

Generalization of CNNs on Relational Reasoning with Bar Charts

Zhenxing Cui, Lu Chen, Yong Wang, Daniel Haehn, Hanspeter Pfister, and Yunhai Wang

Abstract—This paper presents a systematic study of the generalization of convolutional neural networks (CNNs) and humans on relational reasoning tasks with bar charts. We first revisit previous experiments on graphical perception and update the benchmark performance of CNNs. We then test the generalization performance of CNNs on a classic relational reasoning task: estimating bar length ratios in a bar chart, by progressively perturbing the standard visualizations. We further conduct a user study to compare the performance of CNNs and humans. Our results show that CNNs outperform humans only when the training and test data have the same visual encodings. Otherwise, they may perform worse. We also find that CNNs are sensitive to perturbations in various visual encodings, regardless of their relevance to the target bars. Yet, humans are mainly influenced by bar lengths. Our study suggests that robust relational reasoning with visualizations is challenging for CNNs. Improving CNNs' generalization performance may require training them to better recognize task-related visual properties.

Index Terms—Convolutional Neural Networks, Generalization Evaluation, Graphical Perception, Relational Reasoning.

I. INTRODUCTION

DEEP neural networks, especially convolutional neural networks (CNNs), are increasingly being adopted in the visualization community for many tasks such as visual question answering [33], [34], automatic visualization design [3], and chart captioning [35], [44]. Despite their widespread use, the crucial question of how well these models generalize to previously unseen visualizations remains less explored. Understanding and enhancing this generalization ability is crucial for the real-world deployment of CNNs.

Graphical perception [5] refers to the human ability to decode visually encoded quantities in visualizations. It plays a foundational role in understanding the relations between visual elements, such as the bar length ratios in bar charts. Seminal work by Cleveland and McGill [5], along with subsequent studies [26], [55] highlighted human proficiency with aligned bars over stacked configurations. Recent findings by Haehn et

Zhenxing Cui & Lu Chen contribute equally to this work, and Yunhai Wang is the corresponding author.

Zhenxing Cui is with the School of Computer Science and Technology, Shandong University, China. E-mail: cuizhenxing0423@gmail.com

Lu Chen is with the State Key Lab of CAD&CG, Zhejiang University, China. E-mail: lu.chen@zju.edu.cn

Yong Wang is with the College of Computing and Data Science, Nanyang Technological University, Singapore. E-mail: yong-wang@ntu.edu.sg

Daniel Haehn is with the College of Science and Mathematics, University of Massachusetts Boston, USA. E-mail: haehn@cs.umb.edu

Hanspeter Pfister is with the School of Engineering and Applied Sciences, Harvard University, USA. E-mail: pfister@seas.harvard.edu

Yunhai Wang is with the School of Information, Renmin University of China, China. E-mail: wang.yh@ruc.edu.cn

Manuscript received May 22, 2023; revised September 6, 2024.

al. [20] indicate that CNNs do not match human accuracy in interpreting bar charts, raising questions about the discrepancy in graphical perception between CNNs and humans.

Our research aims to address two primary shortcomings of the current evaluation of CNNs in graphical perception. First, a common limitation of existing studies is their reliance on oversimplified stimuli that fail to capture the complexity of standard visualizations. For example, the stimuli used by Haehn et al. [20] were all black-and-white images devoid of axes, legends, or titles, which are essential chart components according to the Grammar of Graphics [59]. We address this limitation by evaluating the performance of CNNs and humans using more realistic bar charts that better reflect standard visualization design.

Furthermore, although CNNs have demonstrated remarkable performance in object recognition, their robustness to visual changes is limited, as highlighted by Geirhos et al. [15]. We hypothesize this limited robustness extends to relational reasoning tasks, particularly in the context of commonly used bar charts, which have not been thoroughly investigated. Our study seeks to systematically evaluate the generalization capabilities of CNNs in visual relational reasoning, focusing on their adaptability to novel charts with variations in key visualization parameters such as stroke width, title position, and background color, which reflect a wide range of design choices driven by individual preferences.

In this paper, we present a comprehensive study of the graphical perception capabilities of CNNs and humans. We focus on the ratio estimation task in bar charts and aim to answer the following research questions:

- (i) How well do CNNs perform and generalize on standard visualizations that include all necessary components specified by a visualization grammar?
- (ii) How robust are CNNs to perturbations of different visual properties in standard visualizations?
- (iii) What are the major differences between CNNs and humans in visual relational reasoning?

Our answers to these questions will help us better understand the potential generalization ability of CNNs and humans and provide guidelines for training and designing robust CNNs for relational reasoning tasks involving data visualizations.

We first revisit the experiments conducted by Haehn et al. [20] and find that a CNN model trained with the identified (near) optimal configurations can estimate bar length ratios more accurately than humans when the training and test data have similar visual encodings. Moreover, this CNN model can also evaluate bar charts with an expanded parametric

space, such as an increased number of possible ratios in ratio estimation tasks.

Next, we conduct a structured generalization analysis of CNNs and humans on standard visualizations with a wider range of visual appearances, synthesized using Vega-lite [52], a high-level visualization grammar. We evaluate the performance of CNNs on visualizations without and with progressive perturbations of various visual parameters, such as title position, background color, bar width, and bar length.

We also conduct a user study to measure human performance under the same conditions. The results show that: (i) CNNs perform better than humans in relational reasoning on bar charts when the testing data mirrors the visual encodings of the training data. However, their performance can quickly decline below human levels on charts that incorporate perturbations in certain visual encodings; and (ii) CNNs exhibit less robustness to perturbations of various visual encodings, even those unrelated to the target visual marks. In contrast, humans are mainly influenced by the lengths of the bars in the chart.

Our study suggests that humans and CNNs may use different reasoning processes, with humans primarily focusing on specific parameters of the target bars while ignoring other irrelevant chart details. To understand why CNNs fail to generalize to perturbed visualizations, we visualize the important regions used by CNNs for estimating the ratios using gradient weighted class activation mapping (Grad-CAM) [53] and find that the regions of target bars are rarely used as the key evidence by CNNs trained on this task. However, by providing enough guidance of the target regions, their ability to generalize to perturbed visualizations can be improved. Even so, CNNs still do not use the lengths of the bars as their sole basis for estimating ratios.

In summary, this work makes several novel contributions. We systematically study the generalization performance of humans and CNNs on a large number of standard visualizations synthesized according to the grammar of graphics. We demonstrate that appropriately tuned hyperparameters can enable most CNN models to meet or exceed human performance on relational reasoning tasks when the training and test data have similar visual encodings. Finally, we show that providing ground-truth target region information can improve the generalization performance of CNNs on graphical perception tasks. Data, source code, study materials and additional analyses are included in our supplemental material and can be found at https://osf.io/6dxyz/?view_only=5a62778bf9e744c5bfe8c2a41572f946.

II. RELATED WORK

A. Graphical Perception

The concept of graphical perception was introduced by Cleveland and McGill to describe the visual decoding of information encoded in graphs [5], [6]. By conducting a series of elementary perceptual experiments on different visual encodings, Cleveland and McGill evaluated the perceptual difficulty of different visual encodings for humans. Their work was later replicated by a number of follow-up studies, such as Heer and Bostock's evaluation of human graphical perception

on Mechanical Turk [26] and Harrison et al.'s reproduction of the Cleveland-McGill experiments with a focus on the influence of affective priming [23]. These studies have established a foundation for understanding human capabilities in interpreting information from visualizations. Building on this body of knowledge, Zeng et al. [64] leverage insights from graphical perception to guide visualization recommendation.

Recently, researchers have focused on the building blocks of visualizations, such as position, length, and angle estimation, and have explored how well neural networks can perform these elementary perceptual tasks. Haehn et al. [20] investigated whether off-the-shelf CNNs can predict human responses to graphical perception building blocks by replicating Cleveland and McGill's experiments [5] (see Section III). Haleem et al. [21] explored the potential of using a handcrafted CNN to estimate the readability of graph visualizations and found that machine graphical perception has promise for graph visualizations. Giovannangeli et al. [16] found that current CNNs are able to replicate previously published studies of human graph perception while still exhibiting generalization limitations.

However, all of these studies evaluate the generalization ability of CNNs using data generated from highly controlled settings, which are far from real-world scenarios. We argue that the evaluation should be done with standard visualizations created according to the grammar of graphics [59]. Otherwise, the results may not be representative of meaningful generalization, and such controlled settings may limit the usefulness of CNNs. In addition, we aim to investigate the robustness of CNNs against varying visual properties and compare them with human performance.

B. CNNs for Visualization Generation and Analysis

Inspired by the great success of CNNs on general computer vision and graphics tasks, there has been an increasing interest in applying them to data visualizations. The studies can be categorized into two groups: visualization generation and visualization analysis.

In visualization generation, CNNs are employed to ease the creation of visualizations and encode additional information into visualizations [12], [65]. VizML [30] utilizes a three-layer CNN to automatically recommend a suitable visual encoding of standard charts, including axis specifications and chart types. Chen et al. [3] leverage the ResNeXt architecture [60] to decompose timeline infographics and further construct infographics with a similar design style for new input data.

Prior studies on visualization analysis focus on training CNNs to understand data visualizations. A representative task is chart Question Answering (QA) [34], [46]. Since simple CNNs have difficulties solving the visual relation reasoning problem, Kafle et al. [33] extend Stacked Attention Networks [62] with dynamic encoding to enhance the network's understanding of bar charts. VisQA [36] adopts a template-based approach to generate explanations along with automatic question answering. CNNs have also been applied for various other visualization analysis tasks, such as visualization retrieval [50], outlier identification [17], [18], design pattern analysis [63], and chart ensemble exploration [8], [66].

All the above techniques intrinsically rely on the graphical perception capability of CNNs. In this work, we systematically evaluate how different CNN architectures perform in graphical perception tasks while providing implications for training and designing more powerful CNN models for visualization tasks.

C. CNN Generalization for Visualizations

CNNs have shown remarkable generalization abilities on data in which training and test data have similar distributions, outperforming humans in various tasks like image classification and object recognition. Despite these successes, their generalization performance degrades significantly when faced with out-of-distribution (OOD) test data [14], [15], [27], [28].

