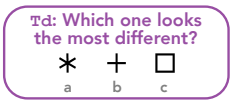
Pairwise rating on a 5-point scale (L5) presents participants with a pair of stimuli and asks them to rate the similarity between these two stimuli on a 5-point Likert scale.[10]

(2) 

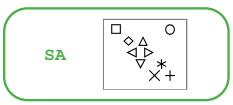
Pairwise rating on a 9-point scale (L9) presents participants with a pair of stimuli and asks them to rate the similarity between these two stimuli on a 9-point Likert scale.[11]

(3) 

Triplet ranking with matching (Tm) presents participants with three stimuli, one of which is the reference, and asks participants to select among the two remaining options which one is most similar to the reference.

(4) 

Triplet ranking with discrimination (Td) asks participants to select the *odd* stimuli out of a triplet of stimuli. This task is otherwise known as the *odd-one-out* task.

(5) 

Spatial arrangement (Sa) presents participants with all stimuli and asks them to arrange them on a 2D plane.

A total of 6 visual variables × 5 judgment types = 30 jobs were run on Amazon Mechanical Turk, with each job completed by 20 Turkers for 600 distinct subjects [13].

6.2 Implementation

We replicate parts of Demiralp et al.'s [13] experiments using deep-feature-based similarity metrics (Table 1) to generate 10×10 and 16×16 distance matrices. For a baseline comparison, we also calculate pairwise distances using Mean Squared Error (MSE), a popular pixel-based metric. Because LAB color space is "perceptually uniform" [9], meaning numerical distances between colors match human perception of those differences, we compute MSE in LAB rather than sRGB space for visual channels that involve color. Following Demiralp et al.'s approach, we normalize each matrix to span the range $[0, 1]$.

6.2.1 Data and Visual Stimuli. Demiralp et al.'s experiments used stimuli of size 36×36, which is too small for certain deep-feature-based similarity metrics to operate on. Since these stimuli consist

[10] These are actually 6-point Likert items (including 0), but we follow the naming scheme from the original work.
[11] Likewise, these are 10-point Likert items (including 0).

of elementary visual channels (single colors, shapes, and sizes) rather than complex data visualizations, we do not expect human perceptual judgments to be affected by increasing the stimulus size. Therefore, we generated 10 (16 for *size-color*) PNG images in RGB color space of size $224 \times 224 \times 3$ for each visual channel.

6.2.2 Performance Evaluation. Demiralp et al. [13] compared the degree of compatibility between the five judgment tasks by calculating **Spearman's rank correlation coefficient** on pairwise distances for the same visual channel. Spearman's rank correlation coefficient, otherwise known as Spearman's ρ, is defined as the Pearson correlation coefficient between the variables' ranks and assesses how well the relationship between two variables can be described using a monotonic function. Spearman correlation ranges between -1 and 1, with 1 indicating perfect rank correlation and -1 indicating perfectly opposite ranks. This metric has also been used to evaluate the performance of computer vision similarity metrics, such as DISTS [16], on Image Quality Assessment datasets.

6.3 Results

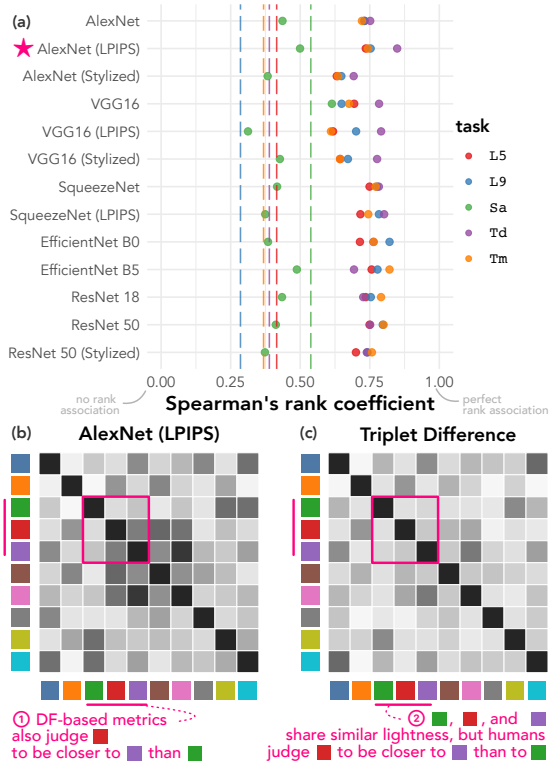


Figure 10: (a) Spearman's rank correlation between distance matrices and Demiralp et al.'s [13] perceptual kernels (dashed: MSE-based distance matrices), (b) `AlexNet (LPIPS)` distance matrix, (c) Demiralp et al.'s Td perceptual kernel.

For each of the four visual channels — *color*, *size*, *shape*, and *size-color* — we analyze the best-performing DL network.[12] For comparison, we show the perceptual kernel most correlated with the distance matrix of the best-performing DL network.[13]

6.3.1 Color. `AlexNet (LPIPS)` outperforms other DL networks. Distance matrices from deep-feature-based similarity metrics show the highest rank correlation with the perceptual kernel from Td and the lowest with Sa (Fig. 10 top).

`AlexNet (LPIPS)` demonstrates good overall performance in measuring relative color distances, but fails to replicate certain human perceptual patterns. For example, humans judge yellow to be more similar to green than teal, but `AlexNet (LPIPS)` "sees" yellow and teal as about equidistant from green.

The baseline MSE performances demonstrate the highest rank correlation with Sa, which makes sense given that MSE, a pixel-level metric, is directly measuring color differences in a perceptually uniform color space. This also suggests that Sa might be a better task to capture the non-linear aspects of human color perception. None of the pre-trained networks outperform the MSE baseline in Sa.

6.3.2 Size. `AlexNet (LPIPS)` outperforms other DL networks. Distance matrices from deep-feature-based similarity metrics and MSE show the highest rank correlation with the perceptual kernel from Sa and lowest with Td (Fig. 11 (a)).

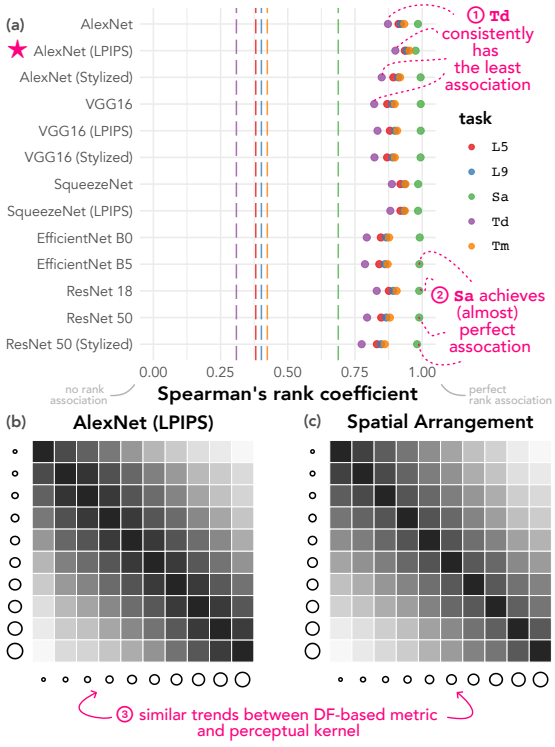


Figure 11: (a) Spearman's rank correlation between distance matrices and Demiralp et al.'s [13] perceptual kernels (dashed: MSE-based distance matrices), (b) `AlexNet (LPIPS)` distance matrix, (c) Demiralp et al.'s Sa perceptual kernel.

