VisImages: A Fine-Grained Expert-Annotated Visualization Dataset

Dazhen Deng, Yihong Wu, Xinhuan Shu, Jiang Wu, Siwei Fu, Weiwei Cui, and Yingcai Wu

Abstract—Images in visualization publications contain rich information, e.g., novel visualization designs and implicit design patterns of visualizations. A systematic collection of these images can contribute to the community in many aspects, such as literature analysis and automated tasks for visualization. In this paper, we build and make public a dataset, VisImages, which collects 12,267 images with captions from 1,397 papers in IEEE InfoVis and VAST. Built upon a comprehensive visualization taxonomy, the dataset includes 35,096 visualizations and their bounding boxes in the images. We demonstrate the usefulness of VisImages through three use cases: 1) investigating the use of visualizations in the publications with VisImages Explorer, 2) training and benchmarking models for visualization classification, and 3) localizing visualizations in the visual analytics systems automatically.

Index Terms—Visualization dataset, crowdsourcing, literature analysis, visualization classification, visualization detection

1 Introduction

Images are crucial to publications in the visualization community (e.g., IEEE VIS), showcasing the visual designs, system frameworks, model details, experiment results, etc. The images contain a rich trove of visual information (e.g., color schemes and shapes of graphical elements) and semantic information (e.g., different combinations of charts) that can advance the understanding of the field and the research of artificial intelligence for visualization (AI4VIS) [1].

First, the information contained in the images can greatly benefit the analysis of visualization literature, which mainly employs publication metadata like keywords, citations, and co-authorship [2]. Incorporating image data into literature analysis can understand the visualization field from more dimensions (e.g., what types of charts are frequently used in different venues across years; how different charts are used in different research topics?) and inspire new research problems (e.g., how different charts can be organized together to better represent data?).

- *Dazhen Deng, Yihong Wu, Jiang Wu, and Yingcai Wu are with the State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310027, China. E-mail: {dengdazhen, wuyihong, wujiang5521, ycwu}@zju.edu.cn.*
- *Xinhuan Shu is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong. E-mail: xinhuan.shu@connect.ust.hk.*
- *Siwei Fu is with the Zhejiang Lab, Hangzhou, Zhejiang 310024, China. E-mail: fusiwei339@gmail.com.*
- *Weiwei Cui is with the Microsoft Research Asia, Beijing 100080, China. E-mail: weiweicu@microsoft.com.*

Manuscript received 8 June 2021; revised 1 February 2022; accepted 14 February 2022. Date of publication 7 March 2022; date of current version 30 May 2023.
This work was supported in part by NSFC under Grants 62072400 and 62002331 and in part by the Collaborative Innovation Center of Artificial Intelligence by MOE and Zhejiang Provincial Government (ZJU). This work was also supported by Zhejiang Lab under Grants 2020KE0AA02 and 2021KE0AC02, and Microsoft Research Asia.
(Corresponding author: Yingcai Wu.)
Recommended for acceptance by A. Endert.
Digital Object Identifier no. 10.1109/TVCG.2022.3155440

Moreover, a visualization dataset from visualization publications affords new opportunities for the application of AI4VIS. Existing studies [3], [4], [5] collect chart images online and train computer vision models for visualization tasks, such as chart classification. The images in these datasets usually comprise common chart types with simple layouts due to the data source. Therefore, the models trained on these datasets might fail when dealing with charts with complex designs, such as system interfaces with multiple charts. Consequently, a new visualization dataset from visualization publications can advance the research in developing machine learning models and serve as a benchmark to test the generalizability and robustness of models.

However, the images from visualization publications cannot be directly utilized for above tasks, as they lack adequate annotations that describe the semantics information in the images. For example, when analyzing visualization literature, the information of the visualization types in the images is necessary to index and search the images of interest for in-depth analysis. In addition, there is a lack of proper labels for applying frontier machine learning models to visualization tasks (e.g., deconstructing visual analytics systems or generating visualizations). In particular, object detection models require the bounding boxes of visualizations in the images, and image-text translation models need textual descriptions about the images.

To facilitate the use of images, we build and make public a visualization dataset, *VisImages*, from visualization publications. The data in VisImages is organized into three levels, namely, papers, images, and visualizations (Fig. 1). The paper data includes metadata of the paper (i.e., title, authors, conference, and year) and image data. The metadata of the paper is coded from vispubdata.org [10]. The image data is a list of data for each image, which includes the image file name, textual caption, image position (i.e., bounding box) in the paper, and visualization data. The visualization data is a list of data for each visualization, including the visualization type and visualization position (i.e., bounding box) in the

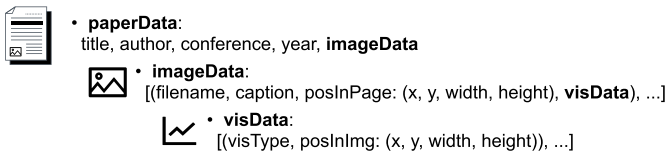


Fig. 1. Data structure of VisImages.

image. In all, the dataset contains the data of 1,397 papers, 12,267 images, and 35,096 visualizations.

Creating such a dataset faces three major challenges. The first challenge is categorizing diverse visualizations. Though various taxonomies have been proposed to define and distinguish different visualization types [11], [12], they cannot fully cover the various designs in visualization publications, such as novel glyphs (Fig. 2a) or the variations of existing visualizations (Fig. 2b). In addition, annotating visualization types requires extensive visualization expertise. Second, the diverse layouts (Figs. 2c and 2d) and large quantity of visualizations in the publications makes identifying the positions of visualizations difficult and tedious. Third, it is also challenging to ensure the quality of the annotations, given the diverse knowledge expertise of annotators and the lack of "the ground truth". Addressing conflicts and reducing biases is critical for data annotation.

To address the first challenge, we use a comprehensive taxonomy proposed by Borkin *et al.* [12] and annotate the visualizations by regarding them as compositions of different visualizations [13]. Based on the taxonomy, we invite visualization practitioners to annotate the visualization types. For the second challenge, we set up a series of criteria for decomposing and localizing visualizations in the images. Based on the criteria, we recruit trained crowd workers to annotate the bounding box for each visualization. To tackle the third challenge, we adopt a series of measures for quality control, including the gold standard [14], majority voting, and sampling test. Our contributions are threefold.

- We build a novel dataset named VisImages from IEEE VAST and InfoVis for the research of literature analysis and AI4VIS. We release the dataset and codes for data collection at https://visimages.github.io/.
- We present an overview of the use of visualizations in visualization publications and compare it with the images collected from public sources. From the analysis, we gain insights into the peculiarities of visualizations in academic publications.
- We showcase the usefulness of VisImages through three use cases, namely, 1) exploring and analyzing the evolution of visualizations in publications with VisImages Explorer, 2) evaluating the generalizability and robustness of visualization classification models, and 3) localizing visualizations in the interfaces of visual analytics systems automatically.

2 Related Work

This section introduces related studies on visualization datasets and visualization literature analysis.

2.1 Image Datasets in Visualization

We first introduce existing visualization image datasets and demonstrate that VisImages can greatly boost the research in visualization because of its unique data source, visualization publications. In addition, VisImages is, to our best knowledge, the most comprehensive one regarding visualization quantity, layout complexity, label types, etc.

The visualization community has built a variety of image datasets of basic charts (such as bar charts or scatterplots) with simple layouts. The images in these datasets are mainly collected from the Internet, such as social media (e.g., Twitter) and media outlet (e.g., BBC), or generated by visualization libraries (e.g., D3 [15], Vega-Lite [16]). For example, Battle *et al.* [3] gathered over 41,000 SVG-based charts, manually labeled each chart one of 24 chart types, and trained classification models to analyze the chart distribution on the web. Jung *et al.* [17] collected 5,659 images consisting of 10 chart types to develop models for chart classification and proposed ChartSense for data extraction. Savva *et al.* [4] delivered a dataset containing 2,601 single-chart images in 10 categories. The dataset is used to develop a system called ReVision to redesign the charts for better visual styles. Similarly, Poco *et al.* [18] collected more than 5,000 bitmap images and annotated chart types (area, line, bar, and scatter) and textual annotations (labels and titles of axes and legend) of the charts. The dataset is used for reconstructing the original charts with declarative grammars, such as Vega-Lite [16]. Borkin *et al.* [12], [19] developed MassVis for memorability study. They collected more than 2,000 single-chart visualizations and categorized them into 12 categories. For each image, they evaluated the data-ink ratio and visual density through crowdsourcing. They also annotated a subset of 396 images for detailed information (e.g., annotations, axis, and data). Lee *et al.* [20] collected the images from scientific publications and categorized them into equation, photo, diagram, etc.

