

VisRecall: Quantifying Information Visualisation Recallability via Question Answering

Yao Wang, Chuhan Jiao, Mihai Băce, and Andreas Bulling

Abstract—Despite its importance for assessing the effectiveness of communicating information visually, fine-grained recallability of information visualisations has not been studied quantitatively so far. In this work, we propose a question-answering paradigm to study visualisation recallability and present VisRecall — a novel dataset consisting of 200 visualisations that are annotated with crowd-sourced human ($N = 305$) recallability scores obtained from 1,000 questions of five question types. Furthermore, we present the first computational method to predict recallability of different visualisation elements, such as the title or specific data values. We report detailed analyses of our method on VisRecall and demonstrate that it outperforms several baselines in overall recallability and FE-, F-, RV-, and U-question recallability. Our work makes fundamental contributions towards a new generation of methods to assist designers in optimising visualisations.

Index Terms—Information Visualisation, Recallability, Memorability, Machine Learning

1 INTRODUCTION

Memorability is an intrinsic, global, and stimulus-driven perceptual property that is important for better comprehension of visual stimuli [1, 2]. A growing body of work has studied image recognisability – one of the most fundamental attributes of memorability, both from a perceptual [1, 3] and a computational [4, 5] perspective. Recognisability has also been studied in information visualisations and previous work has revealed specific attributes that make visualisations memorable [6]. Recognisability measures whether a visualisation looks familiar or novel [3]. A visualisation that has unique features may stand out more and may therefore be more memorable. However, recognisability does not capture how effective a visualisation is in conveying information to observers. Other works have therefore studied *recallability* – a concept that goes beyond memorability, yet is complementary to it [7], by quantifying *what* viewers remember from a visualisation [8]. Despite its importance and potential for designing better information visualisations, a deeper understanding of which characteristics of visualisations influence recallability, and in which way, is currently missing.

Current methods to assess recallability rely on visualisation experts to assign a qualitative score to self-reported free-text descriptions of viewers [7]. This approach is cumbersome and only provides a single score representing overall recallability while hiding the contribution of individual visualisation characteristics. While Borkin et al. [7] noted the importance of titles for recallability on visualisations, Polatsek et al. [9] conducted three low-level analytical tasks, focusing on visual elements with extrema, or specific values. These works inspired us to quantify visualisations'

recallability by looking into specific types of visualisation elements, such as the title, elements with extrema, or distinct data points.

To quantify recallability, we propose to adopt a question-answering paradigm, similar to visual question answering (VQA) [10] that has become widely popular in computer vision. While originally introduced for natural images [10, 11], a VQA dataset was also proposed for information visualisations [12]. In our work, instead of proposing computational models of reasoning and correctly answering questions about images, we evaluate the performance of human observers in answering questions about visualisations and use their performance as a subjective measure of information visualisation recallability.

In this work, to quantify fine-grained recallability of information visualisations, we design and execute a question-answering based study to collect VisRecall: a novel visualisation dataset with 200 visualisations, which contains 1,000 high-quality questions annotated by visualisation experts and crowd-sourced human recallability scores. Our work is inspired by and extends prior task taxonomy on visualisations [9, 13] to define fine-grained recallability scores through five question types: identifying the title or theme, finding extrema, filtering data elements, retrieving values, and understanding structure (subsection 3.1). Through our analyses of VisRecall, we make several interesting findings: the highest recallability across question types occurs in questions that are about the title or the general theme (T-question), which is significantly higher than other question types. Moreover, a 10-second encoding duration is sufficient for most visualisation types, including *bar*, *pie*, *line*, and *scatter plots*. Based on VisRecall, we further present RecallNet, a novel method based on convolutional neural networks (CNNs) to predict one overall and five fine-grained recallability scores, one for each question type.

Our contribution is threefold: (1) We adapt a question-answering paradigm to quantify overall and fine-grained recallability of information visualisations. (2) We collect

- Yao Wang, Mihai Băce, and Andreas Bulling are with the Institute for Visualisation and Interactive Systems, University of Stuttgart, Germany, E-mail: {yao.wang, mihai.bace, andreas.bulling}@vis.uni-stuttgart.de.
- Chuhan Jiao is with Aalto University, E-mail: chuhan.jiao@aalto.fi
- Yao Wang is the corresponding author.

Manuscript received xx xx, 2021; revised xx xx, 202x.

VisRecall, a novel visualisation dataset with human recallability scores ($N=305$) from 1,000 questions and five question types. (3) We propose a computational model that predicts fine-grained recallability of visualisations. As such, our work points the way towards new methods and tools to create more effective information visualisations.

2 RELATED WORK

Our work is related to previous works on 1) image memorability, 2) perception and memorability of visualisations, and 3) chart question answering (chart QA) datasets.

2.1 Image Memorability

A pioneering study [3] reported a strong capability of humans to recognise what they have seen before even up to 10,000 images, which is denoted as “image recognition memory”. Isola et al. [4, 14] have demonstrated that memorability is an observer-independent property, which only depends on images [15, 16]. Furthermore, previous studies have proven that memorability could be reliably quantified for individual images by asking subjects to report whether images are novel or familiar. Large-scale memorability datasets have been collected for natural images, such as SUN-Mem [4], Figrim [17] and LaMem [5]. With the rise of deep learning, deep convolutional neural networks were proposed as computational methods to predict image memorability [5, 18, 19]. Recent work also integrated visual attention into the memorability prediction model [20]. Meanwhile, recallability is a complementary memory task to visual recognition [21], which requires subjects to view images and then recall what they have seen [22]. One previous work found that sketch-based methodologies can improve the recall of a sampling distribution from an experiment [23]. Several recent studies are consistent with the conclusion that image memorability variation may be distinct for recognition and recall tasks [8, 24]. Based on this, our work is the first to improve understanding of recallability characteristics and the factors that influence it on information visualisations.

2.2 Perception and Memorability of Visualisations

Pioneering works in the visualisation community have examined how different data types and tasks influence human perception [25, 26, 27]. Inbar et al. [28] reported that people prefer over-embellishment (i.e., “chart junk”) instead of Tufte’s minimalist design [29]. Bateman et al. [30] further claimed that the “chart junk” improves recognisability but is not essential for understanding the visualisation. This triggered a series of studies evaluating the impact of style on memorability and comprehensibility [31, 32, 33]. The effect of specific factors or components on recall memory has been investigated, such as interaction [34], prior knowledge [35], title [7, 36] and text redundancy [7]. Borkin et al. [6] studied visualisation memorability on the MASSVIS dataset, and their follow-up work [7] further conducted online crowdsourcing studies to quantify both recognisability and recallability. However, there are two main drawbacks to the previous recallability quantification procedure. Firstly, the method used to recall quality annotations is subjective and

cumbersome. In addition, visualisation experts are necessary to attribute these scores. Secondly, the description quality score scale with only four possible values is too coarse to represent a visualisation. To overcome these limitations, we introduce question answering as a powerful paradigm to quantify the recallability of information visualisations. Through multiple questions and answers on different visualisation characteristics, we propose a novel computational model to predict not only overall but also fine-grained recallability based on five different question types.

2.3 Chart Question Answering (Chart QA) Datasets

Despite the importance of information visualisations, chart QA datasets have only been proposed in recent years. FigureQA [12] was the first chart QA dataset. Images were plotted in simple and fully synthesised visualisations in five visualisation classes, along with polar questions. DVQA [37] focused specifically on the problem of visual reasoning on bar charts, which was used as a corpus for generating the topic of chart QA. PlotQA [38] and LEAF-QA [39] synthesised their question-answer pairs based on crowd-sourced question templates from real-world data sources to increase variety. One follow-up work collected human answers to the DVQA dataset [40] as baselines, which inspired us to use the human performance as a subjective measure of information visualisation recallability.