[12] ... which is the one that achieves the highest average Spearman rank correlation coefficient when compared to Demiralp et al.'s perceptual kernels across their five judgment tasks.

[13] The complete set of distance matrices is available in the supplementary material.

Most DL networks achieve almost perfect rank correlation with Sa (Fig. 11 (a)) while MSE achieves around 0.7, suggesting these metrics capture more fundamental aspects of human size perception than pixel-level differences.

6.3.3 Shape. `VGG16 (LPIPS)` outperforms other DL networks. Distance matrices from deep-feature-based similarity metrics show the highest rank correlation with the perceptual kernel from Tm and lowest with Sa (Fig. 12 (a)). Stylized ImageNet weights weakly improve the correlation for `VGG16` and `ResNet50` but not `AlexNet`.

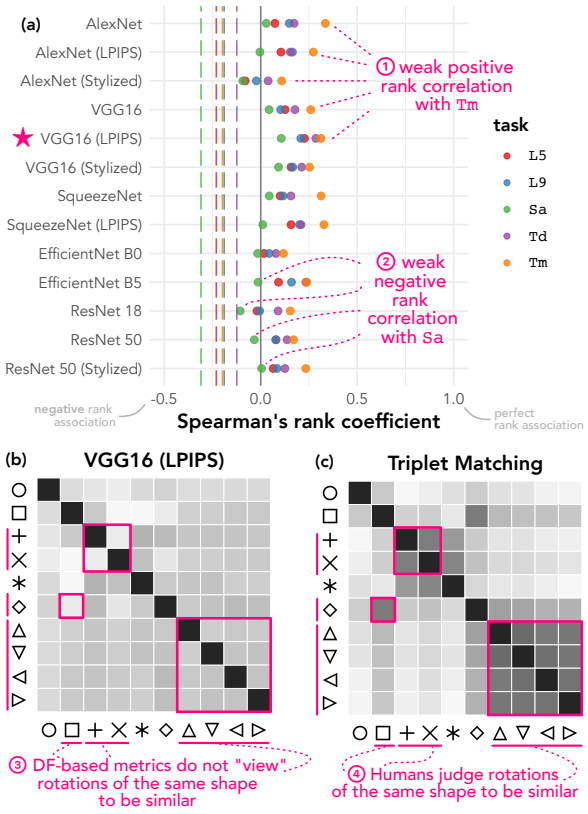


Figure 12: (a) Spearman's rank correlation between distance matrices and Demiralp et al.'s [13] perceptual kernels (dashed: MSE-based distance matrices), (b) `VGG16 (LPIPS)` distance matrix, (c) Demiralp et al.'s Tm perceptual kernel.

All MSE performances show a negative correlation, with Sa having the most negative correlation. Compared to MSE performances, the modest positive associations from CNNs show they are learning something, but there's still a large gap in capturing human shape perception. Specifically, all correlation coefficients are *weak* (i.e., $\in [-0.25, 0.5]$) compared to prior results for color (Section 6.3.1 Fig. 10) and size (Section 6.3.2 Fig. 11), where excluding Sa, rank correlations are around 0.75. Examining the distance matrix and perceptual kernel closely (Fig. 12 (b), (c)) reveals that, unlike humans, deep-feature-based perceptual similarity metrics do not view similarity as **rotation-invariant**. In other words, while humans may judge + and ×, □ and ◇ to be similar, as they are rotated variants of each other, `VGG16 (LPIPS)`, the best-performing DL network, "sees" them as different. We speculate why this is the case in Section 6.5.

6.3.4 Size-Color. `AlexNet` outperforms other DL networks. Distance matrices from deep-feature-based similarity metrics show the highest rank correlation with the perceptual kernel from Tm and the lowest with Sa (Fig. 13 (a)).

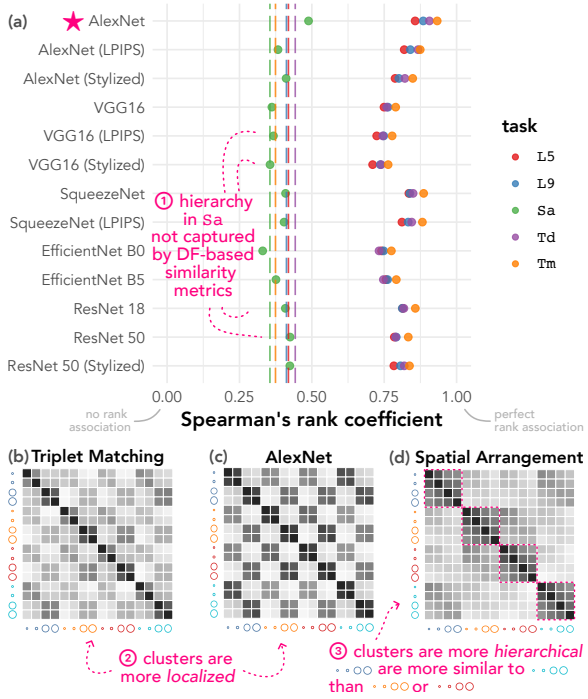


Figure 13: (a) Spearman's rank correlation between distance matrices and Demiralp et al.'s [13] perceptual kernels (dashed: MSE-based distance matrices), (b) Demiralp et al.'s Tm perceptual kernel, (c) `AlexNet` distance matrix, (d) Demiralp et al.'s Sa perceptual kernel.

Interestingly, perceptual distances obtained via Sa (Fig. 13 (d)) demonstrate a clear *hierarchy* — circles are most similar to those sharing the same color, followed by those with similar sizes, and then those with related colors (i.e., blue to teal, red to orange). In contrast, similarity judgments elicited through Tm (Fig. 13 (b)) reveal more *localized* similarity clusters, possibly due to the absence of *context* and its implied constraints when using Tm to elicit similarity judgments. This suggests a potential need to reconsider Demiralp et al.'s advice to favor triplet matching (Tm) judgments [13], especially when collecting visual variables that encode multiple visual channels simultaneously.

This hierarchy of "*color* before *size*" is not captured by either MSE or deep-feature-based similarity metrics (Fig. 13 (c)), and using ImageNet pre-trained weights in general does not improve correlation much against simple pixel-based MSE baseline.

6.4 Robustness Check — Random Weights

Similar to Section 5.4.1, we initialize the same architectures with random weights to determine to what extent ImageNet pre-trained

weights can explain our results. We repeat the process of calculating distance matrices and Spearman's rank correlation coefficients with perceptual kernels ten times and present the 95% non-parametric bootstrap CIs for Spearman's rank correlation coefficient. For each visual channel, we focus on the best-performing DL architecture and present the average distance matrix[14] that achieves the highest average Spearman correlation coefficient across all repetitions and all five judgment tasks. For direct comparison, we also present the perceptual kernel with which the average distance matrix has the highest Spearman correlation.

6.4.1 Color. Correlation results with random weights (Fig. 14 (a)) are only slightly worse than those with ImageNet pre-trained weights (Fig. 10 (a)) in Section 6.3.1, suggesting that existing deep-feature-based similarity metrics may not effectively capture the perception of color similarity.