Other than basic charts, the images in visualization publications contain novel designs created by visualization experts, such as system interfaces, which are the interests

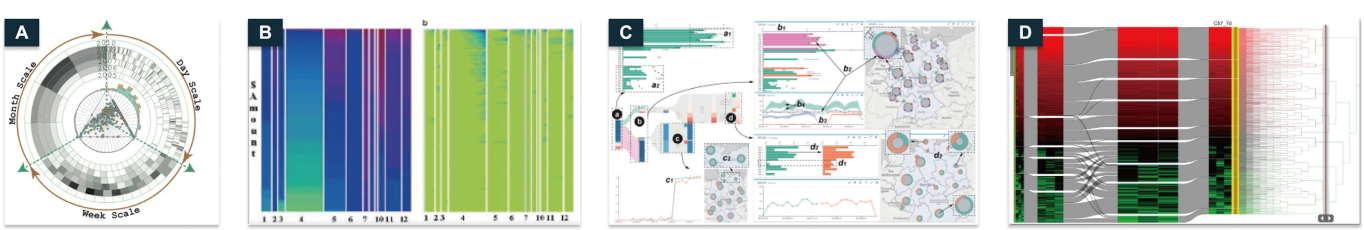


Fig. 2. Novel designs in visualization academic publications. (a) shows the design of OpinionSeer [6], consisting of a triangular scatterplot, a circular bar chart, and donut charts. (b) shows the pixel bar charts [7], a variant of bar chart. (c) shows the interface of TPFlow [8], which is facilitated with a set of views of different visualizations. (d) shows a design by Lex *et al.* [9], which combines a Sankey diagram, heatmaps, and a tree.

TABLE 1
Existing Visualization Datasets in the Visualization Community

Dataset	Audience	Layout	#Annotated Visualizations	#Images	#Annotated Categories	Label Types			How to label?
						type	bbox	caption	
MassVis [12]	general users	single	~2000	~2000	12	✓	-	-	manual annotation
REV [18]	general users	single	~5000	~5000	4	✓	-	-	machine generation + manual refinement
Beagle [3]	general users	single	33,778	33,778	24	✓	-	-	manual annotation
ChartSense [17]	general users	single	~2000	~2000	10	✓	-	-	search engine + manual refinement
VizioMetrics [20]	general users	-	-	~**4,986,302***	5	✓	-	-	object classfication + manual annotation
VIS30K [21]	visualization experts	-	-	~30,000	4	✓	-	-	object detection + manual refinement
MV Dataset [22]	visualization experts	multiple	not reported	360	14	✓	✓	-	manual annotation
VisImages	**visualization experts***	**multiple***	**35,096***	12,267	**34***	✓	✓	✓	manual annotation

many studies lay in. For example, Li *et al.* [23] collected images from IEEE SciVis and conducted user studies to understand the relation between memorability and image characteristics. Chen *et al.* [21] adapted object detection models (Faster-RCNN [24] and YOLOv3 [25]) to crop the figures and tables from visualization publications and proposed VIS30K, a dataset of figures and tables. Similarly, Zeng *et al.* [26] collected figures from IEEE VIS to visualize and analyze the evolution of figures. However, detailed information of the visual designs, such as chart types, chart positions, and captions, are not considered in their dataset. Chen *et al.* [22] collected figures of multiple-view visualizations (MVs) from the publications and annotated the view positions and types. They contributed a corpus of 360 MV images and statistically analyzed the view types and layouts. We conduct a detailed comparison between VisImages and the existing datasets from the perspectives of target audience, visualization layout, quantity, category, and label type, as illustrated in Table 1.

VisImages is novel for visualization research compared to existing datasets because of its usefulness for visualization understanding and its usability due to comprehensive annotations. First, VisImages comprises fresh designs with complex configurations that are created by visualization researchers. Meanwhile, existing datasets, such as VizioMetrics [20], Beagle [3], ChartSense [17], MassVis [12], [19], and REV [18], collect data from public sources and contain a majority of basic charts. The unique data source makes VisImages a compelling dataset for studying the variations or combinations of common visualization types as well as innovative visual designs. Besides, the novel visual designs in VisImages serve as a challenging benchmark for new research methods that boost the study of AI4VIS [1].

Second, VisImages is outstanding in its comprehensive annotations, which are well-suited for various scenarios. 1) *Annotation Granularity*: Compared to VIS30K [21] and VizioMetrics [20] that classify the images by their goals or usages (e.g., diagram, equation, photo, plot, and table) at the image level, VisImages further specify the visualization types and bounding boxes within the images with a fine-grained taxonomy. 2) *Design Complexity:* Compared to MassVis [12], [19] and REV [18] that are mainly composed of single-chart visualizations, VisImages includes detailed annotations on multi-chart visualizations, such as visual analytics systems. 3) *Visualization Quantity:* In addition to the complexity, VisImages also covers a large quantity of visualizations (about 100x of the quantity of MV Dataset [22]), which can fulfill the requirements of training deep learning models. With the above features, VisImages can be directly used in many scenarios, such as classification and localization, which is practical for the research of AI4VIS.

2.2 Visualization Literature Analysis & Datasets

Literature analysis is an important research area for indexing and understanding the publications. Current studies mainly use the following four types of data: text, citations, authors, and metadata [2].

Many datasets of visualization publications [10], [27], [28], [29], [30] are used to support interactive literature analysis. The most up-to-date one is vispubdata.org [10], which contains metadata of publications in IEEE VIS sub-conferences. The publication data, such as authors, references, and keywords, is collected from the electronic proceedings. A series of visual analytics tools, such as CiteVis2, CiteMatrix, and VisList [10], were proposed on the basis of vispubdata.org. Ponsard *et al.* [31] proposed PaperQuest, which is a tool to search for relevant papers that users are interested in. Several studies [32], [33] also attempt to organize publications based on research topics. Chuang *et al.* [32] introduced a framework to use topic modeling for the analysis of InfoVis corpus. Isenberg *et al.* proposed KeyVis [33] that extracts the keywords of visualization papers. However, none of the above studies investigate the image data. VisImages comprises a large dataset of images with rich annotations, which can provide additional dimensions for literature analysis.

3 Dataset Construction

In this section, we overview the construction of VisImages.

3.1 Data Preprocessing

To construct VisImages, we started with collecting images from top-venue visualization publications (Fig. 3a). In this study, we focused on 2D static visualizations and collected the images from VAST (IEEE Conference on Visual Analytics Science & Technology) and InfoVis (IEEE Conference on Information Visualization). We excluded SciVis (IEEE

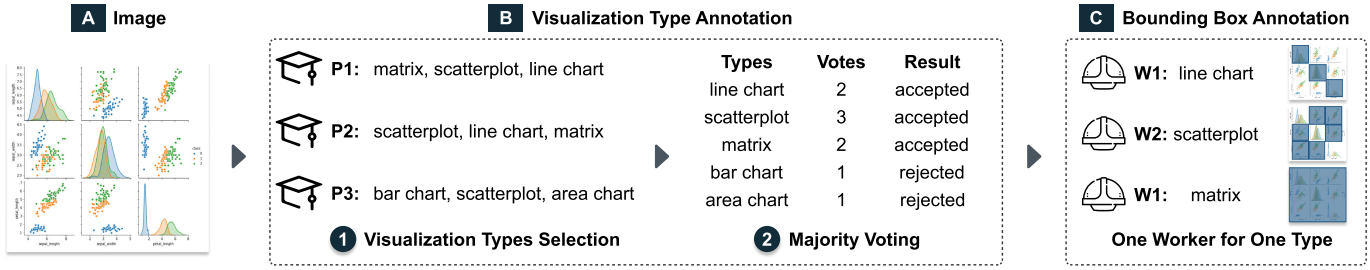


Fig. 3. The pipeline of data annotation. (a) shows an image sample for annotation. (b) shows the process of annotating visualization types, in which three visualization practitioners independently specify the visualization types that the image contains. (c) shows the process of majority voting, in which we accept the visualization types that receive at least two votes from three practitioners. (d) shows the process of annotating bounding boxes, in which each crowd worker focuses on one type of visualization and draws the bounding boxes to specify the positions of visualizations.