As a conclusion, the question-answering setting has not yet been used for memorability studies on visualisations, and current chart QA datasets are synthesised from simple templates with limited content, making it a distance away from real-world visualisations. However, chart QA provides an interesting means to quantify recallability. In our work, we evaluate and obtain recallability scores by asking users questions and validating their answers. Therefore, we present the design of our novel adaptation of a question-answering-based study on information visualisations and our novel VisRecall dataset in the next section.

3 VISRECALL DATASET

The currently available recallability scores on the visualisation dataset MASSVIS [6, 7] are annotated from free-text descriptions. However, its procedure to quantify recallability is coarse and cumbersome. Meanwhile, visual question-answering (VQA) datasets [10] selectively target elements of visualisations in different question-answer pairs, making it a suitable setting to quantify memorability objectively and efficiently. Under the question-answering paradigm, different tasks can be represented as different types of questions to viewers, and consequently, recallability is quantified by the accuracy in answering those questions.

Towards quantifying recallability, we propose the Visualisation Recallability Dataset (VisRecall) — a dataset consisting of 200 real-world information visualisations with crowd-sourced human recallability scores ($N=305$) obtained from 1,000 questions in five question types (see Figure 1). Visualisations in our dataset are mainly sourced from the MASSVIS dataset [6] to enable better alignment with prior works on this topic. The recognisability scores are also collected to replicate the previous memorability studies [6, 7].

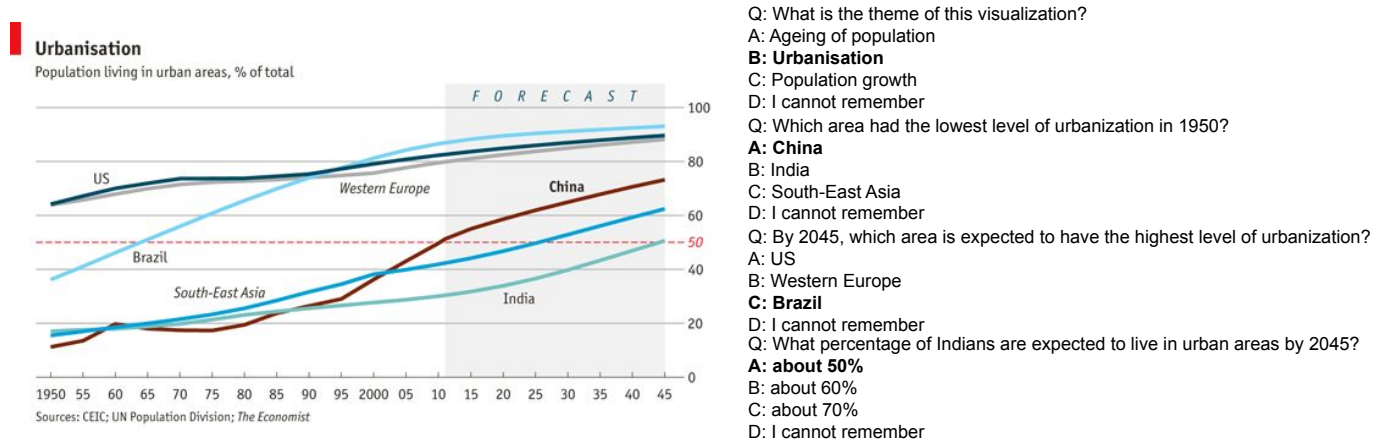


Fig. 1: Sample visualisation with multiple-choice questions from VisRecall. Five types of questions were designed by experts, which are questions regarding the title (T-questions), understanding structure or trend (U-questions), finding extrema (FE-questions), filtering elements (F-questions) and retrieving values (RV-questions). Each figure has at least two question types. The correct answer to each question is shown in **bold**. Image sourced from MASSVIS [6].

Our dataset and code are accessible at: <https://doi.org/10.18419/darus-2826>

3.1 Visualisation Collection and Question Types

We randomly selected a subset of 200 visualisations from the MASSVIS dataset [6]. Notably, we excluded all infographics from our collection, since infographics have the highest recognisability and recallability compared to all other types of visualisations [7]. However, scatter plots represented only 5% of the sampled subset. Therefore, we collected 20 additional scatter plot visualisations by crawling the web through search engines (Google, Bing) using the keyword “scatter plots”. Then, we replaced some bar plots with the web-crawled plots to balance the visualisation type classes. The final distribution of visualisation types is: 56 bar plots, 45 line plots, 27 scatter plots, 22 pie plots, 25 tables and 25 others. Those visualisations that don't belong to any of the first five types are categorised as *others*, including box charts, isotype charts, and other complex visualisations.

When creating the questions for our VisRecall dataset, we had to identify question types that are not only suitable for static information visualisations, since our images are primarily sourced from the MASSVIS dataset but also applicable to all visualisations. Therefore, inspired by prior work, we selected five question types: T-questions, U-questions, FE-questions, F-questions, and RV-questions. T-questions were questions regarding the title or the visualisation theme, which were used to analyse how the existence of the title influenced the description scores of visualisation recallability [7]. U-questions were about understanding the plot structure [38] or the general trend [39]. For the remaining three question types, we followed the work by Polatsek et al. [9] who picked these question types for the visualisation types from MASSVIS. The three question types were FE-questions (finding an extremum attribute value), F-questions (filtering visualisation elements based on specific criteria) and RV-questions (retrieving values for a specific visualisation element).

All question-answering data were created by five data visualisation experts. They were asked to provide five

questions per visualisation, and every visualisation has at least two question types. Each question corresponds to four possible answer options. Only one option is correct, two other options are choices with similar, yet incorrect answers, and the last option is always “I cannot remember”. See supplementary material for question examples. All annotations were saved separately in standard JSON files for each visualisation. There are 193, 150, 178, 99, and 64 visualisations in VisRecall that have at least one T-, FE-, F-, RV-, and U-question, respectively.

T-question. T-questions are about the title or the general theme of the plot and do not require any reasoning. Example questions: *What is the title of the visualisation?*, *What is the theme of the visualisation?* For the incorrect choices in T-questions, we either replaced keywords or phrases with words of similar, but different meanings, such as changing *car thefts* to *car accidents* or *car manufacturers*, or using titles from other visualisations, such as using *Expectations On House Prices Above 2009 Projections* and *HIV Prevalence in Women Aged 15-49 Years by Region, 1990-2007* as incorrect choices for *Covered Transactions by Sector and Year, 2009-2011*.