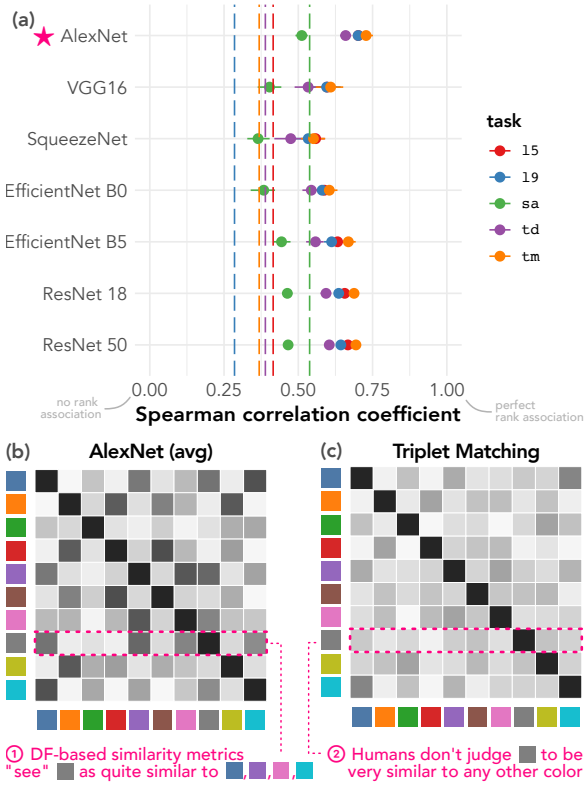


Figure 14: (a) Spearman's rank correlations between random-weight architectures' distance matrices and perceptual kernels (95% bootstrap CI; dashed: MSE-based distance matrices), (b) Random `AlexNet`'s average distance matrix, (c) Demiralp et al.'s Tm perceptual kernel.

6.4.2 Size. The correlation results for size (Fig. 15 (a)), except for `AlexNet` and `SqueezeNet`, are now around 75%, compared to perfect rank correlations for almost all networks (Fig. 11 (a)).

This suggests that deep-feature-based similarity metrics are learning something about size, but it is unclear to what extent this is

[14] All distance matrices can be found in the supplementary material.

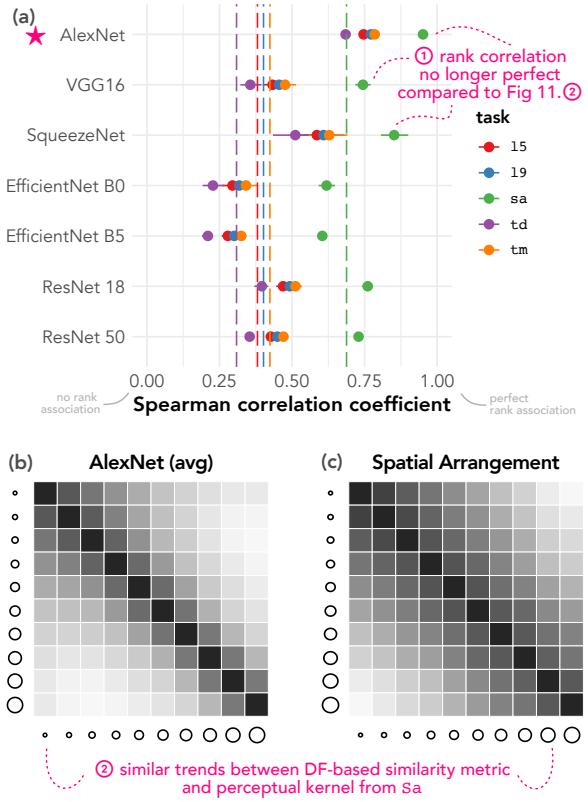


Figure 15: (a) Spearman's rank correlations between random-weight architectures' distance matrices and perceptual kernels (95% bootstrap CI; dashed: MSE-based distance matrices), (b) Random `AlexNet`'s average distance matrix, (c) Demiralp et al.'s Sa perceptual kernel.

attributable to ImageNet pre-trained weights or the functional form of deep-feature-based similarity metrics (Eq. (1)) that relies on taking Euclidean differences between deep feature activations across spatial locations. For all tasks except Sa, using randomly initialized weights result in about similar rank correlation performance as using MSE (except for `AlexNet` and `SqueezeNet`).

6.4.3 Shape. Across all DL architectures, the performance is weakly negatively correlated to all perceptual kernels obtained via different tasks (Fig. 16 (a)) and is worse than the previous correlation results (Fig. 12 (a)) in Section 6.3.3.

The performance when using randomly initialized weights is very similar to that of using MSE. Given the existing functional form of deep-feature-based similarity metrics (Eq. (1)) and its reliance on taking Euclidean differences between deep feature activations spatially, this result suggests that some transfer learning is happening in Figure 11, but not to the degree that it captures what humans perceive as shape similarity. We also suspect that glyph shape similarity judgments are not entirely perceptual but may also depend on context/concept — consider this triplet difference example: (×, □, ◯). One could choose ◯ as the odd one out, since the other shapes have angles and circles do not have angles; or

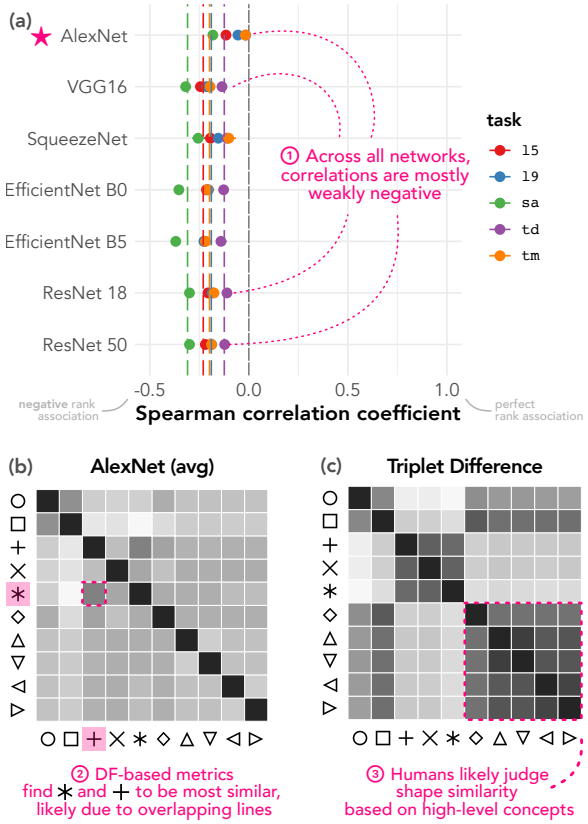


Figure 16: (a) Spearman's rank correlations between random-weight architectures' distance matrices and perceptual kernels (95% bootstrap CI; dashed: MSE-based distance matrices), (b) Random `AlexNet`'s average distance matrix, (c) Demiralp et al.'s Td perceptual kernel.