Only a few studies have been conducted regarding the OOD performance of CNNs for visualization techniques with small perturbations in the input. Lopes and Brodlie [42] evaluated the robustness of isosurfaces by observing how the continuity varies with respect to changes in the data or isosurface level. Correll et al. [7] show that a small perturbation of visual parameters might obscure the important data patterns. Wang et al. [58] examined the robustness of scagnostic measures by conducting a set of empirical studies. Yang et al. [61] revealed that VGG19 excels in predicting human correlation judgments on scatterplots and exhibits generalization ability across different designs.

In this work, we follow the above approach to evaluate the generalization performance of CNNs and humans on relational reasoning tasks involving bar charts. To the best of our knowledge, we construct the first generalization benchmark dataset for graphical perception tasks, which consists of a large set of synthesized standard visualizations and their element-wise perturbed counterparts with exposed and programmable parameters. We analyze the effect of perturbation operations on CNNs and compare the results with human judgments under the same conditions.

III. BACKGROUND

In this section, we briefly review two fundamental experiments that evaluate the graphical perception of humans and CNNs. We then revisit their hypotheses, experimental setups, and derived conclusions and report three major limitations that motivated our work.

A. Cleveland and McGill's Experiments

To study how humans decode graphs, Cleveland and McGill [5] summarized ten elementary graphical encodings from various commonly-used plots and charts: *position along a common scale*, *positions along non-aligned scales*, *length*, *direction*, *angle*, *area*, *volume*, *curvature*, *shading*, and *color saturation*. To evaluate the effectiveness of these visual encodings, they conducted two experiments using bar and pie charts. Here, we focus on the *position-length* experiment, which estimates bar length ratios.

Position-length experiment. Subjects were presented with a set of grouped and stacked bar charts with five different types, where two bars or bar segments were marked with black dots

on each chart. The task was to estimate the ratio between the two marked bars, i.e., the percentage of the smaller value to the larger value. Subjects could make judgments based on the *position along a common scale* or the *length* of the bars.

Measurements. The perception accuracy was measured by the midmeans of log absolute errors (MLAE) for each experimental unit in two experiments:

$$\text{MLAE}(\hat{\theta}) = \frac{2}{n} \sum_{i=0.25 \cdot n+1}^{0.75 \cdot n} \log_2(|\hat{\theta}_i - \theta_i| + 0.125), \quad (1)$$

where $\hat{\theta} = \{\hat{\theta}_1, \dots, \hat{\theta}_n\}$ is a set of ordered estimations, $\hat{\theta}_i$ is a predicted percent given by a subject, and θ_i is the corresponding true percent.

Findings. Cleveland and McGill found that adjacent bars score the best, closely followed by the separated bars and horizontally aligned stacked bars, while unaligned stacked bars and vertically aligned bars are the worst. The quantitative results show that the position judgments were 1.4 to 2.5 times more accurate than length and 1.96 times more accurate than angle. The authors attributed the differences to different visual estimation strategies in humans, where aligned bars involve the judgment of positions along a common scale and unaligned bars involve length judgments.

B. Haehn et al.'s Evaluation of CNNs

To investigate if CNNs meet or outperform humans on graphical perception, Haehn et al. [20] conducted five perceptual experiments using four off-the-shelf neural networks: a multilayer perceptron (MLP), LeNet [39], VGG19 [54], and Xception [4], with increasing sophistication in the architectures. They reproduced Cleveland and McGill's *position-length* and *position-angle* experiments and added three new experiments: (i) Cleveland and McGill's *elementary perceptual tasks* to study if CNNs can extract quantities from basic visual marks, (ii) *bars and framed rectangles* experiment to test if visual cues can help CNNs perform more accurately, and (iii) *Weber's law point cloud* experiment to check whether CNNs and humans have a similar mechanism on perceiving differences. In this work, we mainly focus on the *position-length* experiment.

Measurements. In all experiments, Haehn et al. simplified the charts from Cleveland and McGill's study by removing the axes and text. They presented the CNNs with sketch-style black-and-white images and tested their predictions on charts with the same parameters. Then they compared the results with human performance data collected by themselves using the same MLAE metric as Cleveland and McGill.

Findings. They found that CNNs performed worse than humans on the position-length experiment and had similar accuracy across different bar chart designs. Overall, CNNs may exceed or perform similarly to humans for some experiments but cannot complete tasks such as the ones in the *position-length* experiment, which requires estimating ratios between the lengths of the bars of interest. Therefore, Haehn et al. concluded that "CNNs are not currently a good model for human graphical perception."

C. Study Limitations

We identified the following three limitations of these studies:

- L.1** Haehn et al. explored four network architectures but did not study the influence of hyper-parameters such as different optimization solvers and learning rates. They also did not investigate more advanced CNN model architectures such as ResNet [24]).
- L.2** The charts that were used in the above studies are oversimplified. For example, they do not contain many indispensable elements, e.g., axes, labels, or titles. The graphical perception performance on these charts may not reflect the actual capability of CNNs for standard visualizations, let alone the ones used in real applications.
- L.3** Although humans are known to generalize effectively on common visual relational reasoning tasks, it remains unclear how CNNs perform when there is a discrepancy between the test and training data, as Haehn et al. only evaluated the generalization ability of CNNs when the test visualizations had similar visual encodings as the training ones.

IV. REVISITING HAEHN ET AL.'S EXPERIMENT

In this section, we replicate Haehn et al.'s evaluation of graphical perception with CNNs to address limitation **L.1** and then, while still using the black-and-white chart images, we gradually expand the parameter space for data generation of perceptual tasks in four steps, resulting in an increased task complexity. We measure the performance of each step to see if CNNs can still accomplish the ratio estimation task.

A. Replicating Haehn et al.'s Experiment

Following Cleveland and McGill's task design and Haehn et al.'s experimental design, we reproduce the *position-length* experiments using different network architectures and hyper-parameters.

In Cleveland and McGill's *position-length* experiment, the authors used controlled parameters for the generation of the stimuli. The *length values* involved in subjects' estimations (the marked target bars) were set as 10 real numbers equally spaced on the log scale:

$$s_i = 10^{1+(i-1)/12}, \quad i = 1, \dots, 10. \quad (2)$$

They also fixed the *indices* of the target bars, for example, the second and third ones in the left group of the bar charts. Furthermore, the *pixel dots* (a one-pixel dot for each bar) were all located in a fixed position inside the target bars.

Haehn et al. [20] used the same parameters when conducting the experiments with CNNs. In addition, they randomly sampled length values of non-target bars from an interval [10, 93], guaranteeing that all the local patterns in the image could be detected by the largest convolutional filter. They synthesized 100,000 100 × 100 sized images containing sketch-styled bars in black and white. To avoid that the CNN models simply memorize the data, they added 5% normally-distributed noise with mean = 0 and variance = 1 to each image.

TABLE I
PERFORMANCE OF VGG19 AND RESNET152 WITH DIFFERENT HYPER-PARAMETERS IN TYPE 1 OF THE POSITION-LENGTH EXPERIMENT, WHERE THE CELL WITH \times INDICATES THAT THE CORRESPONDING PARAMETER IS INVALID, AND CI INDICATES CONFIDENCE INTERVAL. NOTE THAT THE HUMAN ESTIMATED MLAЕ VALUE IN THIS TASK REPORTED BY PREVIOUS STUDIES [20] IS 1.4.

Network	Optimizer	Learning rate	Momentum	Weight decay	MLAE	CI
VGG19	SGDM	High	Classic	\times	1.43	0.05
		Nesterov	\times		1.10	0.22
		Low	Classic	\times	1.41	0.15
	AdamW	Nesterov	\times		1.04	0.27
		High	\times	High	-2.32	0.11
			\times	Low	-2.41	0.10
ResNet152	SGDM	High	Classic	\times	-0.61	0.24
		Nesterov	\times		-0.60	0.21
		Low	Classic	\times	0.72	0.24
	AdamW	Nesterov	\times		0.70	0.25
		High	\times	High	-2.76	0.10
			\times	Low	-2.42	0.17
		Low	\times	High	-2.71	0.10
			\times	Low	-2.53	0.24

CNN Architectures. The network architecture is crucial for the performance of CNNs in various applications. Haehn et al. used VGG [54], a simple and effective CNN architecture that supports up to 19 layers. Since their result shows that VGG19 achieves the best graphical perception performance among other networks, we use VGG19 as one of the baselines in our experiments. To investigate if a better performance can be achieved by more recent CNNs, we include a ResNet152 [24] in our experiments, which is 8 times larger than VGG19 and deeper than the Xception126 [4] network also used by [20]. We also test multi-layer perceptron (MLP), AlexNet [38], LeNet [39], DenseNet [31], EfficientNet [56], and a Relation Network [51].

Optimization Solvers. Optimizers also play an important role in the training process of CNNs and can have a big influence on their performance. Stochastic gradient descent (SGD) is a classic iterative method for optimizing deep neural networks. Haehn et al. use SGD with momentum (SGDM) [48] to improve the classic SGD with the advantages of faster convergence and fewer oscillations during training. More recently, there are more powerful and efficient optimization solvers for training deep neural networks. Adaptive moment estimation (Adam) [37] is a representative example, which achieves better performance than SGDM. Loshchilov and Hutter [43] proposed an improved version of Adam called AdamW, which shows better training curves and generalization ability. In this work, we explore the different performances of SGDM and AdamW.

Training Objective. Given the objective of predicting continuous bar length ratios, we train regression models as in previous work [20]. Although MLAЕ (Equation 1) is suitable for perception accuracy, using it as a log-scale loss during network

training leads to biased gradients for different prediction error scales. Thus, similar to Haehn et al. [20], we train our models using the mean squared error (MSE):

$$\text{MSE}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2.$$

Following the suggestion of Haehn et al., we normalize both $\hat{\theta}_i$ and θ_i to the range $[0, 1]$. When MLAE is -3, MSE is around 0; when MLAE is 0, MSE is around 0.76. A negative MLAE score of around -3 indicates high estimation accuracy, and a high positive score indicates poor accuracy.