FE-question. These are questions about finding extreme values in the visualisation that fulfil certain conditions, without asking any exact numbers. Example questions: *Which area had the lowest level of urbanization in 1950?*, and *Which particle is the latest discovered?* We used other elements that appeared in the visualisation as incorrect answer choices. As seen in Figure 1, *India* and *South-East Asia* were the incorrect alternative choices for *China* in the question *Which area had the lowest level of urbanization in 1950?*

F-question. These are questions about filtering data elements based on specific criteria. Example questions: *Which particle is Bosons?* and *What is the source of the data?* For F-questions, we either changed keywords to their synonyms, or used other elements that appeared in visualisations as incorrect alternative choices, such as using *Electron* and *Muon* for *Photon* in the question *Which particle is Bosons?*

RV-question. These are questions about retrieving a specific value located in the plot. Example questions: *What is the maximum percentage of aid allocated?* and *What percentage*

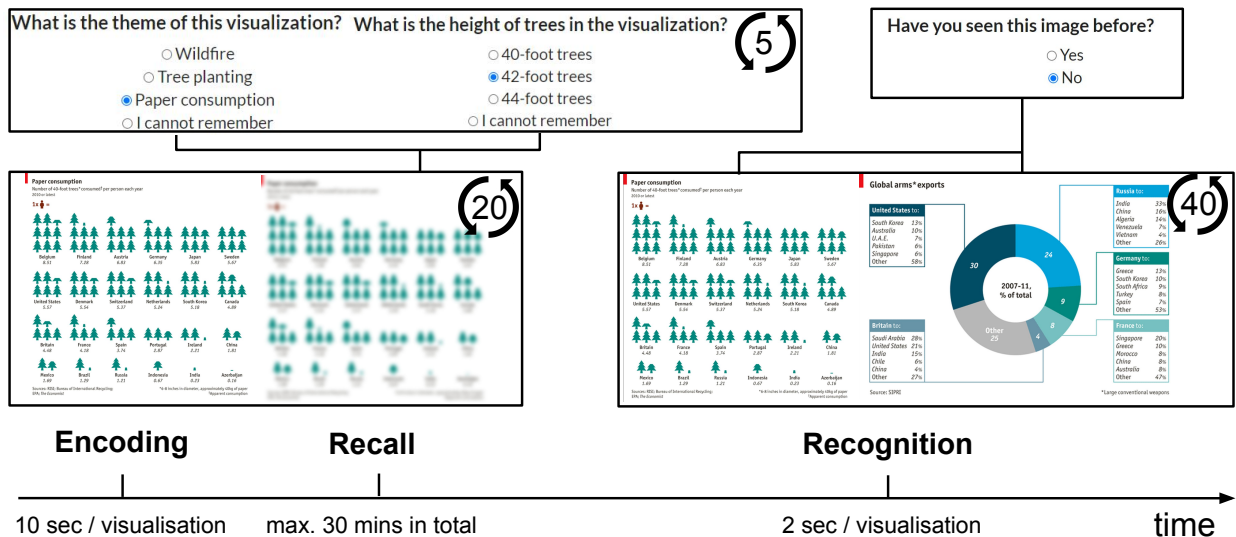


Fig. 2: Experiment design. From left to right: Visualisations are shown to viewers for a fixed duration in the “Encoding” phase. In the “Recall” phase, visualisations are blurred and each has a multiple-choice question next to it with a single correct answer. Finally, visualisations are shown to viewers for 2 seconds in the “Recognition” phase. The numbers in the circular arrows indicate the number of repetitions.

of Indians are expected to live in urban areas by 2045? (see Figure 1). Example incorrect choices: *about 60 %* and *about 70 %* for *What percentage of Indians are expected to live in urban areas by 2045?*, and the correct answer is *about 50 %*.

U-question. These are questions about understanding the structure or the trend of a visualisation. Example questions: *What does the purple curve represent?* and *What decreases as time goes by?* Example incorrect choices: for structure questions, other elements appearing in the visualisation are used, such as using *Red* and *Blue* as incorrect choices for *Green* in the question *What color stands for Residents?* As for questions about understanding trends, the choices are *increasing*, *decreasing* and *almost the same*.

3.2 Crowd-sourcing Study Set-up & Participants

Our study design is illustrated in Figure 2. In the encoding phase of our study, study participants were shown a sequence of visualisations for a fixed duration. We followed the 10-second encoding phase in a prior memorability study [6], and also conducted the study with a 20-second encoding phase to see the impact of encoding duration on recallability. We asked participants to memorise as much of the information presented in each visualisation as possible. To advance from the encoding phase to the recall phase, our study participants had to click on the “next” button. In the recall phase, each visualisation was shown at 50% of the size from the encoding phase and blurred by a 24-pixel Gaussian filter to make the text unreadable. The question orders were predefined to avoid the situation where some questions might provide answers to other questions. The blurry visualisation was shown with a single multiple-choice question. The presentation order of the first three multiple-choice options for each question was randomly shuffled once and fixed for all participants, while the option *I cannot remember* always appeared last. The following question would be shown only if the participant clicked the next button, and

they could not return to the previous question. This setting was to avoid providing hints in upcoming questions. Before running our user study, we did preliminary tests with three designs: showing visualisations one-by-one, two-by-two, and four-by-four. In the one-by-one setting, we found the task too easy with very high recallability scores. The four-by-four task, i.e. first encoding four different visualisations before the recall phase, was found too difficult. Therefore, as a trade-off and to reduce the effect of working memory [41], we empirically selected the two-by-two setting for our user study. In each set, the encoding phase of two images were presented, followed by their recall phase, before repeating the process for the next set of two images. Then, the recognition phase involved an online memorability game similar to prior work [6]. Study participants were presented with a sequence of images, and they had to select if they had seen this visualisation before. In each Human Intelligence Task (HIT), 40 blurred images were shown for 2 seconds each. The images in the recognition phase contained 20 visualisations that were the same in the recall phase, and 20 fillers from a different group. Finally, participants were asked to provide anonymous feedback on the study design in a questionnaire.

To support the study, we implemented the procedures in a web application. We then integrated our application into an existing crowd-sourcing toolbox that worked well with the Amazon Mechanical Turk (MTurk) platform [42]. We deployed our experiment on MTurk to collect recallability and recognisability scores on all 200 visualisations, splitting them randomly into ten groups of 20 visualisations per HIT. Visualisation types were balanced among all groups (see Figure 1 in supplementary material). MTurk workers could participate in multiple HITs. To participate in one of our HITs, a worker had to be a Master Worker approved by MTurk as a quality check. Master Workers are top workers rated by MTurk who have consistently demonstrated high

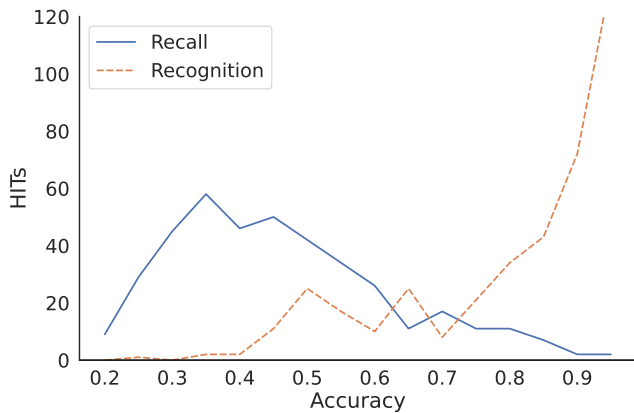


Fig. 3: Recall and Recognition accuracy over all 404 HITs. Participants can recognise most of the visualisations easily, but only answer less than half of the questions correctly.

quality work. Workers were paid \$4.00 for completing each HIT. To ensure data quality, we filtered out 467 HITs (305 workers) if the answers were all “Yes” or “No” in the recognition task. For each visualisation, we received an average of 40.4 ($\sigma = 16.9$) valid responses. All workers were distributed in various educational levels: 8.2% two-year degree, 56.9% four-year degree, 22.3% master’s degree or higher, and 12.6% other / unreported. The age groups were 44.1% in 25 - 34, 28.5% in 35 - 44, 12.4% in 45 - 54 and 9.9% over 55. In the anonymous feedback form at the end of our study, most workers responded positively, with two examples being: “Great self test for capable of memory power” and “After taking survey, I’m really getting interested in learning data plots and visualisations”.