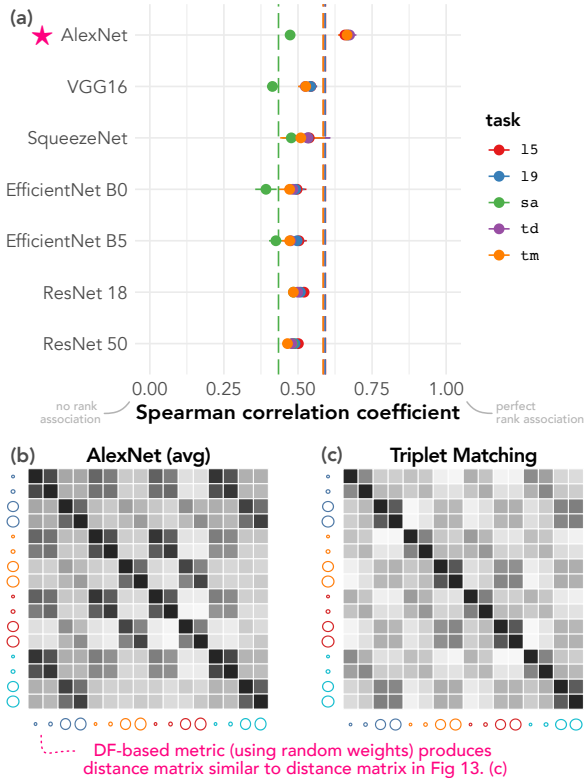


Figure 17: (a) Spearman's rank correlations between random-weight architectures' distance matrices and perceptual kernels (95% bootstrap CI; dashed: MSE-based distance matrices), (b) Radom `AlexNet`'s average distance matrix, (c) Demiralp et al.'s Tm perceptual kernel.

choose × as the odd one out, since both □ and ◯ can be drawn on a page without lifting the pen off the page.

6.4.4 Size-color. Again, we observe similar trends — the rank correlation results (Fig. 17 (a)) are slightly worse compared to the rank correlation results (Fig. 13 (a)) in Section 6.3.4 when using ImageNet pre-trained weights. For all tasks, using randomly initialized weights result in rank correlations about the same as MSE.

Across the examined visual channels, the best-performing architecture, when using randomly initialized weight, is always `AlexNet`. We suspect that the simpler and shallower architecture and the larger filter size in the earlier layers (11×11, 5×5 instead of 3×3) of `AlexNet` might "preserve" more of the raw visual information, which explains its close to 0.5 rank correlation with Sa for visual channels related to spatial information (*size* and *size-color*).

6.5 Discussion of Demiralp et al.

Although deep-feature-based similarity metrics achieve positive (Figs. 10 and 13), if not perfect (Fig. 11), rank correlation for several visual variables, the comparison against baseline MSE performances and robustness check (Section 6.4) reveals that it is unclear whether such performance can be entirely attributed to pre-trained ImageNet or Stylized ImageNet weights.

We suspect the decent correlation in the robustness check stems from both the functional form of deep-feature-based similarity metrics (Eq. (1)) and Demiralp et al.'s [13] choice of visual stimuli. Deep-feature-based similarity metrics, as we implemented, calculate *Euclidean differences between deep feature activations*, and for simple visual stimuli without complex patterns or textures, the feature map difference primarily reflects how well the stimuli *spatially align/structurally correspond*. For example, with rotated shape stimuli (+ and ×), the feature map differences remain almost identical whether using pre-trained ImageNet or random weights, given the same DL architecture.

Across all five similarity judgment tasks (L5, L9, Sa, Td, and Tm), Sa shows the least correlation with deep-feature-based similarity metrics for all visual channels except *size*, and has the least average rank correlation with all other perceptual kernels of the same visual channel (Demiralp et al. [13]). We suspect this occurs because participants see the most visual stimuli simultaneously (all 10/16 of them), and *therefore face more global, hierarchical constraints when judging their similarities*. In contrast, pairwise rating tasks (L5, L9) and triplet judgment tasks (Td, Tm) only require focusing on the local relationships between 2–3 stimuli at a time. This provides

one way to explain the Sa perceptual kernel for *size-color* (Fig. 13), where participants judge similarity first by color, then by size within each same-color group.

Our replication of Demiralp et al. [13]'s experiment demonstrates that deep-feature-based similarity metrics currently *do not approximate the perceptual distance between visual encodings* such as color and shape. Specifically, these metrics are *not rotation-invariant* (Section 6.3.3) and do not capture the *precedence effect* of certain visual channels over others in similarity judgments (Section 6.3.4). Finally, while we explored using Stylized ImageNet weights, we cannot conclude definitively whether this approach improves performance, suggesting the need for further research.

7 General Discussion

7.1 Key Findings

(1) **Effectiveness for complex visualizations**: Our replication of Veras and Collins' study [74] demonstrates that deep-feature-based similarity metrics can outperform traditional computer vision metrics (e.g., MS-SSIM) in aligning with human clustering judgments of scatterplot similarity. We highlight the fact that MS-SSIM parameters are optimized on the domain data (i.e., scatterplots) while deep-feature-based similarity metrics only rely on weights trained on *natural images* and are therefore completely domain-free. This implies that transfer learning from natural image domains to visualizations is feasible and that ImageNet pre-trained weights capture fundamental visual features that generalize across diverse visual domains. This could be good in practice by significantly reducing the effort required to develop and train models from scratch for each new set of domain-specific visualizations.

(2) **Limitations for abstract visual encodings**: In replicating Demiralp et al.'s work [13], we find that deep-feature-based metrics struggle to capture human perceptual similarities for basic visual channels like color and glyph shape, but perform well when assessing size. We hypothesize that part of the reason for this poor performance is because judgments of color and glyph shape similarity are *not purely perceptual*, that is, they also rely on high-level semantics and/or concepts. For example, different cultural associations with color hues [38] may lead people to judge the similarity of color between the same triplet differently. In terms of context/concept, see the example in Section 6.4.3. We also observe that when participants use spatial arrangement to judge visual stimuli that encode multiple channels (e.g., Fig. 13 in Section 6.3.4), they tend to judge color similarity *before* size similarity.

(3) **Architecture and weight sensitivity:** Our results show that the performance of deep-feature-based metrics does not vary significantly across different neural network architectures (Fig. 4) but does vary across different pre-trained weights of varying complexity (e.g., CIFAR10 vs ImageNet-1K in Figure 8). The impact of stylizing ImageNet is less clear.

In summary, our findings suggest a potential for deep-feature-based similarity metrics in tasks involving visualizations that exhibit texture or pattern-like features (e.g., scatterplots) and highlight a need for caution when applying these metrics to primitive visualization glyphs or elements.

7.2 Limitations and Future work

While deep-feature-based similarity metrics offer a promising solution for approximating similarity perception in information visualization, future work should explore the **generalizability** of deep-feature-based similarity metrics, focusing on:

(1) **Perceptual vs. conceptual similarity:** Our current approach focuses on low-level perceptual similarity and does not account for higher-level conceptual or semantic similarities [53] that may be important in visualization interpretation.

(2) **Visualizations are multi-modal and relational:** Our study primarily focuses on static, visual elements. An interesting direction for future work is to investigate how these metrics can be extended to handle different types of visualizations and their encoded visual relations [19, 29, 42], diverse user groups, or multimodal visualizations that incorporate text, interactivity, or other non-visual elements.

(3) **Benchmark development:** Deep-feature-based similarity metrics vary widely in their architectures [49, 87], training/tuning datasets [16], feature extraction layers [55], and evaluation approaches [68]. Comprehensive benchmarks are needed to standardize the evaluation and comparison of these metrics across domains.