Hyper-Parameters Tuning. Another issue for achieving optimal model performance is the selection of hyper-parameters during CNN training, such as learning rate, momentum type, weight decay, batch size, and epoch number. We follow the suggestions by recent guidelines [10], [19], [29] for tuning hyperparameters. Table I shows our chosen values for *learning rate* and a parameter for each of the two main *optimizers*: *momentum type* for SGDM and *weight decay* for AdamW. To find proper values for learning rate and weight decay, we adopt a coarse-to-fine strategy that first searches for coarse ranges of hyperparameter values using a small number (100) of epochs and a small training set (1/3 of the entire dataset) and gradually narrows the range down through binary search. To alleviate falling into a local minimum, we search for the optimal values from two different ranges: one with large values and the other using small values. Hence, we set two learning rates for each combination of network architectures and optimizers to inspect if higher or lower values contribute to better results.

For SGDM, we test two commonly used types of momentum, classic and Nesterov momentum. For AdamW, we examine the effect of weight decay and set two values as high and low, respectively. Based on these parameters, we follow the previous guidelines that train the networks for 500 epochs with a mini-batch size of 16 and 32 for each task and terminate early if the validation loss stays unchanged for ten epochs.

Since Haehn et al. [20] reported poor performance of CNN models that were pre-trained on ImageNet [38], we do not perform any pre-training and train the models from scratch. For all the experiments, we employed 5-fold cross-validation for each model. We split the dataset into five equal-sized subsets and used four subsets for training and one for validation iteratively. This process was repeated five times, ensuring each subset was selected as the validation set exactly once. Our models are implemented in Pytorch and trained on an NVIDIA RTX 3090 GPU with mini-batch size 16 for 500 epochs. All networks are initialized with random parameters.

Results. All models outperform VGG19 (MLAE=3.51) as reported by [20], even the simple MLP (MLAE=0.35). Given this unexpected result, we first reviewed the training process by comparing the training and validation loss curves [2] to rule out issues such as overfitting. The curves are provided in the supplemental material, where the validation loss consistently decreased alongside the training loss, confirming that the training process was conducted properly.

Across the eight tested network architectures, VGG19 and ResNet152 perform significantly better than other networks as shown in Fig. 1. To analyze the influence of hyper-parameters,

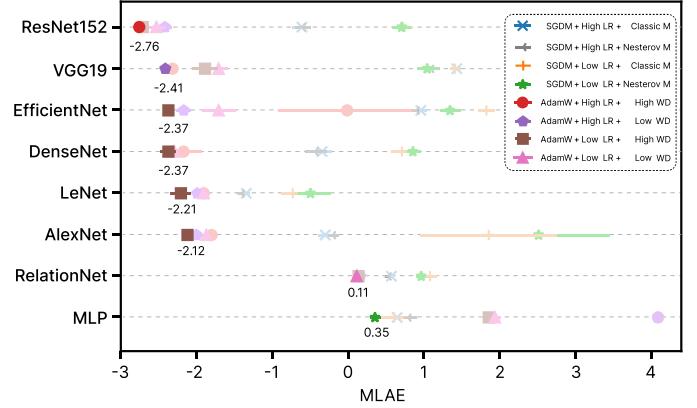


Fig. 1. Performance comparison of eight network architectures trained with eight sets of hyper-parameters in type 1 of the position-length experiment. The best trained model for each kind of network is highlighted and labeled with the corresponding MLAE value, while others are shaded.

we report the detailed results of these two networks for the position-length type-1 experiments in Table I, as similar patterns are observed in the remaining models. Comprehensive results are provided in the supplementary material. Haehn et al. [20] reported the lowest MLAE=3.51 achieved by VGG19 compared to humans (MLAE=1.4) for this task, which led to their conclusion that CNNs *cannot* complete this task. In contrast, we find that **VGG19 and ResNet152 trained with carefully selected hyper-parameters exceed human performance**, as shown in Table I.

Examining the influence of each factor, we find that AdamW enables VGG19 and ResNet152 to produce the best results, with a higher learning rate improving network performance in most cases. In contrast, the influence of momentum and weight decay are both unclear as the MLAE grows higher or lower for different conditions. The batch size has a weak influence on model training since smaller (16) and bigger (32) sizes result in similar MLAEs. Throughout all combinations of networks and hyper-parameters, ResNet152 trained with AdamW with a high learning rate and weight decay performs best for the *position-length* experiments. In particular, the MLAE value for the experiment is -2.76, which is highly accurate and much better than average human performance (1.4). Therefore, we perform further evaluations using this configuration of ResNet152 with AdamW.

B. Expanding the Data Generation Parameter Space

In this experiment, we use the same black-and-white chart images as in Section IV-A, but gradually expand the parameter space for data generation in three steps, resulting in increased task complexity. We measure the performance of each step.

Experimental Setup. As shown in Fig. 2, the appearance of the bars to be estimated is mainly decided by two factors: the number of possible bar heights, and indices of the target bars. Here, we investigate how CNNs perform when expanding the value range of each factor using the following steps:

- (i) We expand the range of target bar lengths from 10 to 20 fixed numbers by modifying Equation 2:

$$s'_i = 10^{1+(i-1)/24}, \quad i = 1, \dots, 20.$$

- (ii) We further expand the length range of all bars to random numbers within the interval [10, 93]. This range is chosen as in [20] to ensure that all content in a 100×100 image can be seen by large convolutional filters. This also allows for testing on out-of-distribution lengths in the future (see Section V-B).
- (iii) Instead of fixing the target bars (indicated by black dots) to the second and third ones, we define them as two arbitrary adjacent bars ($j, j+1$), where j ranges from 1 to 9.

The dataset generated (consisting of 60,000 samples for training, 20,000 for validation, and 20,000 for testing) in each step is fed into the network for model training and evaluation. Later, we test the model with unseen stimuli. We use ResNet152 with the AdamW optimizer and keep all the other hyper-parameter settings the same as in Section IV-A.

Results. The test MLAEs of the CNN models on our expanded parameter space are shown in Fig. 2(d). A reasonable hypothesis is that the model performance will drop as the datasets and tasks become more complicated. This is confirmed by the general trend of our experimental results. Despite a performance decrease, the MLAE value of our CNN models is -2.30, already reaching an extremely high estimation accuracy that is significantly better than that of humans (average MLAE=1.4). Hence, **CNNs perform well in the expanded data generation parameter space and achieve comparable performance as in the previously highly controlled settings.** However, even in such a setting, the stimuli are still far simpler with no essential graphical elements adjunct to the visualizations like axes and texts or using color.

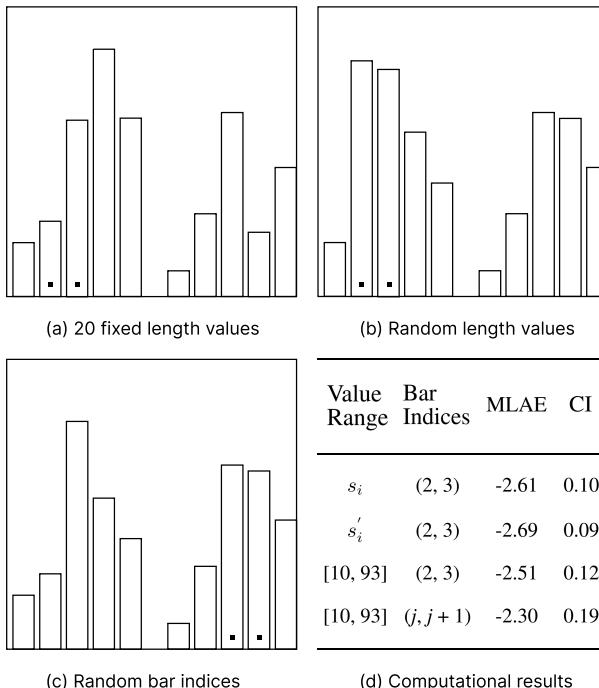


Fig. 2. Sample images of expanding the parametric space in three steps: (a) from 10 fixed length values to 20 fixed values, (b) random length values, and (c) from fixed indices of target bars to random indices; (d) control settings and MLAE values and CIs of three extended steps in (a-c)

V. GENERALIZATION OF CNNS AND HUMANS ON STANDARD VISUALIZATIONS

In this section, we explore how well CNNs and humans perform on standard visualizations. To address L.2 and L.3, we first create **GRAPE** (GRAphical PERception tasks), a synthetic dataset that contains 766,000 (500,000 for training and 266,000 for testing) images of standard visualizations generated by Vega-Lite [52]. Next, we train a CNN model for each task and test how well they perform on unseen visualizations. We also conduct a user study to measure human performance.

Since Haehn et al. [20] only evaluate CNNs with test data that has similar visual encodings as the training data, their results do not allow to reveal how well CNNs handle charts with slight variations in visual encodings (L.3). Hence, we create two kinds of test stimuli based on GRAPE: one sharing similar visual encodings with the training data, and the other with variations in the visual encodings. We refer to these two settings as independent and identically distributed (IID) and out-of-distribution (OOD) test data, respectively [13], [15].

A. GRAPE Dataset

In contrast to the simple black-and-white charts in the previous studies [5], [20], the charts used in our next set of experiments are standard visualizations that were generated by Vega-Lite [52]. Here, we describe the construction of our dataset of standard visualizations, **GRAPE**, and the methodology we followed to create training and test stimuli for CNNs.

Parametric Space. Vega-Lite uses the grammar of graphics [59] with a set of parameters that control the visual appearances of independent chart elements. We programmatically vary commonly used parameters for bar charts, such as bar width, stroke width, title properties (position, font, and size), and colors (background color, bar color, and stroke color). Our parametric model also supports other parameters like tick number, axis label, center position, etc.

Since color is an essential component of standard visualizations, we design a color mapping scheme that satisfies two requirements: adjacent colors have visually noticeable differences, and all colors are nameable to facilitate communication. To do so, we take the 140 named colors from D3 [1], order them by their perceived luminance, and divide them into four non-overlapping color sets. We assign the brightest (set 1) and darkest (set 4) colors to the background and stroke, respectively. Set 3 is used for the bar fill color, while set 2 is used to generate completely unseen color distributions for testing our CNN models. To ensure sufficient color discriminability, we make sure that the luminance difference between the colors of any pair of visual elements (e.g., bars, strokes, and background) in one chart is at least 5 in CIE LAB color space.