3.3 Data Analysis

Recallability Formulation. For each question, we measured the recall accuracy as follows: $Acc = \frac{RA}{RA+WA}$, where RA is the number of correct answers, and WA is the number of wrong answers, including the number of *I cannot remember* answers. If we focus on viewers who have selected choices excluding *I cannot remember*, the accuracy can be computed as: $Acc' = \frac{RA}{RA+WA-CNR}$, where CNR stands for the number of *I cannot remember*. Averaging all questions of type t in a visualisation gives us the recallability by question type and is computed as: $Rec_t = \frac{1}{n} \sum_{i=1}^n Acc(i), question_i \in t$. By averaging all questions in a visualisation, we have the overall recallability of a visualisation as: $Rec = \frac{1}{n} \sum_{i=1}^n Acc(i)$.

HIT-wise Recallability. HIT-wise recallability as well as recognition accuracy across HITs ($N = 404$) are shown in Figure 3. 63.9% of HITs have a recognition accuracy higher than 0.85, and 34.83% are higher than 0.95, which shows that our study participants could easily recognise most of the visualisations ($\mu = 0.83$). Meanwhile, they could only answer less than half of the questions correctly ($\mu = 0.49$).

Fine-grained Recallability by Question Type. Figure 4 illustrates that T-questions have the highest recall accuracy among all question types both when including *I cannot remember* ($\mu = 0.66$), and excluding *I cannot remember* ($\mu = 0.69$). The accuracy of T-questions is significantly higher than

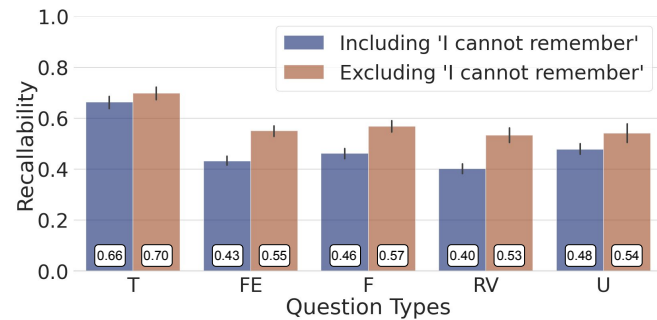


Fig. 4: Recallability scores by question type. T-questions have significantly higher recallability scores than all other question types (FE-, F-, RV-, and U-questions). Additionally, 24.7% of the viewers selected *I cannot remember* in RV-questions, while only 5.1% of the viewers selected *I cannot remember* in T-questions.

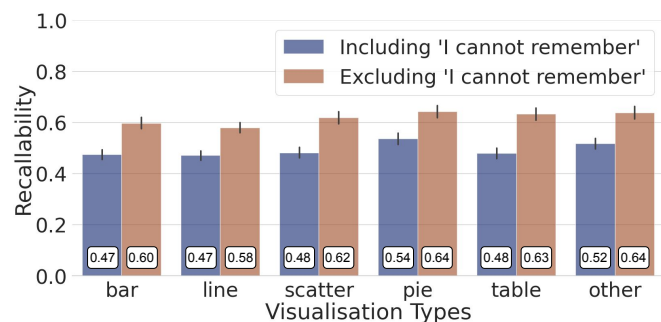


Fig. 5: Recallability scores by visualisation type. Pie plots have significantly higher recallability scores compared with all other visualisation types (bar, line, scatter plots, tables, and others).

other question types ($t(1969) = 18.87, p < 0.001$). 24.7% of viewers selected *I cannot remember* in RV-questions, and 21.4%, 18.8%, 11.7% for FE-, F- and U-questions, respectively. Only 5.1% of the study participants selected *I cannot remember* in T-questions. We observed a mean proportion of 19.1% ($\sigma = 13.0\%$) of study participants who selected *I cannot remember* in all visualisations. The lowest proportion is 3%, while more than 50% of participants selected *I cannot remember* in seven visualisations. Figure 8 shows visualisations with the most and least *I cannot remember* answers from VisRecall. We observe that a high visualisation complexity is common among those visualisations with the most *I cannot remember* answers.

Fine-grained Recallability by Visualisation Type. Figure 5 illustrates the recallability scores by visualisation type. A one-way ANOVA test is applied across visualisation types, and we observed a significant difference for both excluding *I cannot remember* ($F = 4.412, p < 0.001$), and including *I cannot remember* ($F = 6.916, p < 0.001$). Post-hoc analyses with Tukey’s HSD [43] confirmed that the recallability scores of pie plots are significantly higher than any other visualisation types including *I cannot remember* (for all pairs, $p < 0.001$). For excluding *I cannot remember*, line plots are significantly lower than pie plots ($t = 3.725, p = 0.003$) and others ($t = 3.458, p = 0.007$).

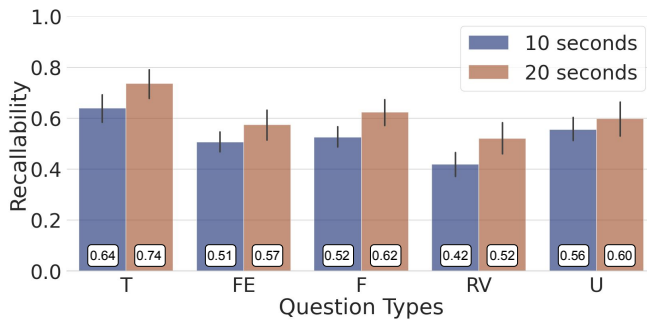


Fig. 6: Recallability scores under a 10-second and 20-second encoding phase by question type in one MTurk group.

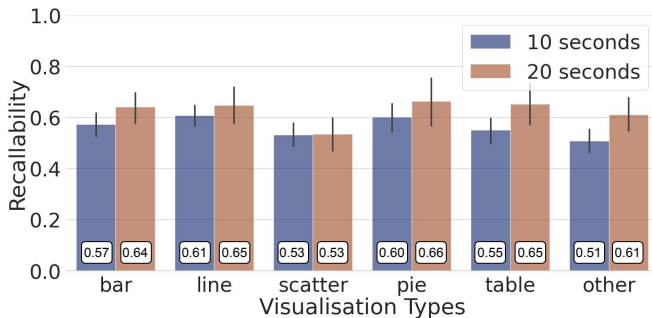


Fig. 7: Recallability scores under a 10-second and 20-second encoding phase by visualisation type in the same MTurk group as Figure 7.

Encoding Duration. The study of the 20-second encoding phase was conducted in two randomly selected MTurk groups. We observed significant improvement of recallability in one group ($t(198)=2.284$, $p=0.023$) from 10-second ($\mu=0.51$, $\sigma=0.51$) to 20-second encoding phase ($\mu=0.57$, $\sigma=0.22$), but not in the other group ($t(198)=1.627$, $p=0.105$). The recallability scores by question type is shown in Figure 6. Prolonging the 10-second encoding phase to 20 seconds, the recallability scores of each question type all increased. No significant improvements of recallability scores were found in T-questions ($t(59)=1.367$, $p>0.05$), F-questions ($t(59)=1.796$, $p>0.05$), FE-questions ($t(59)=1.474$, $p>0.05$), RV-questions ($t(59)=1.951$, $p>0.05$), or U-questions ($t(59)=0.830$, $p>0.05$). Figure 7 illustrates the recallability scores by visualisation type. Significant improvements of recallability scores are found in tables ($t(59)=2.144$, $p=0.036$) and others ($t(59)=2.969$, $p=0.009$), but not in pie plots ($t(59)=1.141$, $p>0.05$), bar plots ($t(59)=1.675$, $p>0.05$), line plots ($t(59)=1.052$, $p>0.05$), or scatter plots ($t(59)=0.817$, $p>0.05$).