Conference on Scientific Visualization) papers since these papers generally comprise a large number of images depicting the results of 3D rendering, which are beyond the scope of this paper. We first downloaded PDF files of the papers according to the digital object identifier (DOI) provided in vispubdata.org [10]. Next, we used PDFFigures 2.0 [34] to extract images and captions from the files, and manually checked and corrected the results. We focused on the figures and tables indexed by *Figure* and *Table* and inline figures without a caption. In total, we processed 1,397 papers in VAST and InfoVis dated from 1996 to 2018 and collected 12,267 images with 12,057 textual captions.

3.2 Visualization Taxonomy

To classify visualizations, we used the taxonomy proposed by Borkin *et al.* [12] that categorizes visualizations into a two-level structure, i.e., 12 categories with sub-types. The taxonomy classifies the visualizations used in the public (i.e., infographics, news media, scientific journals, and government & world organization) based on their visual encodings (e.g., bar and area), visual tasks (e.g., statistics), and visual layouts (e.g., diagram). However, we discovered that the original taxonomy has some sub-types that are similar in definition, such as bar chart and histogram. To avoid ambiguities, we merge these types together. In addition, we also discovered that some visualization types are not listed in the taxonomy, such as icicle plot and glyph-based visualizations. Therefore, we added these types into the original taxonomy. The taxonomy used in our dataset consists of 13 categories and 34 sub-types, as shown in Table 2.

3.3 Pipeline of Data Annotation

To annotate the visualization types and bounding boxes in the images, we designed a pipeline that takes advantage of the expertise of visualization practitioners and the scalability of crowdsourcing.

We first recruited qualified participants from the university for visualization type annotation given the visualization expertise required for recognizing the types of visualizations (Fig. 3b)). To identify different visualizations, we basically used a visualization taxonomy proposed by Borkin *et al.* [12]. Specifically, for each image, three participants were asked to identify all possible visualization types that appear (Fig. 3b1). To address conflicts, we adopted majority voting that a visualization type was accepted only if at least two participants voted for it (Fig. 3b2). After this step, we collected 10,289 images containing visualization types within our taxonomy. The details of visualization type annotation are illustrated in Section 4.1.

Second, we annotated the bounding boxes of visualizations in the images through crowdsourcing (Fig. 3c). To ensure high data quality, we carefully designed the tasks and performed cross-validation (Section 4.2). As a result, we obtained a dataset of 35,096 bounding boxes, each corresponding to a specific visualization. The detailed procedure of bounding box annotation is demonstrated in Section 4.2.

4 Process of Data Annotation

We label the visualization types and bounding boxes based on the refined taxonomy and annotation pipeline.

4.1 Visualization Type Annotation

Identifying visualizations and their variations is challenging and requires extensive knowledge of visualization. Thus, we recruited the researchers and students who were experienced in visualization research to annotate the visualization types that appear in the images. Please note that the term "type" refers to the visualization sub-types in Table 2.

Participants. We recruited 25 participants, including 1 senior visualization expert who had six-year experience in visualization research, 13 PhD candidates with the research interest of visualization, 7 master students majoring in information visualization, and 4 undergraduate students who

TABLE 2
Visualization Taxonomy

Categories	Sub-types
Area	area chart, proportional area chart (PAC)
Bar	bar chart
Circle	donut chart, pie chart
Diagram	flow diagram, chord diagram, Sankey diagram, Venn diagram
Statistic	box plot, error bar, stripe graph
Table	table
Line	contour graph, line chart, storyline, polar plot, parallel coordinate (PCP), surface graph, vector graph
Map	map
Point	scatter plot
Grid & Matrix	heatmap, matrix
Text	phrase net, word cloud, word tree
Graph & Tree	graph, tree, treemap, hierarchical edge bundling (HEB), sunburst/icicle plot
Special	glyph-based visualization, unit visualization

had taken the undergraduate course of data visualization. Most of them (15/25) had published papers in IEEE VIS.

Procedure. The annotation procedure consisted of a training session and a formal study. In the training session, we first introduced the taxonomy and the definition of each visualization sub-type with examples. Then we introduced the details of annotation tasks. Specifically, annotating visualization types for an image is a multi-label task in our study. In such a task, a participant was shown an image and asked to select all the visualization sub-types occurring in the image based on our taxonomy. If the participant thought the image does not contain the visualization types within our taxonomy, they could choose the additional option "others." After the introduction, participants were asked to take a test to ensure that they had correctly understood the taxonomy. The test contained 20 images covering all visualization types (an image might include multiple types), and participants were considered eligible for the formal study only if they correctly annotated more than 18 images. All participants passed the test at their first attempt.

In the formal study, all participants annotated the images independently. Before the annotation, the participants had to enter their names as identification. All annotated data recorded the name of the participant. During the annotation, the participants could not see the results of others. We developed and deployed an online interface for data annotation. The data was stored in a backend server, which recorded the annotation logs and managed the task assignments.

Each round of annotation comprised 40 tasks. It took about 10 minutes to complete a round. Each participant was assigned at most 40 rounds. To avoid overloading, participants were allowed to accomplish all images within five days. We paid \$0.05 for each accepted task.

Quality Control. We adopted the methods of gold standards and majority voting for quality control. The gold standards were the images manually selected and inserted into each round to test whether the participants were focusing on the tasks. The gold standard images contained simple charts placed in an obvious position, and the participants were expected to correctly specify the visualization types easily. Each round included eight images with gold standards. If a participant failed in more than one gold standard in a round, all results from this round would be rejected, and we would reassign these images to other participants. Finally, we obtained 940 accepted rounds and 102 rejected rounds. In the accepted rounds, the participants obtained an accuracy of 96.4% on the gold standards.

In addition, we used majority voting to address ambiguities in the annotation. Specifically, each image would be annotated by three participants independently. The selection of a visualization sub-type by a participant would be regarded as a "vote". For each image, the sub-types with at least two votes would be accepted as the fact that the image contained the instances of these visualization sub-types. Otherwise, the sub-types would be suspended for further discussion. Due to the majority voting, the entire annotation process contained at least 12,267 images $\times$ 3 repetitions = 36,801 annotations. We computed the acceptance rate of a participant to evaluate the similarity between his/her annotations and the accepted results. Overall, the participants gained an average acceptance rate of 88.3%. In addition, we compute the intercoder reliability with Krippendorff's Alpha-Reliability [35], which supports multiple labels and multiple observers in an annotation task. We compute the alpha value using the Python implementation in "*nltk.metrics.agreement*" [36]. Finally, we obtain the intercoder reliability value of 76.8%.

TABLE 3
Distribution of the Visualization Sub-Types

Sub-type	#bbox	#img	Sub-type	#bbox	#img.
bar chart	5053	2058	pie chart	371	153
scatterplot	4269	1754	PAC	288	130
graph	3722	1615	box plot	277	147
heatmap	3202	1187	unit visualization	275	107
line chart	3004	1300	sunburst/icicle	260	120
table	2172	1676	sankey diagram	260	147
map	2106	986	stripe graph	239	123
matrix	1611	656	HEB	185	61
tree	1292	667	chord diagram	128	72
area chart	1125	527	polar plot	123	56
flow diagram	1118	873	storyline	46	25
PCP	975	541	contour graph	16	12
error bar	709	342	surface graph	13	7
treemap	554	268	word tree	9	9
glyph-based	523	259	phrase net	7	7
word cloud	392	184	Venn Diagram	4	4
donut chart	376	143	vector graph	4	2
others	-	1978			

Finally, we found 10,289 out of 12,267 images were assigned labels of visualization sub-types. We investigated the rest of images and discovered that these images are usually the ones with low resolution, the ones explaining methods, models, or algorithms, the ones of photos, 3D rendering, or artistic work, and the ones with pure text. Therefore, we assigned the rest with a label of "others." The distribution of each sub-type is shown in Table 3.

4.2 Bounding Box Annotation

With the specified visualization sub-types in each image, we further annotated bounding boxes (i.e., the positions in the images) for these visualizations. To improve efficiency, we employed crowd workers from a professional data annotation company, who are well-trained for drawing bounding boxes for machine learning tasks.