With the above parameter settings, we generate a set of chart images of size 150×150 pixels. The reason that we increased the image size, compared to the 100×100 pixels used by Haehn et al. [20], is to provide enough space to place the title, axis, and labels. Fig. 3 shows the resulting five types of bar charts for the *position-length* experiments. Like Haehn

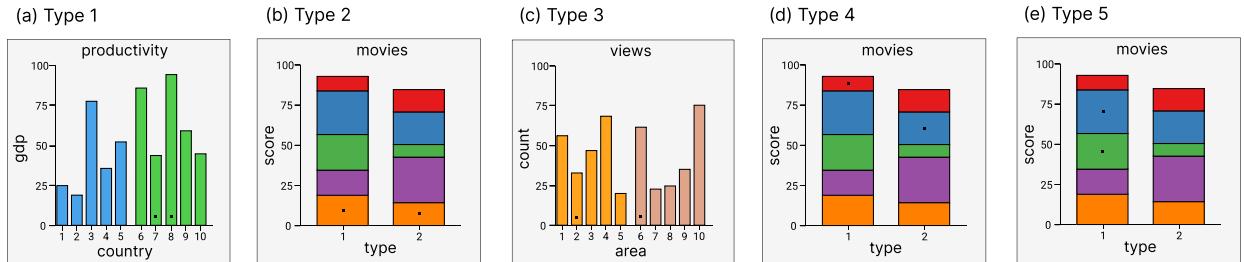


Fig. 3. Our five types of stimuli used in the *position-length* experiment, where each bar chart includes colorized bars, axes, tick labels, and titles.

et al., we add 5% random noise (normally distributed between -0.025 to 0.025) to each pixel to prevent the networks from memorizing each image.

Training Stimuli. We increase the task complexity of the *position-length* experiment setting of Haehn et al. [20] by adding two aspects. (i) One extension is to expand the two parameters length range and indices of target bars defined in Section IV-B. As shown in the last row of Fig. 2(d), the value range of each bar is extended to [10, 93]. For stacked bar charts, we further normalize the total values within each stack. For the indices of the target bars, we follow Cleveland and McGill, who used any two adjacent bars for types 1 and 5, random bars from different groups/stacks for types 3 and 4, and the bottom two bar segments for type 2 (Fig. 3). (ii) The other extension is to sample the ten bar values from more data distributions with different patterns (e.g., long tail, peak, and sine wave) in addition to the random distribution. All stimuli are colored instead of the previous over-simplified black-and-white ones. For each task, we generate 60,000 images for training, 20,000 for validation, and 20,000 for test.

Test Stimuli. Our test dataset comprises two types of stimuli procedurally generated for IID and OOD test respectively. For the IID test, the charts are generated using the same parameters as in the training stimuli. To investigate how CNNs perform on unseen visualizations, especially ones with variations in parametric space, we generate new test stimuli for the OOD test by perturbing the visual encoding parameters. We choose to manipulate 9 parameters that can be divided into three categories: global chart properties (title position, title font size, and background color), *task-unrelated* mark properties (bar width, bar color, stroke width, and stroke color), and *task-related* mark properties (bar length and dot position). Fig. 4 depicts one exemplary perturbation for each of the nine parameters of the type 1 bar chart.

For each parameter, except bar length, we gradually increase and decrease the corresponding value in the training stimuli, such as luminance for color-related parameters. For example, chart titles are center-aligned in the training set, whereas in the test stimuli they are gradually moved to the left or right by 15% of the canvas width each time. Since the color sets assigned to the background and stroke consist of the brightest and darkest colors, respectively, we only decrease the luminance of the background colors and increase the luminance of the stroke colors. If the perturbed value of one parameter ends up out of the valid range, we discard the corresponding stimuli. The bar length is determined by the

encoded data value. Since the range of data values in the training stimuli is [10, 93], we generate the bar lengths for the test stimuli in the value range [1, 9] \cup [94, 100].

By default, the perturbations of each parameter have six levels, as shown on the x-axis of Fig. 5. For example, the title position is moved by 15% of the canvas width three times from the center to both the left and right direction, respectively. To ensure that all marks are within the canvas and do not overlap with each other, we only set four levels of perturbation for bar width on types 1 and 3, and two levels for stroke width on all types of bar charts. As for dot position, we only move it along the horizontal direction for stacked bar charts (type 2,4,5) and the vertical direction for grouped bar charts (type 1,3). For each level of the perturbation, we use the same method described earlier in this section to generate 1,000 chart images. Taking all types and all levels of perturbed charts together results in a total number of 266,000 test stimuli.

B. Generalization Performance of CNNs

With our new GRAPE dataset, we re-train the different model configurations described in Section IV for each type of chart to test the IID generalization on standard visualizations. Then, we apply the trained CNN models to each set of perturbed test visualizations to benchmark their OOD generalization. Fig. 5 shows the results.

IID Test. The MLAE values corresponding to the zero perturbation value on the x-axis of Fig. 5(a-h) and the [10, 93] interval on the value range column of Fig. 5(i) are the average MLAE values of CNN estimations on standard visualizations with our expanded settings. Even though more complicated datasets are fed into the CNNs, our results are significantly better than Haehn et al.'s [20]. The test MLAE values for the five types of bar charts are all between -2.52 and -1.91, whereas even the best performance reported by Haehn et al. for the same tasks is above 3.

Compared with the results in Section IV, where CNNs were trained with black-and-white images under highly controlled settings, the overall MLAEs on standard visualizations only dropped less than .5 for all types of tasks. These average MLAE values (all negative) are also significantly lower than those of humans reported by Haehn et al. [20], which are 1.4 on the over-simplified stimuli. In summary, our results show that **CNNs can achieve extremely high IID generalization performance on graphical perception tasks.** They produce unbiased estimations for all types of bar charts when both training and test data satisfy the IID property.

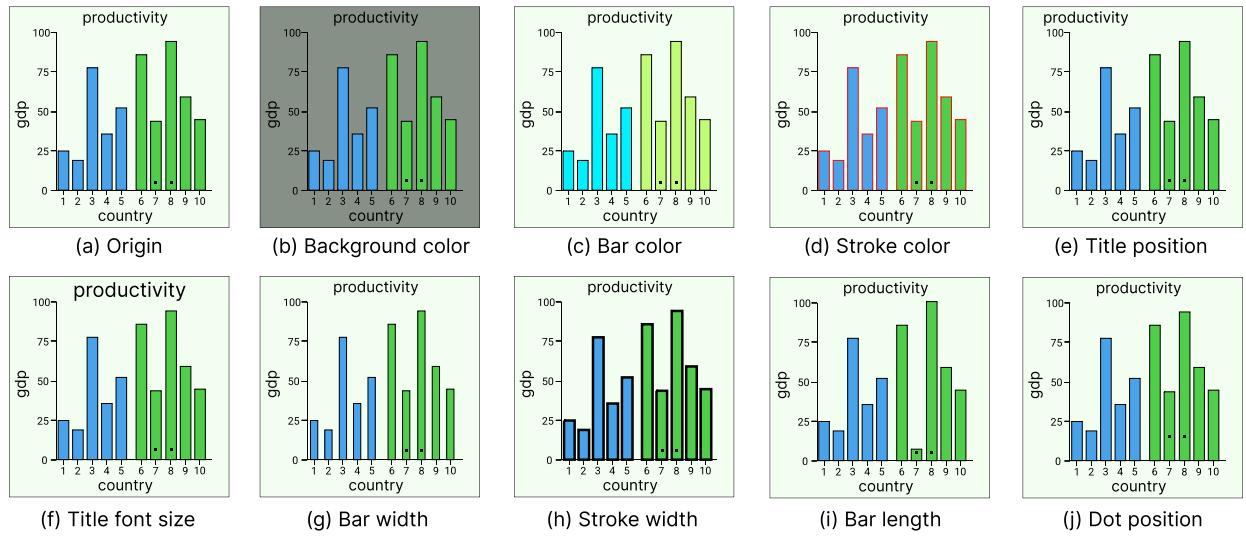


Fig. 4. An exemplar training stimulus (a) of type-1 and the test stimulus (b-j) generated by perturbing eight different parameters of one bar chart of type-1 at one specific level.

OOD Test. The curves in Fig. 5 show how MLAE changes as a function of the perturbation levels of each parameter. We see that MLAE values are almost the same for different perturbations of the title position (Fig. 5 (a)) and title font size (Fig. 5 (b)), except for type 2 and 5 on the -45% perturbation of title position. For the three color parameters (Fig. 5 (c-e)), CNNs are least robust to the background color as the MLAE values of CNNs' estimation on all types of charts change to around 4.0 when the luminance value decreases by 15. This might be reasonable since the background color affects a larger area and can significantly change the appearance of the whole chart compared to the colors of other visual marks.

While the MLAE values do not change much for minor perturbations of color and title parameters, we find that a small positive or negative change to some parameters like bar width, stroke width, and dot position cause significant changes, forming the symmetric curves in Fig. 5(f-h). For example, increasing or decreasing just one pixel of stroke width results in a change from -2.5 to 4.0 of the MLAE value on type-2 bar charts. As for the bar length encoded by the value ranges $[1, 9] \cup [94, 100]$, the MLAE values of types 1 and 3 are even larger than 4.0, whereas for types 2 and 4, the values are less than 2.0. These results indicate that the CNNs might not have modeled the ratio estimation task correctly, as many task-unrelated visual properties heavily impact its relational reasoning.

Among these five types of bar charts, only type 2 is robust to the perturbation of dot position, which is used to indicate the target bars (Fig. 5(h)). We speculate that the reason for this phenomenon is that the target bars of type 2 are restricted to the only two ones on the bottom of the chart, i.e., the indices are actually fixed, which makes it easier for a CNN to model the task. In contrast, the MLAE values of type 2 and 5 all show significant changes when the title is moved to the leftmost position (Fig. 5(a)). This is unexpected since the title position is unrelated to the target bars. For the stroke color and bar color, types 2, 4, and 5 are less robust to large perturbations than the other types. We speculate that the relatively large color

complexity (i.e., more colors are placed closely) in the three stacked bar charts makes CNNs more sensitive to variations of stroke and bar colors.

In summary, we report two findings for OOD generalization: (i) **the relational reasoning ability of CNNs is heavily influenced by most visual parameters, regardless of whether they are related or unrelated to the target visual marks;** and (ii) **the generalization performances of trained CNN models are almost equal on different types of bar charts in the IID setting, but vary significantly in the OOD setting.** Combining these observations, we draw the conclusion that despite their outstanding performance on ordinary IID tests, our used CNNs have not truly learned how to solve the ratio estimation task correctly.