Recallability and Recognisability: a Comparison to Prior Work. To the best of our knowledge, the description quality in [7] is the closest work to ours on the quantification of recallability, where free text descriptions of what participants recall about the visualisations were recorded. Description quality was rated from 0 to 3 where 0 was a completely incorrect description, and 3 was a precise description regarding the topic and at least one detail [7]. We found 31 visualisations with description quality in our VisRecall dataset, and calculated the average descrip-

tion quality scores and the mean visualisation accuracy of all questions in the overlapped visualisations. The Pearson’s correlated coefficient (CC) between description quality scores and mean visualisation accuracy is 0.36 with *I cannot remember*, while it is 0.35 without *I cannot remember*.

For a comparison to prior work on recognisability [6, 7], we also calculated the memorability (or recognisability) score on VisRecall. According to Borkin et al. [6], the hit rate (HR) and false alarm rate (FAR) were computed as: $HR = \frac{HITS}{HITS+MISSES}$ and $FAR = \frac{FA}{FA+CR}$. Then, the recognisability (memorability) of a visualisation was measured as: $d' = Z(HR) - Z(FAR)$, where Z was the inverse cumulative Gaussian distribution. Figure 9 (left) shows the distribution of the raw HR scores of all visualisations from the recognition phase. Figure 9 (right) shows the highest and lowest ranked visualisations across recognisability (memorability) and recallability from our VisRecall dataset. The full memorability (recognisability) and recallability scores of all visualisations are available in supplementary material. Visualisations in each quadrant were ranked highest or lowest 15% among all visualisations.

Data-ink ratio. Data-ink ratio is a commonly used visual attribute introduced by Tufte et al. [44]. A high data-ink ratio visualisation contains a large share of ink presenting information about data. Following the previous annotation procedure in [7], three visualisation researchers independently rated the data-ink ratio for each visualisation in the VisRecall dataset. Data-ink ratio was rated from 1 to 3 where 1 was a low proportion of ink that was related to data, and 3 was high. The ranking was directly applied if more than two researchers agreed. In cases when all three researchers gave different rankings, the visualisation was reviewed and discussed by all three researchers for a consensus. We observed that the high data-ink ratio group has the highest recallability score ($\mu=0.621$), compared with $\mu=0.599$ for the middle data-ink ratio, and $\mu=0.606$ for the low data-ink ratio. A one-way ANOVA test was applied between data-ink ratio groups. Still, no significance was observed in either excluding *I cannot remember* ($F=1.134$, $p=0.322$) or including *I cannot remember* ($F=2.43$, $p=0.088$).

Visual Density. Visual density is another visual attribute to rate the overall density of visual elements without distinguishing between data and non-data elements [6]. The annotation procedure of visual density was the same as data-ink ratio. Visual density was rated from 1 to 3 where 1 was low density of visual elements in the visualisation, and 3 was high. We observed that the high visual density group has the highest recallability score ($\mu=0.619$), compared with $\mu=0.608$ for middle visual density, and $\mu=0.600$ for low visual density. A one-way ANOVA test is applied across visual density groups, but no significance is observed in either excluding *I cannot remember* ($F=0.740$, $p=0.478$) or including *I cannot remember* ($F=0.245$, $p=0.782$).

4 COMPUTATIONAL MODEL FOR PREDICTING FINE-GRAINED RECALLABILITY

Our analyses on VisRecall yielded several insights on recallability in information visualisations. There are currently no baseline methods, for predicting either overall recallability or fine-grained recallability. Existing computational

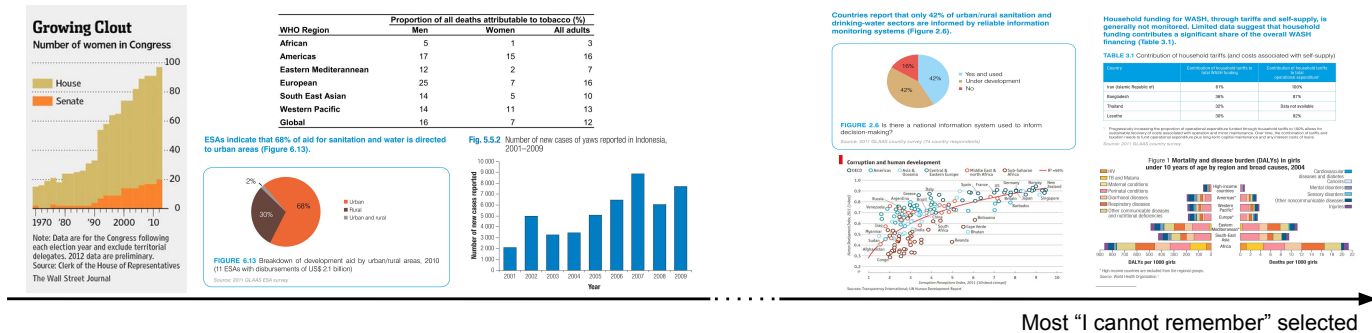


Fig. 8: Example visualisations with the most and fewest answers *I cannot remember* from VisRecall. We observed a higher degree of visualisation complexity for those with multiple *I cannot remember* answers.

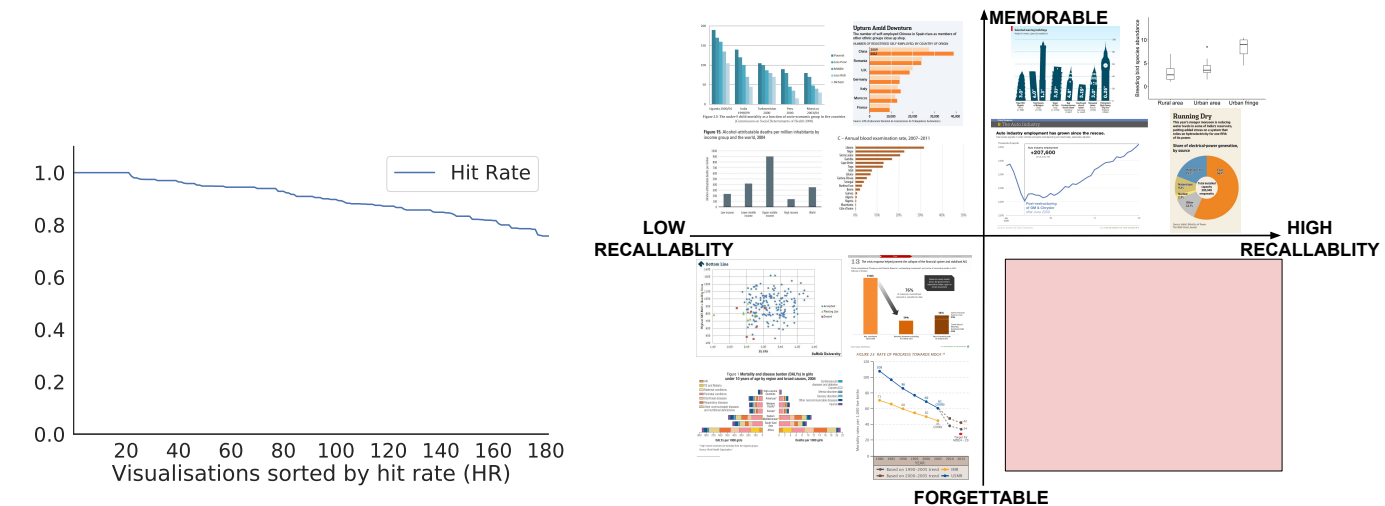


Fig. 9: Left: Raw hit rate (HR) of target visualisations from the recognition phase. Right: The highest and lowest ranked visualisations (within 15%) across recognisability (memorability) and recallability in VisRecall. The x-axis represents the recallability score computed from overall visualisation question accuracy (independent of question type), and the y-axis represents the memorability score from previous work [7].

models only aimed at predicting memorability, also known as recognisability [5, 20]. Therefore, we propose Recallability Network (RecallNet), a lightweight and effective neural network for recallability prediction.