Criteria. Our criteria are based on the composition of the visualizations, i.e., visualization with coordinates or without coordinates. For a visualization with coordinates, the bounding box should cover all components of the coordinates, e.g., axis name, axis labels, chart title, and legends, if they are close to the main bodies of visual representations (Fig. 4b). If multiple sub-types share the same coordinate (e.g., error bar & bar chart in Fig. 4b), the area of their bounding boxes should be the same. For the visualizations without coordinates, we identify two situations, i.e., 1) independent visualizations without any connection or overlapping with other visualizations and 2) the visualizations connected to or overlapped with other visualizations. For the first case, the contents are the visualization itself (Fig. 4a1). For the second case, we only focus on the contents of the requested subtype. For example, the tree in Fig. 4c is connected to the Sankey diagram, and the word cloud in Fig. 4d overlays on the

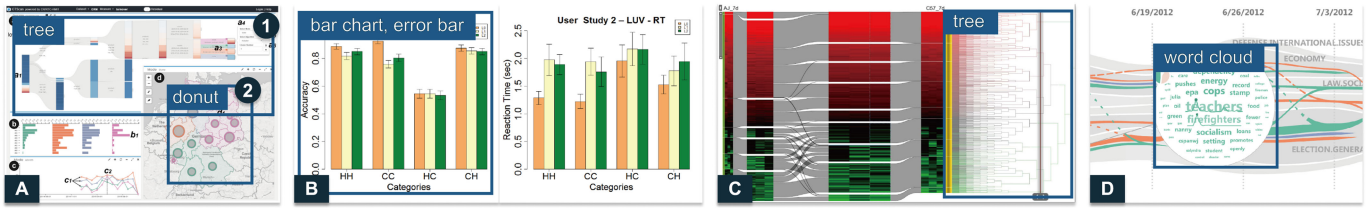


Fig. 4. Criteria for bounding box drawing. (a) shows how to draw bounding boxes for an independent visualization (a1) and multiple identical visualization sub-types (a2) [8]. (b) shows the bounding boxes of multiple sub-types (bar charts and error bars) which are included in the same coordinate. (c) shows how to draw bounding boxes for visualizations (tree) which are closely connected to other visualizations [9]. (d) shows the bounding box of the target visualization (word clouds) which is overlaid on another visualization [37]

area charts. The bounding boxes should only cover the contents of the tree and word cloud, respectively. In addition, there is an exception that requires further specification. Some visualizations contain multiple smaller visualizations of identical sub-type (e.g., the donut charts in the map in Fig. 4a2). In this case, when the number is larger than 10, we annotate them integrally with a single box with a label of "small multiples."

Procedure. The annotation procedure consisted of a training session and a formal annotation. To reduce the training load, each crowd worker was asked to focus on only one sub-type. Therefore, in the training session, a crowd worker was introduced one specific visualization sub-type, including the definition, examples, and the above annotation criteria. After that, the crowd workers were asked to take a test to ensure that they understood the sub-type and requirements. Only crowd workers who passed the test proceeded to the formal annotation. They were then assigned images containing specific visualization sub-types and required to draw the bounding boxes for this type of visualization. During the annotation, sampling tests were adopted to ensure quality.

Quality Measurement & Control. We evaluated the correctness of bounding boxes and tasks to control the annotation quality. The correctness of a bounding box was measured by intersection over union [38] (IoU, illustrated in Fig. 5a) with the ground-truth bounding box. Only when the IoU of the bounding box and the ground truth is higher than 0.9, the bounding box was accepted. Besides, the quality of a series of tasks was measured by the F_1 score, a metric balancing the recall and precision. The calculation of recall, precision, and F_1 score is presented in Fig. 5b.

To ensure quality, we adopted a sampling test on both batch level and worker level. We divided the 10,289 images equally into five batches and performed annotations batch by batch. The batch-level sampling test was performed after completing a batch of annotations. We randomly sampled 10% of the results and evaluated the F_1 score. If the F_1 was lower than 95%, the whole batch of annotation would be rejected. The rejected batch would be annotated again until the F_1 score reached 95%. The worker-level sampling test was conducted during one batch of annotations, where 15% annotations of a worker would be randomly sampled for F_1 score evaluation. If the F_1 was lower than 95%, all finished tasks of this worker in this batch would be rejected and annotated again. For the workers who failed the sampling test, their sampling rate would increase by 5% at the next test. Each accepted bounding box was paid with 0.03$.

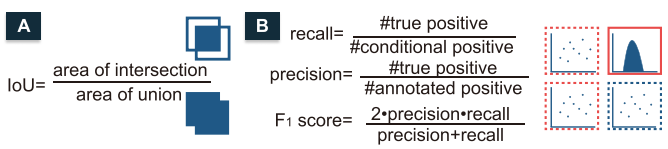


Fig. 5. Quality measurement for bounding box annotation. (a) shows how to compute the IoU of two bounding boxes. (b) shows the computation of recall, precision, and F_1 score. Red boxes represent the boxes labeled by crowd workers (annotated positive), dotted boxes represent the ground truth (condition positive), and red-dotted boxes are the ground truth correctly labeled by the crowd workers (true positive).

5 VISIMAGES

In this section, we present an overview of VisImages data and compare the distribution of visualizations in VisImages with that from other sources [12].

5.1 Overview of the Data

VisImages contains 12,267 images from 22-year VAST and InfoVis publications with 12,057 captions and 35,096 visualization bounding boxes. Table 3 shows the numbers of images (#img) and bounding boxes (#bbox) of each sub-type. For the frequent types, we observe that the number of bounding boxes of some sub-types (e.g., bar chart and scatterplot) is about two times more than the number of images. That is, multiple instances of these sub-types appear in one image simultaneously. For bar charts and scatterplots, the reason might be that they are basic charts and commonly serve as units of small multiples (e.g., scatterplot matrix). On the contrary, tables and flow diagrams have similar numbers of bounding boxes and images. We find that they usually occupy the entire image, since the tables are used independently to show the results of experiments or studies, and the flow diagrams are used to show the pipeline or framework of the methods.

Fig. 6 depicts the distribution of each sub-type from 1996 to 2018 using horizon charts, with color darkness encoding the number of bounding boxes. Several visualization sub-types are becoming increasingly popular, such as bar chart, area chart, scatterplot, matrix, line chart and heatmap (Fig. 6a). We notice that the dark area of graph visualizations distributes evenly across years (Fig. 6b), indicating that graph visualization has long been frequently used in the visualization community [39], [40]. Similarly, tables have always been a common visualization type in publications (Fig. 6b). Besides, we observe that the area of treemap becomes abruptly larger in 2005 while the area of tree reaches a peak in 2003 and 2005 (Fig. 6d). The increase of tree and treemap visualizations implies a more popular investigation

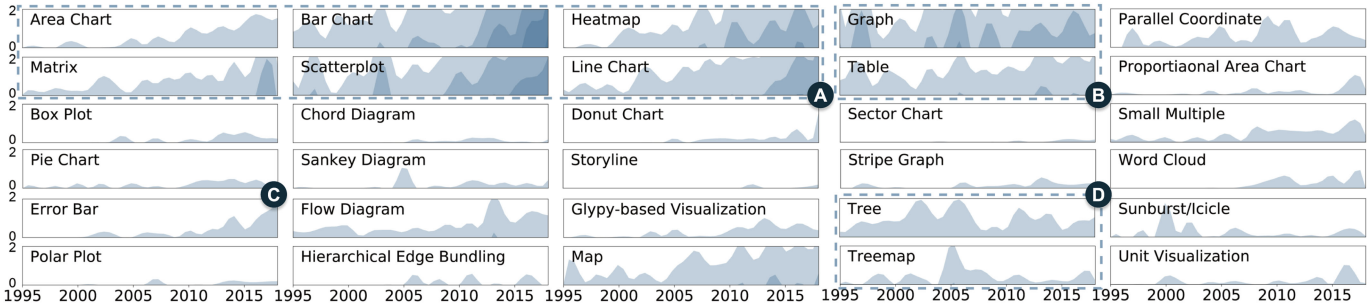


Fig. 6. Horizon charts demonstrate the average numbers of visualization bounding boxes in a paper over years.

in hierarchical data [41]. Moreover, the number of error bars continuously increases in recent years (Fig. 6c), indicating that statistical analysis on the error is increasing, such as user studies and model experiments.

5.2 Comparing Visualizations in Different Fields

Furthermore, we analyze the visualization distribution in academic visualization publications and materials that target to general audience. We use the statistics in MassVis [12] for comparison (Table 4), which collects images from four different sources, i.e., scientific publications (Nature), infographics, news, and government & world organization.

The comparison can be conducted directly because the taxonomies are the same at the category level, as shown in Table 2. From Table 4, we notice that the distribution in visualization publications is more balanced compared to the others. *Graph and Tree* occupy the largest share in visualization publications, which do not frequently appear in other sources. The reason might be that a quantity of research in our community focuses on visualizing hierarchical or network data. On the other hand, news media and government & world organizations prefer basic visual representations such as *Bar*, *Table*, and *Line* because the data they mostly present is relatively simple and in the form of tabular. Scientific papers prefer *Diagram*, *Line*, and *Point* for the presentation of methodology and experiment results. We notice that *Text*, which includes word clouds, word trees, and phrase nets, accounts for a portion in visualization publications but rarely appears in other sources. A lot of visualization research investigates variations of word cloud to make it more informative and effective, such as ManiWordle [42] and dynamic word cloud [43]. However, in public, given that the most commonly used media is text, the authors might expect to use graphical elements other than text to improve the expressiveness.