C. Comparison with Human Performance

To further check whether the generalization performance of CNNs aligns with humans, we conduct a user study to perform the same tasks with human subjects in the IID and OOD settings as described in Section V-B and then compare their results. For obtaining high-quality human data, we perform the ratio estimation task in a well-controlled lab environment with sufficient control over the physical workspace, display calibration, and observer attention.

Measure. Before describing our study, we first define *MLAE deviation* to quantitatively measure the influence of different perturbations on CNNs and humans:

$$\Delta\text{MLAE}(\hat{\theta}) = \text{MLAE}(\hat{\theta}_{ood}) - \text{MLAE}(\hat{\theta}_{iid}), \quad (3)$$

where $\hat{\theta}_{iid}$ is an estimated value of the chart without perturbation, $\hat{\theta}_{ood}$ is the corresponding predicted value of the same target after a perturbing operation is performed. A larger deviation means the perturbation has a greater influence.

Experiment Design. In this experiment, each participant is asked to estimate length ratios of bar charts generated by applying different levels of perturbations on eight parameters.

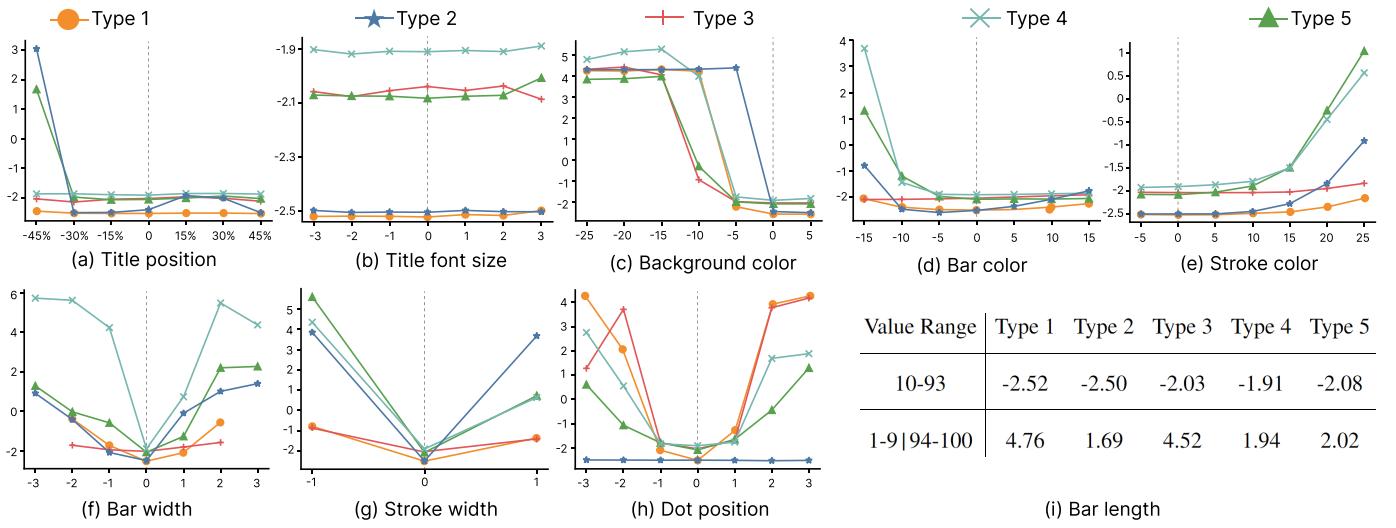


Fig. 5. The mean MAE values produced by CNNs on performing generalization tests of eight parameters on five types of bar charts. (a-h) Each curve shows how the MAE values change by increasing or decreasing the corresponding parameter values. The dotted lines indicate the MAE value computed for the stimuli with non-perturbed parameters. (i) Estimated MAE values for bar charts with bar lengths encoded by different value ranges.

The purpose is to examine whether these perturbations affect human perception in a similar way as they do for CNNs. To this end, there are two kinds of conditions: parameters and perturbation levels. Note that we do not include the title font size, because its perturbation almost brings no change to the MAE values of CNNs (Fig. 5(b)).

However, the number of all stimuli used for evaluating the generalization performance of the CNNs is too large for a human subject study. For example, in our automatic CNN tests, the background color has 7 levels. If we select 5 images for each level of all five types, we end up with $5 \times 7 \times 5 = 175$ bar charts, which would be excessively demanding for human subjects if presented directly.

To address this issue, we introduce a *two-phase* experimental scheme based on the previous finding that small perturbations do not affect human judgment significantly [15]. Generally, we first identify the visual parameters that have the greatest impact on human judgments and then evaluate how humans respond to different levels of perturbations applied to the identified parameters. If human subjects consistently perform well even as the most influential parameters are altered, we can infer that humans are robust to variations in visual parameters in relational reasoning tasks on bar charts.

Specifically, in the first phase, we evaluate human performance on five types of bar charts with the largest perturbations to each of the visual parameters in Fig. 4. Here, we have eight conditions, where each condition corresponds to one visual parameter. To further reduce the number of trials, we ask each participant to complete the tasks of all eight parameters for one type of chart and recruit the same number of different participants for each chart type. Then, we identify the top three (type, parameter) pairs that affect human performance the most, and conduct progressive perturbing experiments on these pairs as in Section V-A in the second phase. Since our goal is to investigate how different perturbation levels of the three most influential parameters affect human estimation accuracy, we ask each participant to complete the trials generated by all

perturbation levels of all three pairs. Thus, the first phase of our experiment is a *between-subject* design where participants are divided into five groups and each group completes the trials of one type of bar charts, while the second phase is a *within-subject* design to evaluate the effect of different perturbation operations.

Device. The study was run on a quad-core PC with a 23-inch color-calibrated screen monitor (120 Hz refresh rate) using a mouse as input and a 1920×1080 pixel display. All participants were seated with their eyes approximately 60 cm from the display in a quiet room.

Hypotheses. As we saw, CNNs perform significantly better than humans for the over-simplified charts in Section IV. Other studies have shown, however, that they are less robust than humans to image manipulations [15]. Thus, we formulate the following two hypotheses:

H1: Human performance is inferior to CNNs in terms of estimation accuracy on standard visualizations; and

H2: Human performance is more robust to perturbations of various visual parameters than that of CNNs.

Participants. We recruited participants with color-normal vision from our local universities and communities. Since bar charts are common visualizations, we follow the conclusion from Hall et al. [22] that task performance would not have significant differences across professions and set no condition on the occupation of participants. We had 20 participants to complete the corresponding tasks for each of the five types of bar charts in the first phase, and 20 participants for all three (type, parameter) pairs in the second phase. Note that none of the participants acted repeatedly either in different phases or groups. Thus, we have $20 \times 5 + 20 = 120$ participants in total (ages between 18 and 45, 50 females).

Procedure. We applied the following procedure in the lab study: (i) explaining the tasks by the researcher, followed by training; (ii) performing the study; and (iii) a short interview. Three training trials were provided to help participants get familiar with the task and our user study system. For each type

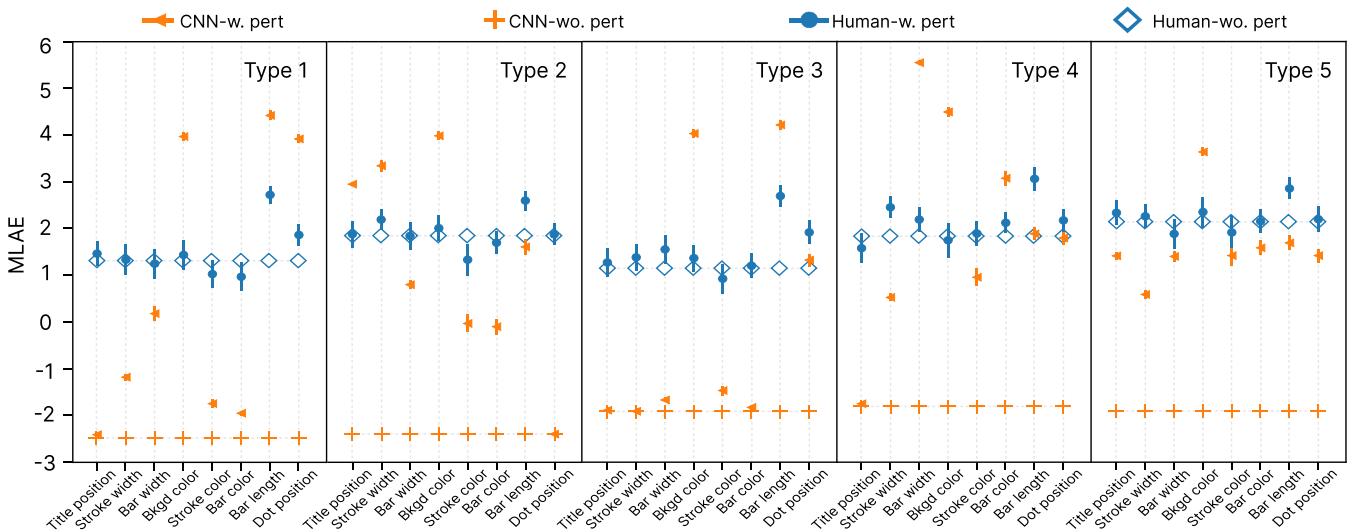


Fig. 6. Comparison of bar chart tasks produced by humans and CNNs: mean MAE and 95% confidence interval values produced by humans and CNNs on five types of bar charts without and with the largest level perturbations on eight parameters.