4.1 Model Architecture

We extend and build on state-of-the-art architectures from other computer vision tasks, such as semantic segmentation [45, 46] and image classification [47, 48], and use such methods as the backbone of our architecture. We design our RecallNet with the specific goal of predicting both overall and fine-grained recallability scores in one single model (see Figure 10 for an overview). Inspired by UMSI [49], the current state-of-the-art architecture for visual importance prediction on graphic designs, we employ the Xception [45] model to effectively encode spatial information. Then, a global average pooling layer, a dense layer with 256 neurons, and finally a dense layer with 2 neurons are sequentially connected. One output neuron predicts the general recallability score, and the other one predicts the fine-grained recallability score.

4.2 Implementation Details & Model Training

We trained RecallNet using weights obtained from the Xception model – which was pretrained on ImageNet [50]. RecallNet was trained with the Adam [51] optimizer with a learning rate of 0.002 and 1:1 Mean Squared Error (MSE) joint loss for the two branches predicting the overall recallability score and the fine-grained recallability score. We averaged all five questions for each image to prepare the ground truth of overall recallability scores. To train our RecallNet to predict fine-grained recallability scores for a certain question type, we only used those visualisations that contained that question type from VisRecall. There are 193, 150, 178, 99, and 64 visualisations with at least one T-, FE-, F-, RV-, and U-question, respectively. Five-fold cross-validation was applied to all evaluation processes. All experiments were conducted on a single NVIDIA 2060 Super GPU with 8GB VRAM.

Baseline Methods. Since no previous computational models focused on predicting recallability on visualisations, we designed three methods as baselines. We replaced the Xception feature encoder in RecallNet with VGG-16 [47] and ResNet-34 [48] as the two baselines. We trained all baseline models for 10 epochs on VisRecall starting from

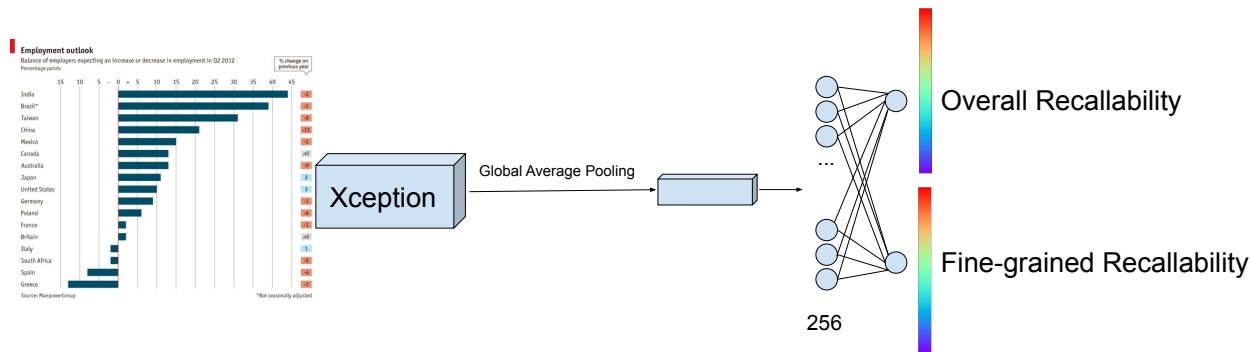


Fig. 10: Method overview. RecallNet leverages the Xception model [45] to effectively encode spatial information. Then, a global average pooling layer, a dense layer with 256 neurons, and finally a dense layer with 2 neurons are sequentially connected. One output neuron predicts the general recallability score, and the other one predicts the fine-grained recallability score.

ImageNet [50] pretrained weights. We used the Adam optimizer [51] with a learning rate of 0.002 and MSE loss for training.

4.3 Model Evaluation

The prediction error is calculated as the mean squared error between the human and the predicted recallability scores. We compared the prediction error of our RecallNet method to the two baselines VGG-16 and ResNet-34. Table 1 summarises fine-grained recallability prediction error on VisRecall under a 5-fold cross-validation evaluation. We used the MSE to evaluate the prediction error. Results showed that RecallNet outperformed the baselines under overall recallability and four fine-grained recallability scores, with a MSE of 0.035 for overall recallability, and 0.021, 0.022, 0.017, 0.043 for FE-, F-, RV-, and U-questions respectively. ResNet-34 was the best performing method for T-questions with a MSE of 0.047, while our RecallNet was second with a MSE of 0.052.

Ablation study. We further carried out an ablation study to investigate how each fine-grained recallability score influences overall recallability (see Table 2). With RecallNet, the overall recallability trained with T-questions has the lowest MSE of 0.030 and the most stable variance of 0.006. With ResNet-34 [48], the overall recallability trained with RV-questions has the lowest MSE of 0.029 and the most stable variance of 0.008. With VGG-16, the overall recallability trained with T-questions has the lowest MSE of 0.037 and the most stable variance of 0.007.

5 DISCUSSION

First, we underline the novelty of VisRecall and its potential in applications such as chart QA [12]. Second, we discuss how recallability and recognisability are different yet connected. Then, several interesting insights from our analyses are reported. Finally, the limitations and future work are discussed.

VisRecall Dataset. VisRecall is the first dataset to introduce fine-grained recallability on an information visualisation dataset as well as high-quality question-answering annotations. The recallability scores are metrics that reveal human performance with a specific type of question. With rich

annotations of the elements necessary for the answers, the recallability score of a certain question could be converted into 2D spatial representations (e.g. recallability heatmaps). The recallability maps could be introduced as an additional feature input to downstream tasks, such as chart QA. Additionally, VisRecall is a novel visualisation question-answering dataset that uses real-world, visually rich visualisations coming in part from the MASSVIS dataset. The questions for chart QA datasets [38, 39] were collected by regular crowd workers. In contrast, all the questions in our VisRecall came from visualisation experts, which promises a higher quality of questions than chart QA datasets. Moreover, most visualisations in chart QA datasets [38] are generated pragmatically. However, when it comes to real-world visualisations, the structural information is usually missing, and researchers have to retrieve it, often by manual annotation [7], which is time-consuming and constrains the dataset size. The introduction of recallability to the question-answering setting and the high quality of visualisations and questions enable VisRecall to trigger fundamental studies on chart QA.