(4) **Explore more DL architectures, learning frameworks, and training datasets:** While this work explores five architectures (trained for image classification), future work can investigate Vision Transformers (ViT) [73], alternative learning frameworks (e.g., self-supervised learning), and low-bit quantization. However, adapting ViT presents a fundamental challenge because the Transformer architecture relies on fixed-size sequences of patches and positional encodings. For example, the ViT-B/16 model processes 224×224 images into 196 patches (14×14), while our 64×64 resolution yields only 16 patches (4×4). This significant mismatch in sequence lengths makes it impractical to transfer pre-trained ViT weights to our deep-feature-based similarity metric without developing novel positional encoding mappings — an algorithmic challenge beyond this paper's scope.

(5) **Replication scope and inherited design choices:** By replicating prior studies, our work inherits their insights and methodological constraints in comparing human and machine behavior. While we show deep-feature-based metrics can fit into existing frameworks (as shown in Figure 1), we are limited by these studies' choices in stimulus presentation, task design, and inference procedures. Future research should validate these metrics through new experimental designs that systematically vary how human similarity judgments are elicited, representations are inferred, and human-machine behaviors are compared.

7.3 Similarity, Constraints, and Inferred Representations

Examining perceptual kernels from *size-color* (Section 6.3.4) reveals that different similarity judgment tasks likely impose *different constraints* on participants. These different constraints are then encoded in different outcome variables and translated into different inferred representations. Beyond traditional tasks such as triplet matching and triplet difference, there are also quadruplet matching [25], octet matching [63], and the lineup task [4]. Beyond the contextual constraints imposed by different similarity judgment tasks, human cognitive limitations also play a crucial role — just as Hosseinpour et al. [33] demonstrated that increasing the number of frames in a small multiples visualization leads to a linear decline in accuracy, the complexity of similarity judgments may also tax our limited cognitive processing capacity.

Moreover, different visualization tasks may require different similarity models. Existing research has predominantly modeled similarity via a *geometric model* [70]. Such models assume minimality, symmetry, and triangle inequality. However, violations of all these assumptions have been empirically observed [71]. Further research is needed to understand how different judgment tasks affect behavioral outcomes, how task constraints influence representation inference, and how different similarity models perform across visualization contexts.

7.4 Potential Applications and Broader Implications

While not yet a replacement for human testing, these metrics show promise for both pre- and post-experimental applications. For instance, they could help researchers narrow visualization design spaces before human studies by providing initial estimates of regions likely to be perceived as similar or dissimilar. They could also enable systematic comparison and evaluation of designs, such as complementing qualitative findings from human studies by quantifying how similar human-recreated charts are against the original charts they were shown [58].

This work contributes to the broader discussion about the role of computational models in understanding and automating aspects of visualization design and evaluation [14]. While our results show promise, they also highlight the complexity of human visual perception and the challenges in creating truly generalizable models. As the field progresses, interdisciplinary collaboration between visualization researchers, cognitive scientists, and machine learning experts will be crucial to developing more perceptually aligned computational models for visualization analysis and design.

8 Conclusion

In this paper, we explore the application of deep-feature-based similarity metrics to the domain of information visualization. Through two replication studies, we investigate how these metrics compare to traditional similarity measures and human judgments across scatterplots and different visual encodings. Our work demonstrates that deep features trained on diverse, large-scale natural image datasets (e.g., ImageNet-1K) transfer remarkably well to analyzing data visualizations like scatterplots, where spatial distributions are key. However, we observe limitations when applying these features to abstract visual primitives like glyph shapes or colors. This suggests that while deep features effectively capture the spatial and structural aspects of visualizations that encode data, they may be less suited for analyzing fundamental visual elements in isolation. These findings help delineate where deep perceptual metrics are most applicable in visualization analysis and design. By demonstrating both the potential and limitations of deep-feature-based similarity metrics, we contribute to the growing body of research at the intersection of machine learning and information visualization [84] and provide valuable insights for researchers and practitioners seeking to develop more sophisticated deep-learning-based tools for visualization analysis and design.

Acknowledgments

We gratefully acknowledge Noah Shen for initiating the preliminary work that led to this project, though the research ultimately took a different direction. We thank Enrico Bertini for providing the 247 scatterplots originally used in Pandey et al. [53], and Fumeng Yang for her invaluable feedback throughout our research process. This work would not have been possible without the publicly available data from Veras and Collins [74] and Demiralp et al. [13], for which we are deeply appreciative. We also thank the members of the Mu Collective for their continued support and the reviewers for their constructive feedback. This work was partially supported by the National Science Foundation under award No. IIS-1930642.

References

[1] Ioanna Maria Attarian, Brett D Roads, and Michael Curtis Mozer. 2020. Transforming neural network visual representations to predict human judgments of similarity. In *NeurIPS 2020 Workshop SVRHM*. NeurIPS. doi:10.48550/arXiv.2010.06512

[2] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. 2018. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology* 14, 12 (2018), e1006613. doi:10.1371/journal.pcbi.1006613

[3] Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. 2023. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences* 46 (2023), e385. doi:10.1017/s0140525x22002813

[4] Andreas Buja, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F Swayne, and Hadley Wickham. 2009. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367, 1906 (2009), 4361–4383. doi:10.1098/rsta.2009.0120

[5] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*. 97–104. doi:10.1109/CVPR.2011.5995413

[6] Qing Chen, Shixiong Cao, Jiazhe Wang, and Nan Cao. 2024. How Does Automation Shape the Process of Narrative Visualization: A Survey of Tools. *IEEE Transactions on Visualization and Computer Graphics* 30, 8 (2024), 4429–4448. doi:10.1109/TVCG.2023.3261320

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 149, 11 pages.

[8] Julien Chiquet, Guillem Rigaill, Martina Sundqvist, Valentin Dervieux, and Florent Bersani. 2020. Package 'aricode'. *R package version* (2020). doi:10.32614/CRAN.package.aricode

[9] C CIE. 1978. Recommendations on uniform color spaces, color difference equations, psychometric color terms. *CIE Publication* 15 (1978), 9–12. doi:10.1002/j.1520-6378.1977.tb00102.x

[10] William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554. doi:10.2307/2288400

[11] Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 215–223.