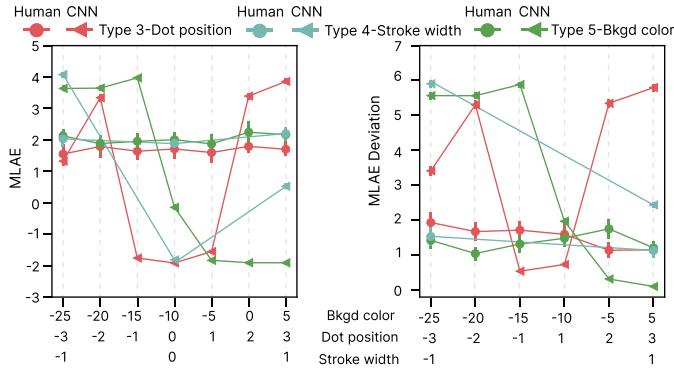


Fig. 7. Comparison of bar chart tasks produced by humans and CNNs: The curves show how the mean MAE values (a) and mean MAE deviations (b) vary along the perturbation levels on the eight parameters.

of bar chart, we selected 5 random images without perturbation and then generated 5 images for each one of the 8 visual parameters by perturbing the corresponding parameter of those selected images, resulting in $5 + 5 \times 8 = 45$ stimuli shown to each participant in the first phase. In the second phase, each participant needed to complete 5 trials for each level of perturbations to three visual parameters identified as most influential in the previous phase, i.e., $5 + 5 \times (6 + 6 + 2) = 75$ stimuli. In the interviews, we asked participants how they performed the task and which factors influenced their estimation through questionnaire. Overall, it took about 15 minutes for each participant to finish the whole study, including 10 minutes for completing the task and 5 minutes for interview.

Results Analysis. We first conducted a preliminary analysis of our collected data. We used the D'Agostino-Pearson test to quantitatively confirm the normality of log errors' distribution. The results show that the log error distributions of human and CNN results are not normal distributions. Thus, following prior studies [5], [20], we compute the bootstrap distribution of the means of log errors using the 95% confidence interval. Note that the results from the between-subject study are treated as

independent groups and do not support cross-group analysis.

Fig. 6 summarizes the results of humans and CNNs for five bar chart tasks without and with the largest perturbations, respectively. We see that the MAE values of humans are significantly larger than the ones of CNNs for all parameters without perturbation ($p < 0.01$). Yet, the MAE values of humans are smaller than the ones of CNNs for some parameters (e.g., stroke color) with perturbations in the OOD setting. We make similar observations in Fig. 7(a). Thus, we partially accept H1. Fig. 7(b) present the summarized average MAE deviation of CNNs and humans on all bar charts. We see that **humans are more robust than CNNs with smaller MAE deviation for all tested levels of the most influential parameters**. Thus, we accept H2.

By using the Student's t-test to detect significant differences among parameters, we found that the human MAE deviation shows significant differences for the bar length compared to the other parameters ($p < 0.05$), whereas the MAE deviation of the CNNs shows significant differences between all pairs of parameters ($p < 0.01$). These statistics indicate that humans and CNNs might use different reasoning processes. Thus, we speculate that **humans are mainly influenced by the bar length but CNNs are influenced (and disturbed) by many other factors**.

This speculation is also supported by the interview results. Most participants mentioned that they did not notice the perturbations of parameters. They said it was hard to estimate the ratio when the difference between two bar lengths was too large or too small. These observations are consistent with the results produced by humans on the bar charts encoded within the value range [10, 93] and the ones in the range [1, 9] \cup [94, 100], where the latter ones result in significantly larger MAE values.

VI. UNDERSTANDING AND IMPROVING CNNS

Our user study results show that humans tend to estimate bar length ratios based on specific parameters of the target

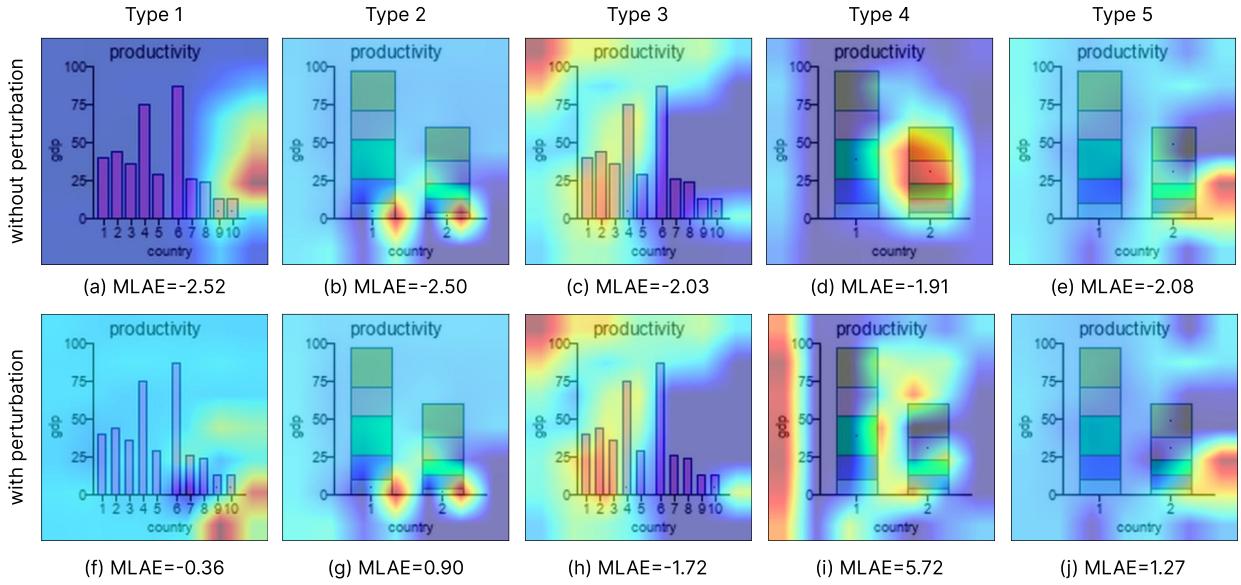


Fig. 8. The Grad-CAM maps show the important regions in the chart image for estimating length ratios across five bar chart types. The two rows correspond to the charts without perturbations (first row) and with minor perturbations of the bar width (second row).

bars while ignoring other task-irrelevant aspects of the chart. In contrast, CNNs inference involves many visual parameters, which may explain why their performance is not as robust as humans. To visually explore the inference of a CNN model, we use the gradient weighted class activation mapping (Grad-CAM) [53] to localize the important regions of the chart image during inference.

Grad-CAM Map. Since the last convolutional layers of the CNN model are expected to retain the task-relevant spatial information, we visualize their corresponding Grad-CAM maps. As shown in Fig. 8, the two rows of Grad-CAM maps correspond to the bar charts without and with the perturbation of bar width, respectively. We see that only type 2 focuses on the target bars (Fig. 8(b,g)) whose indices are fixed. We speculate this is caused by the sufficient capacity of the model in memorizing the indices information of the training data. In contrast, almost all regions of interest in the other types are unrelated to the task. For example, the focus region in type 1 is unrelated to all bars (Fig. 8(a,f)). Comparing the maps without and with perturbations, we see that even a small amount of perturbations on some visual parameters can dramatically change the resulting Grad-CAM maps.

Quantitative Analysis. We further assess the target region localization ability of CNNs by measuring the Intersection over Union (IoU) between the area of target bars T and the high-intensity region of the Grad-CAM map G as follows:

$$\text{IoU} = \frac{|T \cap G|}{|T \cup G|}. \quad (4)$$

We programmatically identify the segmented area of target bars (shown in Fig. 9(a)), and extract the high-intensity regions of the Grad-CAM map (illustrated in Fig. 9(b)) by empirically applying a threshold of 200. We compute the IoU scores for all type-1 stimuli in both IID and OOD tests. The results, visualized as dashed lines in Fig. 9(c), show that mean IoU scores across all conditions are close to zero, indicating

minimal overlap between the two areas. This suggests that **CNNs might be using regions other than the target bars for relational reasoning on bar charts assuming Grad-CAM accurately reveals important features.** However, there is no consensus on the effectiveness of different explainable AI methods [32] (see results in the supplemental material), even though Grad-CAM remains a popular technique.

Data Augmentation. To learn how well a CNN model trained with perturbed bar charts deals with other perturbations, we performed data augmentation using different combinations of the perturbations and then trained the model from scratch. When training a CNN model on two types of parameter perturbations (e.g., bar length and bar width), each type of perturbation was drawn uniformly. Taking the bar charts with the perturbation of bar length and bar width as an example, the bar length in each chart was encoded by the data range [1,100] and the bar width was a random value of the set {4,5,6,7,8}. Fig. 10 visualizes the testing result of the model on charts with the same or different perturbations.

We find that while all models are robust to the perturbation of title positions and those they were explicitly trained on, they all exhibit poor generalization performance on the perturbations that were not included in their training sets, especially the ones of the dot position. Notably, dot position and bar width seem to be the most influential parameters. Given the endless range of possible perturbations, it is impractical to train on every variation. Thus, we conclude that **data augmentation cannot close the human-CNN generalization performance gap in the task of relational reasoning with bar charts.**

Segmentation Mask. In the second attempt to improve CNN performance, we add a segmentation mask of the target bars as an additional alpha channel to each chart image. The segmentation mask is a binary mask of equal size to the original image, where pixels in the segmented foreground are all white and the background is all black. We render

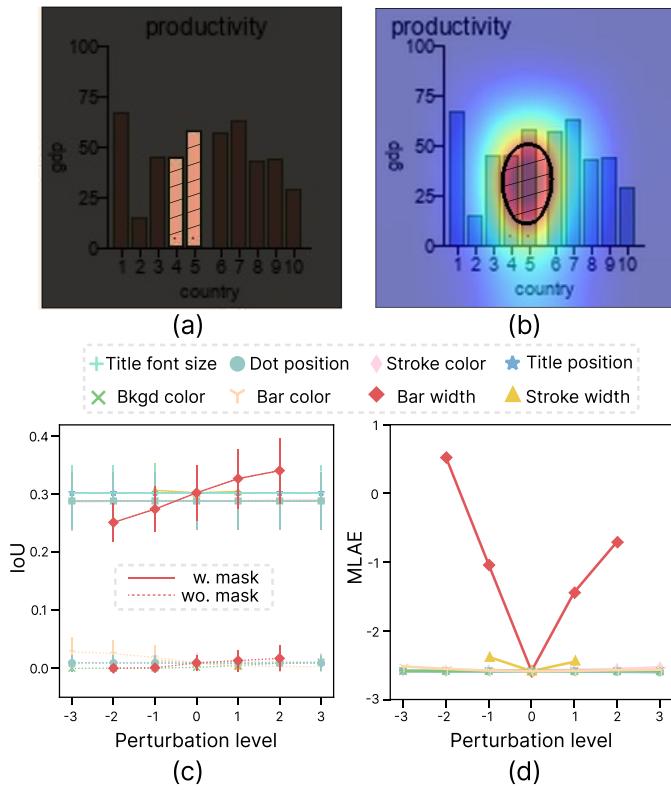


Fig. 9. Improving the generalization performance of CNNs by providing segmentation masks. (a) The blended image of a chart and its segmentation mask, in which the area of target bars is shaded; (b) A perturbed chart blended with its Grad-CAM map, where the high-intensity region is shaded; And curves showing how IoU scores (c) and MLAE values (d) change over different levels of perturbations of one of eight parameters.

these segmentation masks for all bar charts in the GRAPE dataset. After training a CNN model with the corresponding RGB- α images, we tested the target region localization by Equation 4 and the generalization performance of the model as in Section V-B, the results are shown in Fig. 9(c,d). We see that (i) CNNs can better localize the target bars with the provided segmentation masks, as evidenced by the significant increase of the IoU scores, and (ii) CNNs' robustness has improved against most perturbations in visual encodings of bar charts such as title position, background color, and bar color.