Recallability vs. Recognisability (Memorability). The bottom-right quadrant in Figure 9 (right) is completely empty, which means that there are no such visualisations with high recallability (top 15%) and low memorability (bottom 15%) in VisRecall. This means that *visualisations have to be sufficiently memorable before they become recallable*. The visualisations in the top-right quadrant share some characteristics, like a big and highlighted title and some explanatory text. Meanwhile, the visualisations in the top-left quadrant of Figure 9 (right) have high recognisability and low recallability. Compared to the top-right quadrant, visualisations in the top-left quadrant are less recallable. All visualisations in the top-left quadrant are simple monotone plots with few embellishment (e.g. isotype plots). The visualisations in the bottom-left quadrant are easily forgettable and hard to recall. These visualisations are usually overly complex and don't have meaningful titles or additional explanatory text to convey key messages. Compared to the bottom-left quadrant, all the visualisations in the top-left and top-right quadrant are always with titles, which aligns well with the findings in previous studies [6, 7]. However, the recallability

TABLE 1: Prediction error (MSE) of fine-grained recallability on VisRecall under 5-fold cross-validation evaluation. Best results are shown in **bold**, second-best are underlined.

Methods	Overall	T	FE	F	RV	U
RecallNet (ours)	0.035 ± 0.005	0.052 ± 0.009	0.021 ± 0.003	0.022 ± 0.004	0.017 ± 0.004	0.043 ± 0.025
ResNet-34 [48]	0.043 ± 0.013	0.047 ± 0.015	0.068 ± 0.024	0.070 ± 0.042	0.043 ± 0.008	<u>0.050 ± 0.018</u>
VGG-16 [47]	<u>0.036 ± 0.013</u>	0.053 ± 0.017	<u>0.054 ± 0.019</u>	0.076 ± 0.029	0.057 ± 0.010	0.059 ± 0.025

TABLE 2: Ablation study on the prediction error (MSE) of how fine-grained recallability influences the overall recallability. Best results in each row are shown in **bold**.

Methods	T	FE	F	RV	U
RecallNet (ours)	0.030 ± 0.006	0.079 ± 0.052	0.032 ± 0.008	0.035 ± 0.013	0.172 ± 0.215
ResNet-34 [48]	0.043 ± 0.013	0.078 ± 0.087	0.060 ± 0.035	0.029 ± 0.008	0.033 ± 0.013
VGG-16 [47]	0.037 ± 0.007	0.046 ± 0.022	0.041 ± 0.019	0.079 ± 0.053	0.077 ± 0.011

between the data-ink ratio and visual density groups is not significantly different. Either those visual features are not highly correlated with recallability, or the size of our VisRecall prevented the confirmation of significance. Nevertheless, our study on VisRecall validated previous results and provided interesting insights into how recallability and recognisability (memorability) are different yet connected.

Free Recall vs. Question-Answering-cued Recall. The low correlated relationship (CC = 0.35) between description quality [7] and our recallability score drew our attention. The description quality generated from prior work was from free-text descriptions without any context, but our recallability was computed from the mean accuracy of five multiple-choice questions per image. The low correlated relationship suggests that the information (context) in multiple-choice questions might be an essential factor that influenced recallability. One possible explanation is that our study provided cues for visualisations which mitigated the memory decaying process (forgetting) [52].

Impact of Encoding Duration on Recallability. The analysis on encoding duration provided several insights (see Figures 6 and 7). Those text-heavy and complex visualisations (tables, *others*) are more sensitive to viewing duration, and a 10-second encoding phase is sufficient for most basic visualisation types (pie, bar, line, and scatter plots). Filtering data and retrieving value questions (F- and RV-questions) gained quasi-significant improvement in one of the MTurk groups, while no significance was found for the other three question types (T-, FE-, and U-questions). It suggests that a more prolonged encoding phase is more beneficial to those questions that require detailed answers (F- and RV-questions).

Limitations. There is always a trade-off between quality and quantity, which was also the case when designing and collecting our VisRecall dataset. Due to the increasing workload in designing high-quality questions for the question-answering settings specifically targeted for each visualisation, the scale of VisRecall became relatively small. We conducted a preliminary evaluation using Grad-CAM [53], which is a method used for understanding and explaining the predictive behaviour of CNN-based models. However, our qualitative analysis did not reveal any generalisable patterns that can be directly linked to higher-level visualisation features such as visual density or data-ink ratio. To allow more explainable models for recallability prediction, it is

essential to extend the size of VisRecall.

Future Work. How recallability can be applied to reality is a fundamental question. Visualisation type recommendation is one practical use case for a visualisation recommendation system [54]. Prior research has proposed ways to decide whether line or scatter plots are more suitable for time series data [55]. One possible application is to make use of recallability scores to recommend a visualisation type for given data. For visualisations that demonstrate the same data but with different visualisation types, our RecallNet might be useful in recommending a visualisation type that maximises recallability.

In the future, we plan to enrich VisRecall with more complex visualisation types such as box, radar and combination plots. Furthermore, gaze behaviour analysis in a question-answering setting on information visualisations has not yet been studied. However, it is a fundamental step to understand the human visual attention system while viewing visualisations. While physical laboratory studies require special-purpose eye tracking equipment, online crowd-sourcing studies or gaze estimation from substitution devices (e.g., mouse, web camera) can be used as a proxy to human attention. In the future, we will investigate such methods to collect human attention data and extend VisRecall with such annotations.

6 CONCLUSION

This work presented a novel adaptation of a question-answering-based study to collect VisRecall, a novel visualisation dataset with 200 real-world visualisations annotated with crowd-sourced human recallability scores in five question types, along with a deep convolutional network to predict fine-grained recallability of visualisations. Overall, this work made a substantial leap towards quantifying fine-grained recallability scores on information visualisations and envisions several potential applications, such as visualisation optimisation.

ACKNOWLEDGMENTS

Y. Wang was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 251654672 - TRR 161. M. Bâce was funded by a Swiss National Science Foundation (SNSF) Early Postdoc. Mobility

Fellowship (grant number 199991). A. Bulling was funded by the European Research Council (ERC; grant agreement 801708).

We would like to thank Sruthi Radhakrishnan for the creation of visualisations for our application, Maurice Koch, Guanhua Zhang, Adnen Abdessaied, and Florian Strohm for data annotation, Dominike Thomas for paper editing support, as well as Yue Jiang, Daniel Weiskopf and Nils Rodrigues for helpful comments on this paper.