[12] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. 2015. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM multimedia systems conference*. 219–224. doi:10.1145/2713168.2713194

[13] Çağatay Demiralp, Michael S Bernstein, and Jeffrey Heer. 2014. Learning perceptual kernels for visualization design. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1933–1942. doi:10.1109/TVCG.2014.2346978

[14] Çagatay Demiralp, Carlos E. Scheidegger, Gordon L. Kindlmann, David H. Laidlaw, and Jeffrey Heer. 2014. Visual Embedding: A Model for Visualization. *IEEE* 34, 1 (2014), 10–15. doi:10.1109/MCG.2014.18

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255. doi:10.1109/CVPR.2009.5206848

[16] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. 2020. Image Quality Assessment: Unifying Structure and Texture Similarity. (2020), 1–1. doi:10.1109/TPAMI.2020.3045810

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV] https://arxiv.org/abs/2010.11929

[18] Ahmet M Eskicioglu and Paul S Fisher. 1995. Image quality measures and their performance. *IEEE Transactions on communications* 43, 12 (1995), 2959–2965. doi:10.1109/26.477498

[19] François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. 2011. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences* 108, 43 (2011), 17621–17625. doi:10.1073/pnas.1109168108

[20] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344* (2023). doi:10.48550/arXiv.2306.09344

[21] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. *A Neural Algorithm of Artistic Style*. doi:10.48550/arXiv.1508.06576

[22] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. 2020. Beyond Accuracy: Quantifying Trial-by-Trial Behaviour of CNNs and Humans by Measuring Error Consistency. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 13890–13902. https://proceedings.neurips.cc/paper/2020/hash/9f6992966d4c363ea0162a056cb45fe5-Abstract.html

[23] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* (2018). doi:10.48550/arXiv.1811.12231

[24] Bernd Girod. 1993. What's wrong with mean-squared error? In *Digital images and human vision*. 207–220. doi:10.1109/MDSP.1991.639240

[25] Anna Gogolou, Theophanis Tsandilas, Themis Palpanas, and Anastasia Bezerianos. 2018. Comparing similarity perception in time series visualizations. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 523–533. doi:10.1109/TVCG.2018.2865077

[26] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014). doi:10.48550/arXiv.1412.6572

[27] Shreyank N Gowda and Chun Yuan. 2019. ColorNet: Investigating the importance of color spaces for image classification. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*. Springer, 581–596. doi:10.1007/978-3-030-20870-7_36

[28] Guo-Dong Guo, Anil K Jain, Wei-Ying Ma, and Hong-Jiang Zhang. 2002. Learning similarity measure for natural image retrieval with relevance feedback. *IEEE Transactions on Neural Networks* 13, 4 (2002), 811–820. doi:10.1109/TNN.2002.1021882

[29] Daniel Haehn, James Tompkin, and Hanspeter Pfister. 2019. Evaluating 'Graphical Perception' with CNNs. 25, 1 (2019), 641–650. doi:10.1109/TVCG.2018.2865138

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778. doi:10.1109/CVPR.2016.90

[31] Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. 2019. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one* 14, 10 (2019), e0223792. doi:10.1371/journal.pone.0223792

[32] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour* 4, 11 (2020), 1173–1185. doi:10.1038/s41562-020-00951-3

[33] Helia Hosseinpour, Laura E Matzen, Kristin M Divis, Spencer C Castro, and Lace Padilla. 2024. Examining Limits of Small Multiples: Frame Quantity Impacts Judgments with Line Graphs. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[34] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510. doi:10.48550/arXiv.1703.06868

[35] Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. 2013. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on visualization and computer graphics* 19, 12 (2013), 2406–2415. doi:10.1109/TVCG.2013.119

[36] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360* (2016). doi:10.48550/arXiv.1602.07360

[37] Ulf Geir Indahl, Tormod Næs, and Kristian Hovde Liland. 2018. A similarity index for comparing coupled matrices. *Journal of Chemometrics* e3049 (2018). doi:10.1002/cem.3049

[38] Laurence Jacobs, Charles Keown, Reginald Worthley, and Kyung-Il Ghymn. 1991. Cross-cultural colour comparisons: Global marketers beware! *International marketing review* 8, 3 (1991). doi:10.1108/02651339110137279

[39] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 694–711. doi:10.1007/978-3-319-46475-6_43

[40] Alboukadel Kassambara. 2023. *ggpubr: 'ggplot2' Based Publication Ready Plots*. https://rpkgs.datanovia.com/ggpubr/ R package version 0.6.0.

[41] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology* 10, 11 (2014), e1003915. doi:10.1371/journal.pcbi.1003915

[42] Junkyung Kim, Matthew Ricci, and Thomas Serre. 2018. Not-So-CLEVR: learning same–different relations strains feedforward neural networks. *Interface focus* 8, 4 (2018), 20180011. doi:doi.org/10.1098/rsfs.2018.0011

[43] Younghoon Kim, Kanit Wongsuphasawat, Jessica Hullman, and Jeffrey Heer. 2017. Graphscape: A model for automated reasoning about visualization similarity and sequencing. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2628–2638. doi:10.1145/3025453.3025866

[44] Gordon Kindlmann and Carlos Scheidegger. 2014. An algebraic process for visualization design. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2181–2190. doi:10.1109/TVCG.2014.2346325

[45] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* 2 (2008), 249. doi:10.3389/neuro.06.004.2008

[46] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009). http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012). doi:10.1145/3065386

[48] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. 2018. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv* (2018), 408385. doi:10.1101/408385

[49] Manoj Kumar, Neil Houlsby, Nal Kalchbrenner, and Ekin D. Cubuk. 2022. *Do Better ImageNet Classifiers Assess Perceptual Similarity Better?* doi:10.48550/arXiv.2203.04946

[50] Eric C Larson and Damon M Chandler. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging* 19, 1 (2010), 011006–011006. doi:10.1117/1.3267105

[51] Yuxin Ma, Anthony KH Tung, Wei Wang, Xiang Gao, Zhigeng Pan, and Wei Chen. 2018. Scatternet: A deep subjective similarity model for visual analysis of scatterplots. *IEEE transactions on visualization and computer graphics* 26, 3 (2018), 1562–1576. doi:10.1109/TVCG.2018.2875702

[52] Douglas L. Medin, Robert L. Goldstone, and Dedre Gentner. 1993. Respects for Similarity. 100, 2 (1993), 254–278. doi:10.1037/0033-295X.100.2.254

[53] Anshul Vikram Pandey, Josua Krause, Cristian Felix, Jeremy Boy, and Enrico Bertini. 2016. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3659–3669. doi:10.1145/2858036.2858155

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019). doi:10.48550/arXiv.1912.01703

[55] Gustav Grund Pihlgren, Konstantina Nikolaidou, Prakash Chandra Chhipa, Nosheen Abid, Rajkumar Saini, Fredrik Sandin, and Marcus Liwicki. 2023. A Systematic Performance Analysis of Deep Perceptual Loss Networks: Breaking Transfer Learning Conventions. *arXiv preprint arXiv:2302.04032* (2023). doi:10.48550/arXiv.2302.04032

[56] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication* 30 (2015), 57–77. doi:10.1016/j.image.2014.10.009

[57] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. 2018. *PieAPP: Perceptual Image-Error Assessment through Pairwise Preference*. doi:10.48550/arXiv.1806.02067 arXiv:1806.02067 [cs]

[58] Rifat Ara Proma, Michael Correll, Ghulam Jilani Quadri, and Paul Rosen. 2024. Visual Stenography: Feature Recreation and Preservation in Sketches of Line Charts. https://ieeevis.org/year/2024/program/poster_v-vis-posters-1100.html

[59] Zening Qu and Jessica Hullman. 2016. Evaluating visualization sets: Trade-offs between local effectiveness and global consistency. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*. 44–52. doi:10.1145/2993901.2993910

[60] Zening Qu and Jessica Hullman. 2017. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 468–477. doi:10.1109/TVCG.2017.2744198

[61] Brett D Roads and Bradley C Love. 2021. Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 3547–3557. doi:10.48550/arXiv.2011.11015

[62] Brett D Roads and Bradley C Love. 2024. Modeling similarity and psychological space. *Annual Review of Psychology* 75, 1 (2024), 215–240. doi:10.1146/annurev-psych-040323-115131