However, this mask-enhanced CNN model is still sensitive to bar width and stroke width as shown in Fig. 9(d). This suggests that **while segmentation masks may address the generalization issues related to texture/color in bar charts, generalizing to shape variations remains a difficult task for CNNs.** This observation is consistent with the findings reported in [14] that CNNs rely heavily on texture information when performing object recognition tasks and that models trained with stronger shape bias can exhibit greater robustness. Moreover, the resulting mean MLAE value is 2.28 for the bar lengths encoded in the range $[1, 9] \cup [94, 100]$, indicating that this model does not only use bar lengths to estimate ratios. In other words, even though the CNNs are provided with a segmentation mask, they cannot estimate length ratios as well as humans in certain cases.

Because the segmentation masks are based on the complete and accurate segmentation of the target bars, removing the

pixel dot in the RGB image (see the dots in Fig. 9(a)) does not degrade the model's performance. Yet, the CNNs cannot complete the task when the mask is incomplete. For example, if we only denote the pixel dot in the alpha map, the training process does not converge at all. Even replacing the pixel dot with a small Gaussian blob results in an MLAE value of 4.32.

VII. DISCUSSION

Evaluating Vision Models with Grammar of Graphics. For over a decade, the computer vision community has been using ImageNet [9], a large-scale dataset that contains 3.2 million natural images, as the benchmark dataset for training and evaluating the generalization performance of vision models. Compared with natural images, the design space of data visualization contains more complex and hierarchical semantic information. Each visual element in visualizations has its own semantic meaning that contributes to a specific visual expression. For example, even the smallest components of a diagram, such as tick marks, are important indications of the correspondence between numbers and the coordinate axis. This makes it harder for neural networks to learn appropriate features and relations between different visual marks. Synthesizing large-scale visualization datasets using the grammar of graphics, such as GRAPE¹, will allow the research community to conduct more systematic evaluations of graphical perception of machines and humans.

Network Architecture and Task-oriented Attention. Our experiments show that title position affects the robustness of CNNs on certain bar charts, even though it is irrelevant to graphical perception tasks. In contrast, humans do not have this issue due to our robust attention mechanism. The Transformers architecture [57] uses self-attention mechanism that has proven highly effective in [11], [41] and may be robust to the perturbations on certain parameters. Since visualizations are often task-dependent, different tasks have different relevant visual encodings. For example, reading values from a bar chart requires understanding the axis, while estimating the ratio between bars requires comparing bar lengths. Hence, we believe that a task-oriented attention mechanism that can understand and reason about task-related visual encodings will enhance performance in graphical perception and computational visualization analysis.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a comprehensive generalization analysis for CNNs and humans on graphical perception tasks with bar charts. We find that appropriately trained CNNs can outperform humans on ratio estimation tasks. Based on the carefully selected configuration (ResNet152 and AdamW solver), we show that CNNs also outperform humans on graphical perception tasks using standard visualizations specified by Vega-Lite where the visual parameter space is greatly expanded compared to previous studies [5], [20].

¹We put GRAPE and all source code into the supplemental material and will open-source upon acceptance of this paper.

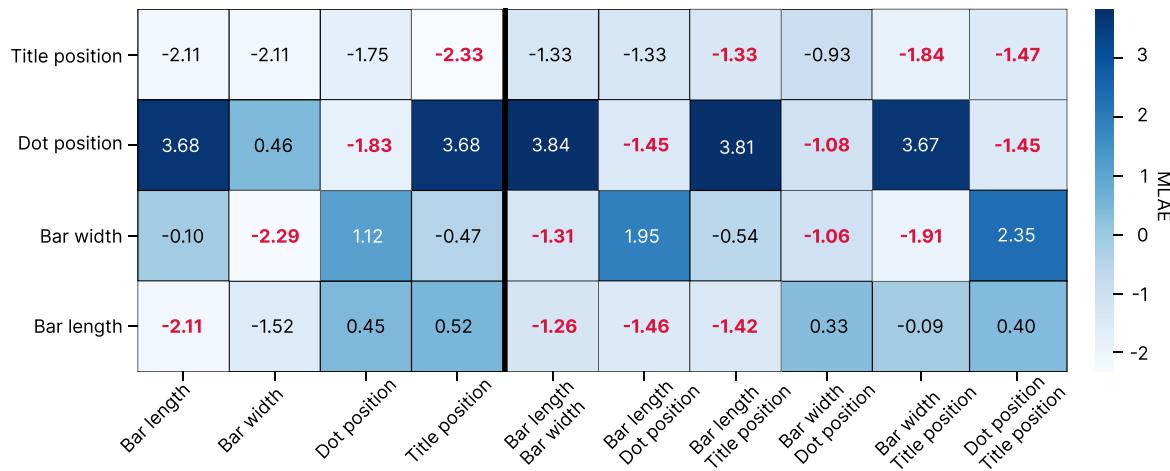


Fig. 10. Improving the generalization performance of CNNs by augmenting training data with different perturbations. The heatmap shows the average MLAE values for testing CNNs on the chart images with different perturbations, where columns correspond to differently trained networks and rows show different test conditions. The cell highlighted by a bold red font indicates that the testing stimuli have similar perturbations as the ones in the training stimuli.

Furthermore, we compare the generalization performance of CNNs and humans by progressively perturbing the design space of the standard visualizations and conducted a user study. We created a large dataset of visualizations called GRAPE by programmatically exploring the grammar of graphics using Vega-Lite. The results of our user studies show that (i) CNNs outperform humans but are less robust when the training data and test data are similar in visual encodings (IID), and (ii) human performance is more robust to perturbations of various visual parameters than that of CNNs (OOD). Hence, we conclude that CNNs' inference process involves many visual variables, yet they lack robust recognition of the target visual marks. Note that the human performance is collected from the user study with a small number of participants, we will conduct a larger-scale study to further confirm this conclusion in the future.

This work is the first step in the systematic study of the generalization behavior of CNNs for visualization applications. There is still a lot to explore. First, we identified our hyperparameters in an ad-hoc way. It would be better to use recently proposed AutoML techniques [25] to automatically build and train optimal neural networks for graphical perception tasks. Second, our analysis results reveal that some factors like background color and stroke color that are unrelated to the graphical perception tasks heavily impact the OOD generalization of CNNs. We plan to explore more recently proposed deep learning frameworks, such as Vision Transformers [11] and CLIP [49], to see if they are able to close the gap between humans and CNNs for visualization graphical perception tasks. Lastly, there are many visualizations (e.g., infographics [3], [47]) that do not comply with the grammar of graphics and still effectively communicate information. We would like to examine the CNNs' graphical perception abilities on such creative visualizations.

While this work focuses on graphical perception tasks, we believe that many applications regarding data visualizations should be evaluated from the generalizability perspective. For example, building on the recent success of LLMs, many chart-centered large multimodal models (LMMs) [40], [45]

have been developed. While these models demonstrated strong performance on several chart QA benchmarks, their sensitivity to various visual variations may still fall short of matching human-level robustness in relational reasoning tasks. Our study highlights the necessity of evaluating the OOD generalization of such LMMs. In the future, we plan to create a research benchmark for evaluating the generalization performance for different visualization techniques and analysis tasks.

REFERENCES

- [1] M. Bostock, V. Ogievetsky, and J. Heer. D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. DOI: 10.1109/TVCG.2011.185
- [2] J. Brownlee. *Better deep learning: train faster, reduce overfitting, and make better predictions*. Machine learning mastery, 2018.
- [3] Z. Chen, Y. Wang, Q. Wang, Y. Wang, and H. Qu. Towards automated infographic design: Deep learning-based auto-extraction of extensible timeline. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):917–926, 2020. DOI: 10.1109/TVCG.2019.2934810
- [4] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017. DOI: 10.1109/CVPR.2017.195
- [5] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. DOI: 10.2307/2288400
- [6] W. S. Cleveland and R. McGill. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833, 1985. DOI: 10.1126/science.229.4716.828
- [7] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger. Looks good to me: Visualizations as sanity checks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):830–839, 2019. DOI: 10.1109/TVCG.2018.2864907
- [8] W. Dai, M. Wang, Z. Niu, and J. Zhang. Chart decoder: Generating textual and numeric information from chart images automatically. *Journal of Visual Languages and Computing*, 48:101–109, 2018. DOI: 10.1016/j.jvlc.2018.08.005
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. DOI: 10.1109/CVPR.2009.5206848
- [10] G. I. Diaz, A. Fokoue-Nkoutche, G. Nannicini, and H. Samulowitz. An effective algorithm for hyperparameter optimization of neural networks. *IBM Journal of Research and Development*, 61(4/5):9–1, 2017. DOI: 10.1147/JRD.2017.2709578