TABLE 4
The Distribution of Visualizations in VisImages and MassVis

Source	VisImages	MassVis			
		Scientific	Infographics	News	Government
Area	4.0%	1.9%	4.4%	4.4%	3.5%
Bar	12.1%	6.4%	5.9%	40.2%	36.9%
Circle	1.8%	0.3%	4.7%	1.3%	6.6%
Diag.	6.3%	27.4%	30.6%	7.2%	5.0%
Stat.	3.7%	3.2%	0.3%	0.3%	1.3%
Table	9.8%	8.3%	32.8%	8.2%	21.5%
Line	11.2%	19.1%	1.6%	19.1%	12.9%
Map	6.0%	9.2%	9.1%	13.5%	7.3%
Point	10.6%	16.6%	2.8%	5.0%	0.5%
Grid	7.2%	2.5%	1.9%	0%	0%
Text	1.1%	0%	0%	0.5%	0%
Graph	16.7%	5.1%	5.9%	0.3%	0%
Special	9.5%				
#Vis.	10,289	348	490	704	528

6 Use Cases

We present three use cases in this section to show the usefulness of VisImages. Specifically, the first case demonstrates the use of all metadata and annotations, the second case demonstrates the use of visualization type data, and the third case demonstrates the use of bounding box data.

6.1 Investigating the Use of Visualizations

VisImages contains rich information from IEEE VIS publications, including images, visualization types and their bounding boxes, captions, and publication metadata. The information can be used to understand the visual designs and the papers, which might be useful for junior visualization researchers and novice visualization designers. Therefore, to assist these users in data exploration, we develop *VisImages Explorer* for efficient data filtering and searching.

VisImages Explorer consists of a paper search panel, an image gallery view, a visualization distribution view, and a word cloud view. The *paper search panel* (Fig. 8a) allows users to filter papers by titles, years, conferences, and authors. A histogram is displayed to show the number of publications from different conferences across years. The *image gallery view* (Fig. 8b) then exhibits all images based on the paper searching results. Users are allowed to further filter the images by visualization types (Fig. 8b1) and keywords in the captions. Users can examine the annotations of each image in a detailed view by clicking the image. Moreover, the explorer has a *visualization distribution view* (Fig. 8c) showing the number of visualizations and a *word cloud view* (Fig. 8d) visualizing the word frequencies in the captions of the filtered images. To demonstrate the usefulness of VisImages, we present two scenarios for junior visualization researchers and novice visualization designers.

Scenario 1: Reviewing Papers of Graph Visualizations. Suppose James, a junior PhD student, is surveying papers about graph visualizations in IEEE VIS to understand the field. Papers might not explicitly mention "graph" in the title, so he decides to search the images directly. He turns to the image gallery view and filters the images whose captions contain the keyword of "graph" (Fig. 7a2). For the reason

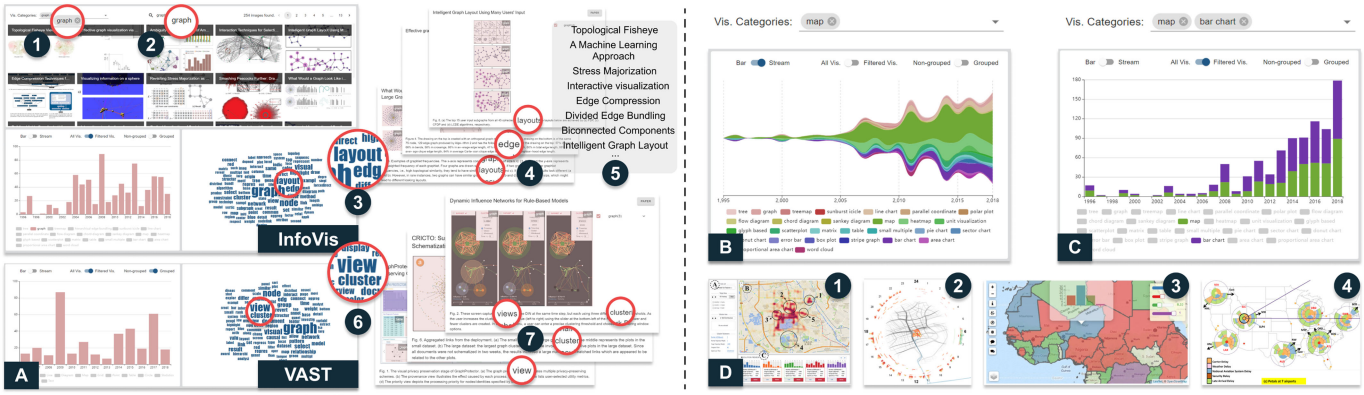


Fig. 7. Usage scenarios for junior visualization researchers and novice visualization designers: reviewing papers of graph visualizations (a) and inspiring visual designs (b, c, d).

that graph data can be represented with a matrix or node-link diagram, he first filters the images that contain the sub-type of "matrix," but only retrieves 58 images. Then he filters the images with the sub-type of "node-link diagram" (denoted as "graph" in this work, as shown in Fig. 7a1) and retrieved 385 images, which means that node-link diagram might be a much more popular representation type for graph data. Due to different research targets of VAST and InfoVis, James further investigates the images of two conferences respectively. He discovers that the top words in the word cloud views of two conferences are different. In addition to the word "graph" and "visual," the top words of InfoVis are "layout" and "edge" (Fig. 7a3) while the top words of VAST are "view" and "cluster" (Fig. 7a6).

When exploring the captions of InfoVis in depth, James discovers that the images are mostly about the resulted graph visualizations after applying layout optimization algorithms (Fig. 7a4). Clicking on the word "layout" in the word cloud view, the images are further filtered if their captions contain the word. He immediately discovers that the titles of these papers contain keywords describing the algorithms, such as "topological fisheye," "stress majorization," and "divided edge bundling" (Fig. 7a5). Therefore, he obtains a paper collection about graph layout optimization for further reading, which is a good starting point for his research career.

James further explores the images in VAST, where the word "view" and "cluster" frequently occur in the captions. Different from InfoVis, graph visualizations in VAST images usually appear in the views of visual analytics (VA) systems. As a result, the words "view" commonly co-occurs with the word "graph" in the captions. Besides, the "cluster" patterns of graphs are the patterns described the most in the VA systems with graphs. He then clicks on the word "cluster" in the word cloud view and investigates the captions. From the captions, he understands that clusters imply patterns of interest and help to guide users for further exploration.

Scenario 2: Inspiring Visual Designs. Suppose Mary, a junior visualization designer, is designing a map visualization for geospatial data. She wonders if there are any design strategies to improve the map visualization to represent more data attributes. She first filters the images that contain a map. The stream graph in the visualization distribution view indicates that heatmaps and bar charts are the most common sub-types that co-occur with maps (Fig. 7b). It is intuitive that heatmaps can visualize density data on maps, but Mary has no clear idea how bar charts can display together with maps. Therefore, she additionally selects bar charts, and the images containing both maps and bar charts are filtered. The histogram shows the distribution of bar charts and maps in the filtered images (Fig. 7c). By investigating the images, she discovers that bar charts and maps can be organized without overlapping, such as positioning them side-by-side (Fig. 7d1) or surrounding the map with circular bar charts (Fig. 7d2). When there is overlapping, the bar charts can be tooltips floating on the map (Fig. 7d3) or glyphs positioned at the places of interest (Fig. 7d4). Mary is inspired by the designs in the retrieved images and decides on a suitable one depending on design requirements.