[63] Brett D Roads and Michael C Mozer. 2017. Improving human-machine cooperative classification via cognitive theories of similarity. *Cognitive science* 41, 5 (2017), 1394–1411. doi:10.1111/cogs.12400

[64] Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. 2009. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing* 18, 11 (2009), 2385–2401. doi:10.1109/TIP.2009.2025923

[65] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv* (2018), 407007. doi:10.1101/407007

[66] Hamid R Sheikh. 2003. Image and video quality assessment research at LIVE. http://live.ece.utexas.edu/research/quality/

[67] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. doi:10.48550/arXiv.1409.1556

[68] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. 2023. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018* (2023). doi:10.48550/arXiv.2310.13018

[69] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114. doi:10.48550/arXiv.1905.11946

[70] Warren S Torgerson. 1965. Multidimensional scaling of similarity. *Psychometrika* 30, 4 (1965), 379–393. doi:10.1007/BF02289530

[71] Amos Tversky. 1977. Features of similarity. *Psychological review* 84, 4 (1977), 327. doi:10.1037/0033-295X.84.4.327

[72] Kevin Ushey, JJ Allaire, and Yuan Tang. 2024. *reticulate: Interface to 'Python'*. https://rstudio.github.io/reticulate/ R package version 1.39.0, https://github.com/rstudio/reticulate.

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[74] Rafael Veras and Christopher Collins. 2019. Discriminability tests for visualization effectiveness and scalability. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 749–758. doi:10.1109/TVCG.2019.2934432

[75] Catherine Wah, Subhransu Maji, and Serge Belongie. 2015. Learning Localized Perceptual Similarity Metrics for Interactive Categorization . In *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE Computer Society, Los Alamitos, CA, USA, 502–509. doi:10.1109/WACV.2015.73

[76] Qianwen Wang, Zhutian Chen, Yong Wang, and Huamin Qu. 2021. A survey on ML4VIS: Applying machine learning advances to data visualization. *IEEE transactions on visualization and computer graphics* 28, 12 (2021), 5134–5153. doi:10.1109/TVCG.2021.3106142

[77] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861

[78] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. Ieee, 1398–1402. doi:10.1109/ACSSC.2003.1292216

[79] Felix A Wichmann and Robert Geirhos. 2023. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science* 9, 1 (2023), 501–524. doi:10.1146/annurev-vision-120522-031739

[80] Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

[81] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4, 43 (2019), 1686. doi:10.21105/joss.01686

[82] Claus O Wilke and Brenton M Wiernik. 2020. ggtext: Improved text rendering support for 'ggplot2'. *R package version 0.1* 1 (2020). https://wilkelab.org/ggtext/

[83] Jeremy M Wolfe. 2020. Visual search: How do we find what we are looking for? *Annual review of vision science* 6, 1 (2020), 539–562. doi:10.1146/annurev-vision-091718-015048

[84] Aoyu Wu, Yun Wang, Xinhuan Shu, Dominik Moritz, Weiwei Cui, Haidong Zhang, Dongmei Zhang, and Huamin Qu. 2021. Ai4vis: Survey on artificial intelligence approaches for data visualization. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2021), 5049–5070. doi:10.1109/TVCG.2021.3099002

[85] Fumeng Yang, Yuxin Ma, Lane Harrison, James Tompkin, and David H. Laidlaw. 2023. How Can Deep Neural Networks Aid Visualization Perception Research? Three Studies on Correlation Judgments in Scatterplots. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 822, 17 pages. doi:10.1145/3544548.3581111

[86] Zehua Zeng, Phoebe Moh, Fan Du, Jane Hoffswell, Tak Yeon Lee, Sana Malik, Eunyee Koh, and Leilani Battle. 2021. An evaluation-focused framework for visualization recommendation algorithms. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 346–356. doi:10.48550/arXiv.2109.02706

[87] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. doi:10.48550/arXiv.1801.03924 arXiv:1801.03924 [cs]

A Neural Network Architecture and Extraction Points

We followed the implementation of Zhang et al. [87] when extracting deep features from `AlexNet`, `SqueezeNet`, and `VGG16`. We followed the implementation of Kumar et al. [49] when extracting deep features from `ResNets` and `EfficientNets`. See Table 2 for details. Other works, such as Pihlgren et al. [55], have extracted four points for each network such that the chosen points represent "early, semi-early, middle, and late" layers in the convolutional layers. While Pihlgren et al.'s approach offers a more systematic way to sample features across network depth, we opted to align our extraction points with Zhang et al. [87] and Kumar et al. [49] to enable direct comparisons, though it's worth noting that there is no broad consensus in the field regarding optimal feature extraction locations.

Table 2: Network architecture, feature extraction points, ImageNet top-1 accuracy, and model size

Network architecture	Feature extraction points	Top-1 accuracy (%)	Model size (MB)	Num of Parameters
`AlexNet`	1st, 2nd, 3rd, 4th, and 5th ReLU	56.522	233.1	61, 100, 840
`VGG16`	2nd, 4th, 7th, 10th, and 13th ReLU	71.592	527.8	138, 357, 544
`SqueezeNet`	1st ReLU, 2nd, 4th, 5th, 6th, 7th, and 8th Fire	58.178	4.7	1, 235, 496
`ResNet18`,	1st Conv2d, 1st MaxPool2d,	69.758	44.7	11, 689, 512
`ResNet50`	2nd, 3rd, and 4th Block Stack	76.13	97.8	25, 557, 032
`EfficientNet B0`,	1st Conv2d, 2nd, 3rd, 4th, and 6th MBConv	77.692	20.5	5, 288, 548
`EfficientNet B5`	(mobile inverted bottleneck)	83.444	116.9	30, 389, 784

B Neural Network Weights

We obtained Stylized ImageNet weights from Geirhos et al. [23]. Stylized ImageNet is generated by AdaIn style transfer [34] and the code is available at https://github.com/rgeirhos/Stylized-ImageNet. We utilize the Stylized ImageNet weights for `AlexNet`, `VGG16`, and `ResNet50` trained by Geirhos et al. [23] at https://github.com/rgeirhos/texture-vs-shape/. For `ResNet-50`, Geirhos et al. [23] used the standard `ResNet-50` architecture from `PyTorch` [54] (i.e., the `torchvision.models.resnet50` implementation). Geirhos et al. [23] used batch size of 256 and trained on Stylized ImageNet for 60 epochs with Stochastic Gradient Descent (`torch.optim.SGD`) using a momentum term of 0.5, weight decay `1e-4` and a learning rate of 0.1 which was multiplied by a factor of 0.1 after 20 and 40 epochs of training. For `AlexNet` and `VGG16`, they used model architectures from `torchvision.models` and trained the networks under the identical circumstances as `ResNet-50`. Identical hyperparameter setting except for the learning rate — the learning rate for `AlexNet` was set to 0.001 and for `VGG16` was 0.01 initially. Both learning rates were multiplied by 0.1 after 20 and 40 epochs of training (60 epochs in total).

We obtained CIFAR-10 weights [46] trained for `ResNet18` and `ResNet50` from https://github.com/huyvnphan/PyTorch_CIFAR10. We obtained the SimCLR weights [7] trained for `ResNet18` and `ResNet50` from https://github.com/sthalles/SimCLR. We use the same feature extraction points as specified in Table 2.