- [11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. DOI: 10.48550/arXiv.2010.11929
- [12] J. Fu, B. Zhu, W. Cui, S. Ge, Y. Wang, H. Zhang, H. Huang, Y. Tang, D. Zhang, and X. Ma. Chartem: Reviving chart images with data embedding. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):337–346, 2020. DOI: 10.1109/TVCG.2020.3030351
- [13] R. Geirhos, K. Narayananappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021. DOI: 10.48550/arXiv.2106.07411
- [14] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018. DOI: 10.48550/arXiv.1811.12231
- [15] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018. DOI: 10.48550/arXiv.1808.08750
- [16] L. Giovannangeli, R. Bourqui, R. Giot, and D. Auber. Toward automatic comparison of visualization techniques: Application to graph visualization. *Visual Informatics*, 4(2):86–98, 2020. DOI: 10.1016/j.visinf.2020.04.002
- [17] L. Giovannangeli, R. Bourqui, R. Giot, and D. Auber. Color and shape efficiency for outlier detection from automated to user evaluation. *Visual Informatics*, 6(2):25–40, 2022. DOI: 10.1016/j.visinf.2022.03.001
- [18] L. Giovannangeli, R. Giot, D. Auber, J. Benois-Pineau, and R. Bourqui. Analysis of deep neural networks correlations with human subjects on a perception task. In *2021 25th International Conference Information Visualisation (IV)*, pp. 129–136, 2021. DOI: 10.1109/IV53921.2021.00029
- [19] G. Grinstein, D. Keim, and M. Ward. Simple guide to hyperparameter tuning in neural networks. *Towards Data Science*, 2002.
- [20] D. Haehn, J. Tompkin, and H. Pfister. Evaluating ‘graphical perception’ with CNNs. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):641–650, 2019. DOI: 10.1109/TVCG.2018.2865138
- [21] H. Haleem, Y. Wang, A. Puri, S. Wadhwa, and H. Qu. Evaluating the readability of force directed graph layouts: A deep learning approach. *IEEE Computer Graphics and Applications*, 39(4):40–53, 2019. DOI: 10.1109/MCG.2018.2881501
- [22] K. W. Hall, A. Kouroupis, A. Bezerianos, D. A. Szafir, and C. Collins. Professional differences: A comparative study of visualization task performance and spatial ability across disciplines. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):654–664, 2021. DOI: 10.1109/TVCG.2021.3114805
- [23] L. Harrison, D. Skau, S. Franconeri, A. Lu, and R. Chang. Influencing visual judgment through affective priming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 2949–2958. Association for Computing Machinery, 2013. DOI: 10.1145/2470654.2481410
- [24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90
- [25] X. He, K. Zhao, and X. Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021. DOI: 10.1016/j.knosys.2020.106622
- [26] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 203–212. Association for Computing Machinery, 2010. DOI: 10.1145/1753226.1753357
- [27] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021. DOI: 10.1109/ICCV48922.2021.00823
- [28] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. DOI: 10.48550/arXiv.1903.12261
- [29] T. Hinz, N. Navarro-Guerrero, S. Magg, and S. Wermter. Speeding up the hyperparameter optimization of deep convolutional neural networks. *International Journal of Computational Intelligence and Applications*, 17(02):1850008, 2018. DOI: 10.1142/S1469026818500086
- [30] K. Hu, M. A. Bakker, S. Li, T. Kraska, and C. Hidalgo. VizML: A machine learning approach to visualization recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 1–12. Association for Computing Machinery, 2019. DOI: 10.1145/3290605.3300358
- [31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. DOI: 10.1109/CVPR.2017.243
- [32] R. Ibrahim and M. O. Shafiq. Explainable convolutional neural networks: A taxonomy, review, and future directions. *ACM Computing Surveys*, 55(10):1–37, 2023. DOI: 10.1145/3563691
- [33] K. Kafle, B. Price, S. Cohen, and C. Kanan. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656, 2018. DOI: 10.1109/CVPR.2018.00592
- [34] S. E. Kahou, V. Michalski, A. Atkinson, Á. Kádár, A. Trischler, and Y. Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. DOI: 10.48550/arXiv.1710.07300
- [35] S. Kanthara, R. T. K. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022. DOI: 10.18653/v1/2022.acl-long.277
- [36] D. H. Kim, E. Hoque, and M. Agrawala. Answering questions about charts and generating visual explanations. *IEEE Transactions on Visualization and Computer Graphics*, p. 1–13, 2020. DOI: 10.1145/3313831.3376467
- [37] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. DOI: 10.48550/arXiv.1412.6980
- [38] A. Kirzhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. DOI: 10.1145/3065386
- [39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. DOI: 10.1109/5.726791
- [40] F. Liu, X. Wang, W. Yao, J. Chen, K. Song, S. Cho, Y. Yacoob, and D. Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023. DOI: 10.48550/arXiv.2311.10774
- [41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021. DOI: 10.1109/ICCV48922.2021.00986.
- [42] A. Lopes and K. Brodlie. Improving the robustness and accuracy of the marching cubes algorithm for isosurfacing. *IEEE Transactions on Visualization and Computer Graphics*, 9(1):16–29, 2003. DOI: 10.1109/TVCG.2003.1175094
- [43] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. DOI: 10.48550/arXiv.1711.05101
- [44] A. Mahinpei, Z. Kostic, and C. Tanner. Linecap: Line charts for data visualization captioning models. *arXiv preprint arXiv:2207.07243*, 2022. DOI: 10.48550/arXiv.2207.07243
- [45] F. Meng, W. Shao, Q. Lu, P. Gao, K. Zhang, Y. Qiao, and P. Luo. ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *arXiv preprint arXiv:2401.02384*, 2024. DOI: 10.48550/arXiv.2401.02384
- [46] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. Plotqa: Reasoning over scientific plots. pp. 1516–1525, 2019. DOI: 10.1109/WACV45572.2020.9093523
- [47] J. J. Otten, K. Cheng, and A. Drewnowski. Infographics and public policy: using data visualization to convey complex information. *Health Affairs*, 34(11):1901–1907, 2015. DOI: 10.1377/hlthaff.2015.0642
- [48] N. Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999. DOI: 10.1016/S0893-6080(98)00116-6
- [49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. DOI: 10.48550/arXiv.2103.00020
- [50] B. Saleh, M. Dontcheva, A. Hertzmann, and Z. Liu. Learning style similarity for searching infographics. In *Proceedings of the Graphics*

- Interface Conference*, pp. 59–64. Canadian Information Processing Society, 2015. DOI: 10.5555/2788890.2788902
- [51] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. DOI: 10.5555/3295222.3295250
- [52] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, 2016. DOI: 10.1109/TVCG.2016.2599030
- [53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017. DOI: 10.1109/ICCV.2017.74
- [54] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015. DOI: 10.48550/arXiv.1409.1556
- [55] J. Talbot, V. Setlur, and A. Anand. Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics*, 20(12):2152–2160, 2014. DOI: 10.1109/TVCG.2014.2346320
- [56] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6105–6114. PMLR, 2019.
- [57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, 2017. DOI: 10.5555/3295222.3295349
- [58] Y. Wang, Z. Wang, T. Liu, M. Correll, Z. Cheng, O. Deussen, and M. Sedlmair. Improving the robustness of scagnostics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):759–769, 2019. DOI: 10.1109/TVCG.2019.2934796
- [59] L. Wilkinson. The grammar of graphics. In *Handbook of Computational Statistics*, pp. 375–414. Springer, 2012.
- [60] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017. DOI: 10.1109/CVPR.2017.634
- [61] F. Yang, Y. Ma, L. Harrison, J. Tompkin, and D. H. Laidlaw. 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*, 2023. DOI: 10.1145/3544548.3581111
- [62] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016. DOI: 10.1109/CVPR.2016.10
- [63] L. Yuan, W. Zeng, S. Fu, Z. Zeng, H. Li, C.-W. Fu, and H. Qu. Deep colormap extraction from visualizations. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021. DOI: 10.1109/TVCG.2021.3070876
- [64] Z. Zeng, J. Yang, D. Moritz, J. Heer, and L. Battle. Too many cooks: Exploring how graphical perception studies influence visualization recommendations in draco. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1063–1073, 2024. DOI: 10.1109/TVCG.2023.3326527
- [65] P. Zhang, C. Li, and C. Wang. Viscode: Embedding information in visualization images using encoder-decoder network. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):326–336, 2020. DOI: 10.1109/TVCG.2020.3030343
- [66] J. Zhao, M. Fan, and M. Feng. ChartSeer: Interactive steering exploratory visual analysis with machine intelligence. *IEEE Transactions on Visualization and Computer Graphics*, 28(3):1500–1513, 2022. DOI: 10.1109/TVCG.2020.3018724



Zhenxing Cui is currently a first-year Master student in School of Computer Science and Technology, Shandong University. His research interests include information visualization and machine learning.



Lu Chen is currently a Ph.D. student at the State Key Lab of CAD&CG, Zhejiang University. He obtained his B.Eng. degree in computer science from Shandong University in 2022. His research interests lie primarily in computer vision and information visualization.



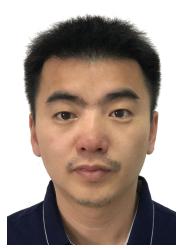
Yong Wang is a tenure-track assistant professor at College of Computing and Data Science, Nanyang Technological University (NTU), Singapore. His research interests include data visualization, visual analytics and explainable machine learning. He obtained his Ph.D. in Computer Science from Hong Kong University of Science and Technology in 2018. He received his B.E. and M.E. from Harbin Institute of Technology and Huazhong University of Science and Technology, respectively.



Daniel Haehn is an assistant professor of Computer Science at the University of Massachusetts Boston (UMB) and Associate of the Harvard Paulson School of Engineering and Applied Sciences. Prior to joining UMB, Daniel was a researcher at Brigham and Women's Hospital, the University of Pennsylvania, and Boston Children's Hospital. Daniel holds a PhD in Computer Science from Harvard University and a Diploma (BS/MS) in Medical Computer Science from the University of Heidelberg in Germany.



Hanspeter Pfister received the PhD degree in computer science from the Stony Brook University and an MS degree in electrical engineering from ETH Zurich, Switzerland. He is the Academic Dean of Computational Sciences and Engineering and An Wang Professor of Computer Science in the Harvard John A. Paulson School of Engineering and Applied Sciences. He is also an affiliate faculty member of the Center for Brain Science at Harvard. His research lies at the intersection of visualization, computer graphics, and computer vision.



Yunhai Wang is a professor in School of Information, Renmin University of China. He serves as the associate editor of *IEEE Transactions on Visualization and Computer Graphics*, *IEEE Computer Graphics and Applications*, and *Computer Graphics Forum*. His interests include scientific visualization, information visualization, and computer graphics.