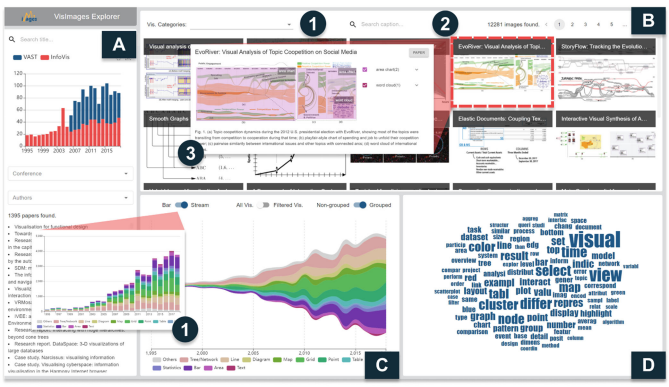


Fig. 8. VisImages Explorer. (a) is a paper search panel filtering papers by title, year, conference, and author. (b) is an image gallery view displaying the images in the filtered papers. The gallery is facilitated with a visualization selector (b1) and a caption searcher (b2) for further filtering of the images. Users can click the images of interest to view the detailed annotations (b3). (c) is the visualization distribution view showing the numbers of visualizations. Users can switch between a stream graph and a bar chart for different preferences (c1). (d) is a word cloud view exhibiting the word frequencies in the captions.

6.2 Classification Benchmarking With VisImages

Object classification has been adopted in many visualization scenarios, such as visualization reverse-engineering [18], [44], [45], visualization demographic analysis [3], and chart data extraction [17]. In this case, we show how VisImages can serve as a benchmark for visualization classification models using the annotations of visualization sub-types.

TABLE 5
The Top-1 and Top-3 Accuracies (%) of Different Models on Visualization Classification Under Different Situations

Training Set	(A) Beagles		(B) VisImages	
Test Set	Beagle	VisImages (Acc↓)	VisImages	Beagle (Acc↓)
ResNet-50	79.3/99.0	32.6/38.8	78.3/92.6	**10.6/9.5**
ResNet-101	80.6/98.9	26.0/33.2	78.9/94.4	**11.5/9.4**
VGG-16	79.9/98.7	36.7/43.8	77.5/92.7	**8.7/7.1**
VGG-19	80.1/98.7	34.8/40.4	76.9/92.2	**8.3/6.0**

The underlined numbers denote the highest accuracies among the models.

6.2.1 Experiment Setup

To compare VisImages and other datasets in training classification models, we set up experiments to mutually evaluate the models trained on different datasets. Specifically, we train the models on one dataset and evaluate them on another, and investigate their performance in different situations. We select Beagle [3] as a baseline dataset to train classification models because Beagle has the most classes in common with VisImages and the largest sample number and class number among existing datasets.

Data Processing. We select 17 common classes from VisImages and Beagle. We convert the images in Beagle into bitmap images for model training and evaluation since they are in the SVG format originally. For the images in VisImages, we crop the visualizations from the images and categorized them by sub-types (denoted as VisImages-cropped). Before the experiment, we randomly divide Beagle and VisImages into training (75%) and test (25%) sets.

Models. In the experiments, we select two widely-used object classification models with different numbers of layers, i.e., ResNet-50, ResNet-101 [46], VGG-16, and VGG-19 [47]. Taking an image as input, the models will output the probabilities of specific classes the image belongs to.

Training. We follow a similar training process described by Krizhevsky *et al.* [48] that all models are trained in two stages with weights pre-trained on ImageNet [49]. In the first stage, we freeze the weights of convolutional layers and train the classification heads. In the second stage, we unfreeze the convolutional layers and finetune the overall weights. In each stage, the models are trained with stochastic gradient descent (SGD) with $1e^5$ steps. The initial learning rates of ResNets and VGGNets are $1e^{-3}$ and $1e^{-5}$, respectively. We use categorical cross-entropy as loss function. Because of the imbalanced distribution between different classes, we introduce class weights for loss computation to reduce over-fitting on specific classes.

Metrics. Following the standard evaluation protocol of multi-classification problems [46], we use the top-k accuracy as a metric, which means, if the ground-truth label appears in the top k predictions, the classification is regarded correct. The top-1 and top-3 accuracies are shown in Table 5.

6.2.2 Classification Performance Analysis

We analyze the model performance from different aspects.

CNN Model Comparison. We first observe the performance of different models, and discover that ResNet-101 achieves the best performance on VisImages and Beagle during training (underlined values in the second row). The results conform to the conclusion drawn by He *et al.* [46] that deeper architectures of ResNets make the models achieve better results than VGGNets in object classification.

Although CNNs achieve satisfactory performance on classifying bitmap visualizations, there might be some setbacks of the models when compared to other methods. Battle *et al.* [3] classified the visualizations in Beagle with decision trees and achieved 86.5% top-1 accuracy, which is better than the performance of CNNs. Differently, the decision trees take in hand-crafted features of SVG visualizations, instead of visual features extracted from bitmap images. Specifically, the features include style features (e.g., number of fill and border colors and stroke width) and per-element features (e.g., CSS class names, circle elements, and rect elements). These features might provide semantic information about the visualizations that are useful for visualization classification. Therefore, there might be some limitations of CNNs for visualization classification, which conforms to the insights provided by Haehn *et al.* [50] that "*CNN architecture performance on natural images is not a good predictor for performance on graphical perception tasks.*"

Training Dataset Comparison. Training data is critical to model generalizability. Overall, we discover that models trained on Beagle have higher accuracies than the models trained on VisImages. However, the comparison is not fair because the test sets are not the same. Therefore, we conduct cross-evaluation by training the models on one dataset but testing them on the other one. First, we evaluate the models trained on Beagle with VisImages. The models encounter a steep decrease on both top-1 (26%-34.8%) and top-3 (33.2%-40.4%) accuracies when tested on VisImages (Table 5 a). Second, we evaluate the models trained on VisImages on Beagle and discover that models also encounter a decrease on the top-1 (8.3%-11.5%) and top-3 (6.0%-9.5%) accuracies (Table 5 b). However, the decrease is much smaller compared to the previous ones. The model performance might decrease when testing the model on a dataset with a different "data generating process" because of the generalization error [51]. The inhibition of model performance decrease indicates that models trained on VisImages might have better generalizability compared to Beagle.

Confusion Analysis. To further investigate the reason for the performance decrease of the models trained on Beagle when testing on VisImages, we visualize the confusion matrix of ResNet-101, which achieves the best performance (Fig. 9, left). Rows of the matrix encode the ground-truth labels and columns encode the prediction labels. We arrange the row and column labels by the sample number of classes in Beagle decreasingly. The matrix is normalized row-wise, and thus the cells on the diagonal represent the recall rates.

In the matrix, while most cells on the diagonal are in dark blue, there are some cells in light colors, such as heatmap, Sankey diagram, and word cloud, indicating low recall rates for these classes. By randomly selecting the samples of these classes from both datasets, we discover that the visualizations of these classes in VisImages are more diverse in appearance, such as layout, color, and shape, but the ones in Beagle are similar in design, as shown in Fig. 9, right. The reason might be that the charts in Beagle are generated by similar visualization libraries (e.g., D3 and Plotly) with

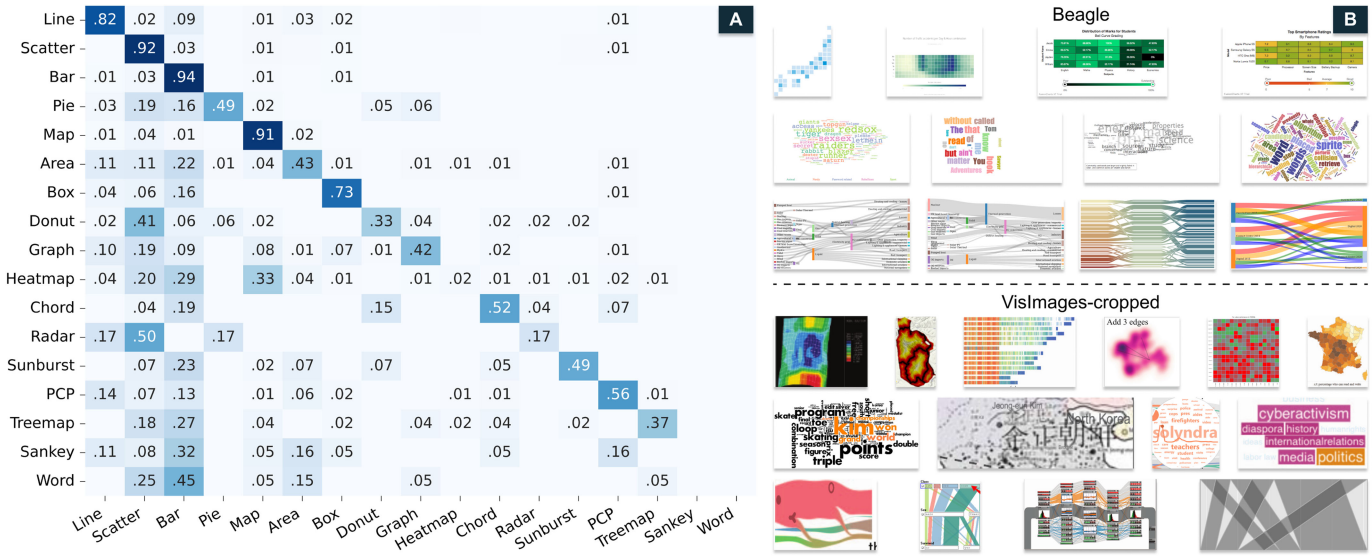


Fig. 9. (A) Confusion matrix of ResNet-101 trained from Beagle and tested on VisImages (on the left). Rows of matrix represent the ground truth, and columns represent the predictions. The confusion matrix is normalized by rows. (B) Examples of heatmap, word cloud, and Sankey diagram from Beagle and VisImages-cropped, respectively.

default settings of styles. However, in VisImages, many of charts in VisImages are created from scratch using design tools (e.g., Adobe Illustrator) or low-level programming languages (e.g., Javascript). Due to the higher diversity in layout and design of the samples, VisImages can be a good benchmark and complementary for existing datasets.