C Pre-processing with Resolution 224×224

Similar to prior observations [49], we also found that using standard ImageNet resolution (224×224) yields slightly worse performance compared to a resolution of 64×64.

D Clustering of Scatterplots

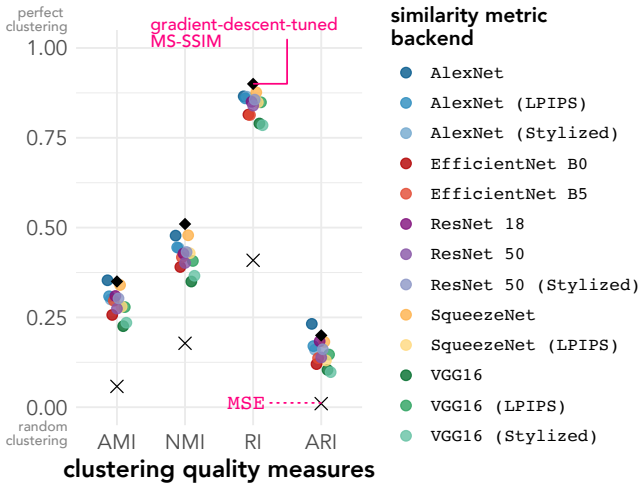


Figure 18: Black diamonds represent Veras and Collins' [74] results using gradient-descent-tuned MS-SSIM. Points are slightly offset for clarity. Results when replicating Pandey et al's experiment using resolution $224 \times 224 \times 3$.

Table 3: Clustering quality measures obtained from 64×64, highlighting the top three performing backbones for each measure.

Backbone	RI	ARI	AMI	NMI
Veras and Collins	0.90	0.20	0.35	0.51
Mean Squared Error	0.409	0.010	0.0579	0.178
AlexNet	0.8878905	0.2378972	**0.4078301**	**0.5328779**
AlexNet (Stylized ImageNet)	0.8517824	0.1814445	0.3521323	0.4876277
AlexNet (LPIPS)	**0.8935190**	**0.2419678**	0.3943937	0.5241815
VGG16	0.8688325	0.1978015	0.3345852	0.4622434
VGG16 (Stylized ImageNet)	0.8759751	0.2099710	0.3494193	0.4757726
VGG16 (LPIPS)	0.8363122	0.1822110	0.3643023	0.4839279
SqueezeNet	**0.8917086**	0.2286150	0.3841792	0.5179834
SqueezeNet (LPIPS)	**0.8939469**	**0.2433410**	**0.4053757**	**0.5353521**
EfficientNet B0	0.8833152	**0.2424203**	**0.4153032**	**0.5400465**
EfficientNet B5	0.8613278	0.1856533	0.3402646	0.4677815
ResNet 18	0.8375959	0.1617837	0.3518211	0.4793912
ResNet 50	0.8516507	0.1918074	0.3637146	0.4907449
ResNet 50 (Stylized ImageNet)	0.8531648	0.2176265	0.3707914	0.4893624

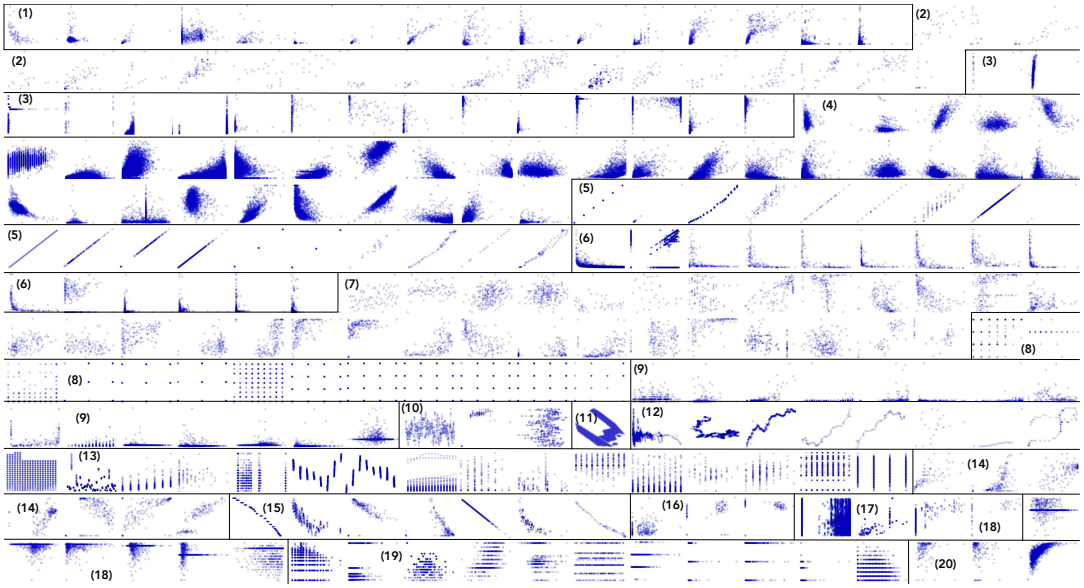


Figure 19: Consensus clustering of scatterplots from Pandey et al. [53].

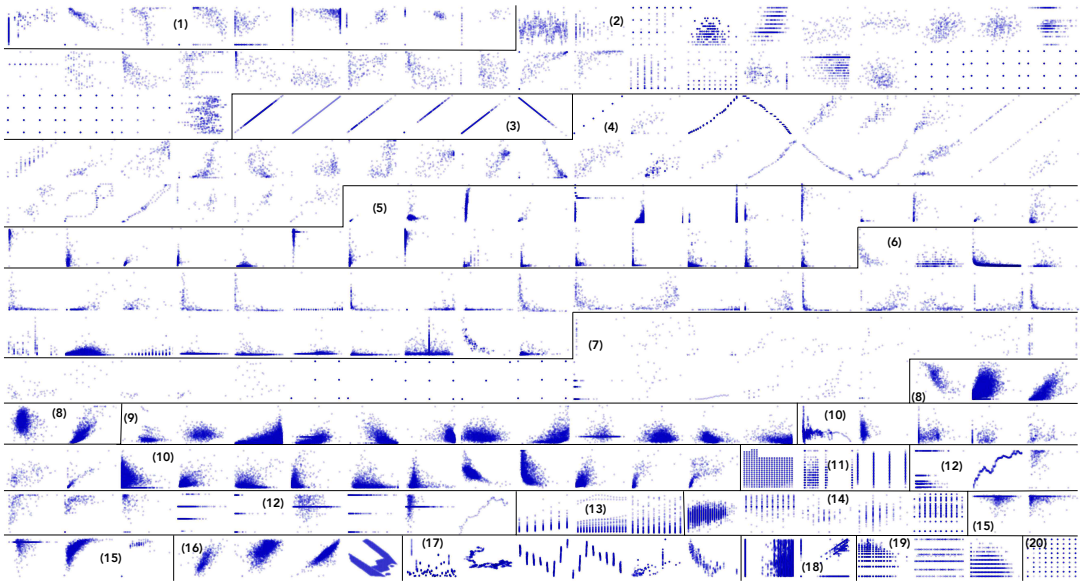


Figure 20: Clustering labels of scatterplots from Pandey et al. [53] using `SqueezeNet (LPIPS)` as the backbone network for the deep-feature-based similarity metric.