REFERENCES

- [1] W. A. Bainbridge, D. D. Dilks, and A. Oliva, "Memorability: A stimulus-driven perceptual neural signature distinctive from memory," *NeuroImage*, vol. 149, pp. 141–152, 2017.
- [2] Z. Bylinskii, L. Goetschalckx, A. Newman, and A. Oliva, "Memorability: An image-computable measure of information utility," *Human Perception of Visual Information*, pp. 207–239, 2021.
- [3] L. Standing, "Learning 10000 pictures," *The Quarterly Journal of Experimental Psychology*, vol. 25, no. 2, pp. 207–222, 1973.
- [4] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 145–152.
- [5] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2390–2398.
- [6] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister, "What makes a visualization memorable?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2306–2315, 2013.
- [7] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond memorability: Visualization recognition and recall," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 519–528, 2015.
- [8] N. C. Rust and V. Mehrpour, "Understanding image memorability," *Trends in Cognitive Sciences*, vol. 24, no. 7, pp. 557–568, 2020.
- [9] P. Polatsek, M. Waldner, I. Viola, P. Kapec, and W. Benesova, "Exploring visual attention and saliency modeling for task-based visual analysis," *Computers & Graphics*, vol. 72, pp. 26–38, 2018.
- [10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.
- [11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6904–6913.
- [12] 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.
- [13] R. Amar, J. Eagan, and J. Stasko, "Low-level components of analytic activity in information visualization," in *IEEE Symposium on Information Visualization*. IEEE, 2005, pp. 111–117.
- [14] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1469–1482, 2013.
- [15] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva, "Visual long-term memory has a massive storage capacity for object details," *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, pp. 14 325–14 329, 2008.
- [16] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "Understanding the intrinsic memorability of images," Massachusetts Inst of Tech Cambridge, Tech. Rep., 2011.
- [17] M. Mancas and O. Le Meur, "Memorability of natural scenes: The role of attention," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 196–200.
- [18] S. Perera, A. Tal, and L. Zelnik-Manor, "Is image memorability prediction solved?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1–9.
- [19] A. Jaegle, V. Mehrpour, Y. Mohsenzadeh, T. Meyer, A. Oliva, and N. Rust, "Population response magnitude variation in inferotemporal cortex predicts image memorability," *Elife*, vol. 8, p. e47596, 2019.
- [20] J. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, "Amnet: Memorability estimation with attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6363–6372.
- [21] F. Haist, A. P. Shimamura, and L. R. Squire, "On the relationship between recall and recognition memory," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 18, no. 4, pp. 691–702, 1992.
- [22] A. P. Yonelinas, "The nature of recollection and familiarity: A review of 30 years of research," *Journal of Memory and Language*, vol. 46, no. 3, pp. 441–517, 2002.
- [23] J. Hullman, M. Kay, Y.-S. Kim, and S. Shrestha, "Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 446–456, 2017.
- [24] W. A. Bainbridge, E. H. Hall, and C. I. Baker, "Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory," *Nature Communications*, vol. 10, no. 1, pp. 1–13, 2019.
- [25] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984.
- [26] S. M. Kosslyn, "Understanding charts and graphs," *Applied Cognitive Psychology*, vol. 3, no. 3, pp. 185–225, 1989.
- [27] S. Pinker, "A theory of graph comprehension," *Artificial Intelligence and the Future of Testing*, pp. 73–126, 1990.
- [28] O. Inbar, N. Tractinsky, and J. Meyer, "Minimalism in information visualization: attitudes towards maximizing the data-ink ratio," in *Proceedings of the 14th European Conference on Cognitive Ergonomics: Invent! Explore!*,

- 2007, pp. 185–188.
- [29] E. R. Tufte, “The visual display of quantitative information,” *The Journal for Healthcare Quality*, vol. 7, no. 3, p. 15, 1985.
- [30] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks, “Useful junk? the effects of visual embellishment on comprehension and memorability of charts,” in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 2573–2582.
- [31] R. Borgo, A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, L. Floridi, and M. Chen, “An empirical study on using visual embellishments in visualization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2759–2768, 2012.
- [32] A. V. Moere, M. Tomitsch, C. Wimmer, B. Christoph, and T. Grechenig, “Evaluating the effect of style in information visualization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2739–2748, 2012.
- [33] X. Shu, A. Wu, J. Tang, B. Bach, Y. Wu, and H. Qu, “What makes a data-gif understandable?” *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1492–1502, 2020.
- [34] S.-H. Kim, Z. Dong, H. Xian, B. Upatising, and J. S. Yi, “Does an eye tracker tell the truth about visualizations?: Findings while investigating visualizations for decision making,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2421–2430, 2012.
- [35] Y.-S. Kim, K. Reinecke, and J. Hullman, “Explaining the gap: Visualizing one’s predictions improves recall and comprehension of data,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 1375–1386.
- [36] H.-K. Kong, Z. Liu, and K. Karahalios, “Trust and recall of information across varying degrees of title-visualization misalignment,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.
- [37] K. Kafle, B. Price, S. Cohen, and C. Kanan, “Dvqa: Understanding data visualizations via question answering,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 5648–5656.
- [38] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar, “Plotqa: Reasoning over scientific plots,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1527–1536.
- [39] R. Chaudhry, S. Shekhar, U. Gupta, P. Maneriker, P. Bansal, and A. Joshi, “Leaf-qa: Locate, encode & attend for figure question answering,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3512–3521.
- [40] K. Kafle, R. Shrestha, S. Cohen, B. Price, and C. Kanan, “Answering questions about data visualizations using efficient bimodal fusion,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1498–1507.
- [41] A. M. Owen, K. M. McMillan, A. R. Laird, and E. Bullmore, “N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies,” *Human Brain Mapping*, vol. 25, no. 1, pp. 46–59, 2005.
- [42] A. Newman, B. Mcnamara, C. Fosco, Y. Zhang, P. Sukhum, M. Tancik, N. Kim, and Z. Bylinskii, “Turkeyes: A web-based toolbox for crowdsourcing attention data,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.
- [43] J. W. Tukey, “Comparing individual means in the analysis of variance,” *Biometrics*, pp. 99–114, 1949.
- [44] E. R. Tufte, N. H. Goeler, and R. Benson, *Envisioning information*. Graphics press Cheshire, CT, 1990, vol. 126.
- [45] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.
- [46] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [47] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2015.
- [48] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [49] C. Fosco, V. Casser, A. K. Bedi, P. O’Donovan, A. Hertzmann, and Z. Bylinskii, “Predicting visual importance across graphic design types,” in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 249–260.
- [50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [51] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [52] J. Jonides, R. L. Lewis, D. E. Nee, C. A. Lustig, M. G. Berman, and K. S. Moore, “The mind and brain of short-term memory,” *Annual Review of Psychology*, vol. 59, pp. 193–224, 2008.
- [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*, 2017, pp. 618–626.
- [54] K. Hu, M. A. Bakker, S. Li, T. Kraska, and C. Hidalgo, “Vizml: A machine learning approach to visualization recommendation,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.
- [55] Y. Wang, F. Han, L. Zhu, O. Deussen, and B. Chen, “Line graph or scatter plot? automatic selection of methods for visualizing trends in time series,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 2, pp. 1141–1154, 2017.



Yao Wang is a PhD student at the University of Stuttgart, Germany. He received his BSc. in Intelligence Science and Technology and MSc. in Computer Software and Theory both from Peking University, China, in 2017 and 2020, respectively. His research interests include computer vision and human-computer interaction, with a focus on visual attention modelling on information visualizations.



Andreas Bulling is Full Professor of Computer Science at the University of Stuttgart, Germany, where he directs the research group "Human-Computer Interaction and Cognitive Systems". He received his MSc. in Computer Science from the Karlsruhe Institute of Technology, Germany, in 2006 and his PhD in Information Technology and Electrical Engineering from ETH Zurich, Switzerland, in 2010. Before, Andreas Bulling was a Feodor Lynen and Marie Curie Research Fellow at the University of Cambridge, UK, and a Senior Researcher at the Max Planck Institute for Informatics, Germany. His research interests include computer vision, machine learning, and human-computer interaction.



Chuhan Jiao is an MSc. student in Computer Science at Aalto University. He received his BEng. in Computer Science and Technology from Donghua University, China. His research interest lies at the intersection of human-computer interaction and computer vision.



Mihai Băce is a post-doctoral researcher in the Perceptual User Interfaces group at the University of Stuttgart, Germany. He did his PhD at ETH Zurich, Switzerland, at the Institute for Intelligent Interactive Systems. He received his MSc. in Computer Science from École Polytechnique Fédérale de Lausanne, Switzerland, and his BSc. in Computer Science from the Technical University of Cluj-Napoca, Romania. His research interests include computational human-computer interaction with a focus on sensing and

modelling user attention.