In addition, we discover that there are blue clusters in the bottom-left of the confusion matrix, especially the first three columns. The clusters indicate that the model usually misclassifies the visualizations to be line chart, scatterplot, and bar chart. We infer that the confusion might be caused by the large numbers of samples of these three classes (about 90%), even though we have introduced class weights to relieve the over-fitting on these classes. Besides, many of the test samples in VisImages might have different styles that are not "seen" by the models during training. Therefore, when these samples occur, the model might tend to "guess" their labels with popular classes. To reduce the confusion, balancing the class distribution and adding more samples to smaller classes might be a practical solution, such as combining Beagle and VisImages as a training set.

6.3 Visualization Localization With VisImages

While VIS30K [21] and Viziometrics [20] only focus on the classification of the image usage in the papers (e.g., equation, diagram, photo, and plot), VisImages further specifies the positions (i.e., bounding boxes) of visualizations in the images. In this case, we exhibit the use of visualization bounding boxes, a featured dimension of VisImages dataset that can be used to train object localization models. Specifically, the models can be used to localize visualizations from the images for reverse engineering of visual analytics (VA) systems and visualization position analysis.

6.3.1 Training Visualization Localization Models

We first train Faster R-CNN [24], one of the most widely-used object localization models, to predict the position of visualizations in the images. We used 80% of images for training and 20% for testing. Following a similar pipeline of Ren *et al.* [24], we train the model using SGD optimizer with a learning rate of $1e^{-3}$ for 15k mini-batches. A momentum of 0.9 and a weight decay of $1e^{-4}$ are adopted.

We use average precision (AP) to evaluate the model performance on object detection [52]. Besides, the IoU is used to measure the overlap between a predicted box and a ground-truth box. For a detailed definition of IoU, please refer to Section 4.2 and Fig. 5a. Table 6 shows the APs of different visualization sub-types, as well as mean average precision (mAP) under different IoU thresholds.

6.3.2 Model Inference on VA Systems

After training visualization localization models, we further apply the models on VA system interfaces. We use the images from the Multiple-View Visualization (MV) dataset [22], which contains 360 VA interfaces cropped from the publication figures. Samples of inference results are exhibited in Fig. 10. Overall, we discover that Faster-RCNN trained with VisImages can successfully localize different visualizations (or views) in the interface. In addition, the model can also handle various challenging cases: 1) visualizations with sub-structures, e.g., graph with sub-graphs

TABLE 6
APs Under Different IoU Thresholds

Sub-type	$AP_{IoU=0.50}$	$AP_{IoU=0.75}$	$AP_{IoU=0.90}$
graph	**0.96**	**0.96**	**0.70**
table	0.95	0.90	0.69
scatterplot	0.83	0.77	0.29
line chart	0.82	0.71	0.23
heatmap	0.80	0.80	0.40
flow diagram	0.75	0.70	0.46
bar chart	0.69	0.58	0.10
map	0.68	0.64	0.54
parallel coordinate	0.67	0.52	0.28
mAP	0.78	0.78	0.52

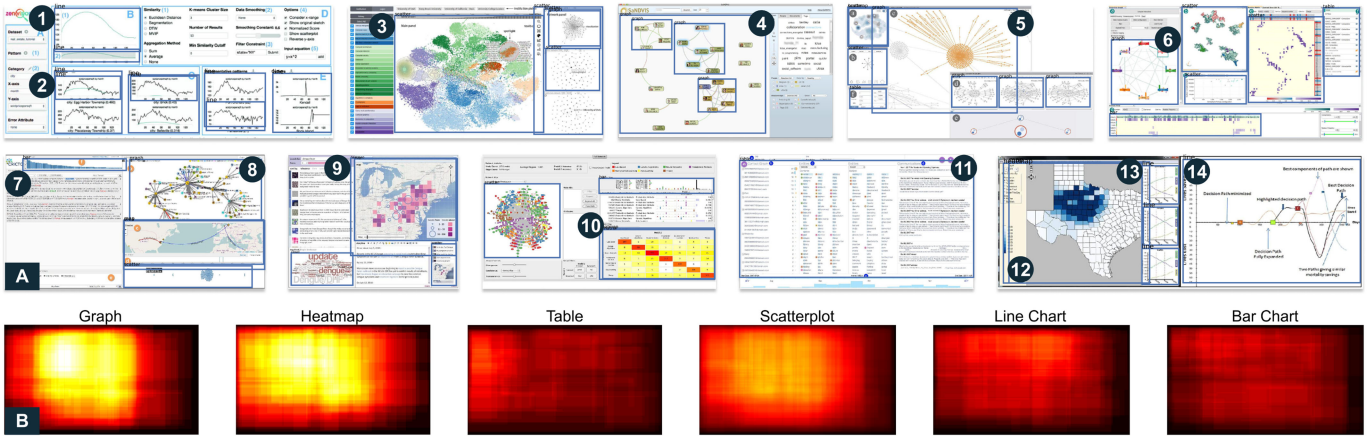


Fig. 10. (a) Results of visualization localization in visual analytics systems and (b) heatmaps showing the spatial distributions of the visualizations.

(Fig. 10a4); 2) visualizations with composite sub-types, e.g., heatmap + map (Figs. 10a9 and 10a13) and heatmap + matrix (Fig. 10a10); and 3) visualizations with extreme aspect ratios (Fig. 10a7). The results demonstrate the potential of using computer vision models for reverse engineering the VA system designs in the publications, by considering the data flow and interactions among views [53], [54], [55].

6.3.3 Localization Performance Analysis

We analyze the localization performance of the model combining the APs and samples of VA systems.

Overall, the mAPs on VisImages are 0.78 (IoU=0.50), 0.78 (IoU=0.75), and 0.52 (IoU=0.90). The graph achieves the highest AP (IoU=0.50) of 0.96, followed by table, scatterplot, and line chart. Interestingly, even though bar chart has the largest portions of samples, its AP is only ranked the 7th (IoU=0.50). When investigating in detail, the APs of basic charts (i.e., bar chart, line chart, and scatterplot, which are commonly facilitated with the coordinate systems) decrease drastically with the rise of IoU threshold, indicating that precisely determining the bounding box of these visualizations are challenging for the model. For example, when the same types of visualizations are aligned closely (Fig. 10a2), the model will fail to determine the area of the visualizations. By exploring the samples, we discover that these sub-types have diverse configurations (with or without the coordinate axes) and sizes, possibly making the model confused about the precise borders for these visualizations. Differently, graph (Figs. 10a4, 10a5, 10a6, and 10a8), table (Fig. 10a11), and map (Figs. 10a9 and 10a13) usually occupy a large area in the images, which might make the localization of these sub-types relatively simpler for the model.

6.3.4 Spatial Distribution of Visualizations in VA Systems

The model inference on VA systems can help us understand how different visualization types are distributed spatially. Although the inference might not be perfectly correct, we argue that the results can indicate general distributions of the visualizations because of the acceptable APs shown in Table 6. We visualize the spatial distribution of specific types of visualization using heatmaps. To plot a heatmap, we transform all VA system images to the same scale, derive the transformed bounding boxes of the visualizations, and overlay all bounding boxes on the same canvas. The heatmap is normalized by the total number of the bounding boxes of this type. To facilitate comparison, we use a consistent brightness scale for different visualization types. In Fig. 10b, we visualize the heatmaps of the most popular visualizations in the VA samples.

From the figure, we discover that some heatmaps have extremely bright areas, i.e., the graph and heatmap visualizations. Therefore, graph and heatmap visualizations have a higher density of visualizations in the bright area compared to other visualizations. In addition, these visualizations are commonly distributed in the top-left of the VA systems, usually as the main view of the system. On the contrary, the table, scatterplot, line chart, and bar chart visualizations do not have an extreme concentration, but these visualizations also have specific distribution patterns. For example, table visualizations appear more on the left side, possibly serving as an auxiliary view for reference of raw data; line charts are more distributed on the top; scatterplots are also distributed more on the top-left; bar charts, the most commonly used visualizations, exhibit a relatively even distribution in the interface.

7 DISCUSSION

We see VisImages as an exciting start point for leveraging the intelligence of the visualization community itself and forge a path to a high-quality, fine-grained, and large-scale visualization dataset. We envision that VisImages can inform opportunities for advancing our knowledge of the field and the research of AI4VIS [1].

Benefits to Literature Analysis. VisImages offers new possibilities to conduct literature analysis in visualization and help understand the evolution of the field. For example, VisImages Explorer (Section 6.1) shows the potential to help users discover the papers of interest and inspire design ideas for novice researchers and designers. We make the explorer available for the community to discover more insights. Furthermore, researchers can investigate VisImages combined with existing publication metadata collections (e.g., keyvis [33]), which is supposed to reveal new insights. For example, analyzing the visual design patterns under different domains and topics.

Opportunities for AI4VIS. VisImages can serve as a useful AI4VIS dataset [1], such as visualization classification (Section 6.2), localization (Section 6.3), visualization-text translation, and recommendation. Visualization-text translation aims to construct relations between textual descriptions and visualizations. Images with elaborated captions in VisImages can naturally be a qualified resource for training such models for the scenarios like visual storytelling [56], [57], [58]. Moreover, a line of research puts efforts on visualization recommendation [59], [60], for example, suggesting potential layouts and combinations of different visualization types in a design [22]. VisImages contains images of well-crafted VA systems for different topics, such as sports analysis [61] and urban planning [62]. Along with annotations of visualization types and positions, the dataset provides a valuable resource for training models on such tasks.

Limitations. Despite the significance and usefulness of VisImages, it still has limitations. First, we tried best to ensure the quality of our annotation with a series of measures, such as the gold standards, majority voting, and sampling test. Mislabeling is inevitable, especially in the situation where recognizing visualization and their variations requires significant expertise. As an alternative, we greatly welcome the visualization community, especially the authors of the publications, to examine and possibly correct the mislabeled visualizations. Second, our dataset currently contains three major types of labels (i.e., image captions, visualization types, and bounding boxes), leaving a wealth of information unexplored, such as axis titles and marks. With richer information, VisImages can be used in a wider range of applications and support more complex tasks. Third, our exemplar use cases only demonstrate featured usage of VisImages. For example, in Section 6.2, we first compare VisImages with Beagle, as Beagle has the most classes in common and the largest samples. Further comparison with more datasets (e.g., ChartSense [17] and VizioMetrics [20]) may also benefit the analysis of low-level and high-level features, as well as the model performance. The exploration on extensive datasets is beyond the scope of the paper, but our use cases provide pointers to designing experiments and conducting analysis for future work. A more in-depth investigation of VisImages can offer illuminating insights and reach broad impacts.

8 Conclusion and Future Work

In this paper, we create and make available VisImages, a visualization dataset from the top-venue visualization publications. VisImages includes 12,267 images with captions from 1,397 papers of IEEE InfoVis and VAST. Each image is annotated with visualization types and their positions in the image, resulting in a total of 35,096 bounding boxes. We further investigate VisImages with an overview of visualization distribution across years and types. Besides, VisImages presents a more balanced distribution in the visualization types, compared to other state-of-the-art datasets in the visualization field. The usefulness and significance of VisImages are demonstrated through three use cases, including visual literature review, visualization classification, and visualization localization. We envision that VisImages can broaden the diversity of visualization research [63] and inspire new research opportunities.

However, VisImages only takes a first step to explore images in the visualization publications. In the future, we intend to expand VisImages to cover more images from other top-notch journals and conferences, such as TVCG, CHI, and EuroVis. Second, given the increasing number of images, we plan to develop a semi-automatic annotation method that leverages human and machine intelligence [64]. Third, we plan to gradually refine and improve the taxonomy to meet the growing diversity of visualization designs.

Acknowledgments

Our deepest gratitude went to the anonymous reviewers for their valuable comments that helped us improve this paper substantially. We sincerely thank the researchers and the students from Zhejiang University and Zhejiang Lab for their time and effort in data annotation. Specifically, We thank Mengye Xu for her contributions in designing the VisImages logo and documenting the training materials for data annotation.

References

[1] A. Wu *et al.*, "AI4VIS: Survey on artificial intelligence approaches for data visualization," *IEEE Trans. Vis. Comput. Graphics*, to be published, doi: 10.1109/TVCG.2021.3099002.
[2] P. Federico, F. Heimerl, S. Koch, and S. Miksch, "A survey on visual approaches for analyzing scientific literature and patents," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 9, pp. 2179–2198, Sep. 2017.
[3] L. Battle, P. Duan, Z. Miranda, D. Mukusheva, R. Chang, and M. Stonebraker, "Beagle: Automated extraction and interpretation of visualizations from the web," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–8.
[4] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer, "ReVision: Automated classification, analysis and redesign of chart images," in *Proc. ACM Symp. User Interface Softw. Technol.*, 2011, pp. 393–402.
[5] N. Siegel, Z. Horvitz, R. Levin, S. Divvala, and A. Farhadi, "FigureSeer: Parsing result-figures in research papers," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 664–680.
[6] Y. Wu *et al.*, "OpinionSeer: Interactive visualization of hotel customer feedback," *IEEE Trans. Vis. Comput. Graphics*, vol. 16, no. 6, pp. 1109–1118, Nov./Dec. 2010.
[7] D. A. Keim, M. C. Hao, U. Dayal, and M. Hsu, "Pixel Bar Charts: A visualization technique for very large multi-attribute data sets," *Informat. Vis.*, vol. 1, no. 1, pp. 20–34, 2002.
[8] D. Liu, P. Xu, and L. Ren, "TPFlow: Progressive partition and multidimensional pattern extraction for large-scale spatio-temporal data analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 1–11, Jan. 2019.
[9] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg, "Comparative analysis of multidimensional, quantitative data," *IEEE Trans. Vis. Comput. Graphics*, vol. 16, no. 6, pp. 1027–1035, Nov./Dec. 2010.
[10] P. Isenberg *et al.*, "Vispubdata.org: A metadata collection about IEEE visualization (VIS) publications," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 9, pp. 2199–2206, Sep. 2017.
[11] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proc. IEEE Symp. Vis. Lang.*, 1996, pp. 336–343.
[12] M. A. Borkin *et al.*, "What makes a visualization memorable?," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2306–2315, Dec. 2013.
[13] W. Javed and N. Elmqvist, "Exploring the design space of composite visualization," in *Proc. IEEE Pacific Vis. Symp.*, 2012, pp. 1–8.
[14] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Proc. AAAI Conf. Artif. Intell. Workshop*, 2012, pp. 40–46.
[15] B. Michael, O. Vadim, and J. Heer, "D3: Data-driven documents," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011.

[16] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer, "Vega-Lite: A grammar of interactive graphics," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 341–350, Jan. 2017.

[17] D. Jung *et al.*, "ChartSense: Interactive data extraction from chart images," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 6706–6717.

[18] J. Poco and J. Heer, "Reverse-engineering visualizations: Recovering visual encodings from chart images," *Comput. Graph. Forum*, vol. 36, no. 3, pp. 353–363, 2017.

[19] M. A. Borkin *et al.*, "Beyond memorability: Visualization recognition and recall," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 519–528, Jan. 2016.

[20] P.-S. Lee, J. D. West, and B. Howe, "Viziometrics: Analyzing visual information in the scientific literature," *IEEE Trans. Big Data*, vol. 4, no. 1, pp. 117–129, Mar. 2018.

[21] J. Chen *et al.*, "VIS30K: A collection of figures and tables from IEEE visualization conference publications," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 9, pp. 3826–3833, Sep. 2021.

[22] X. Chen, W. Zeng, Y. Lin, H. M. AI-maneea, J. Roberts, and R. Chang, "Composition and configuration patterns in multiple-view visualizations," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1514–1524, Feb. 2021.

[23] R. Li and J. Chen, "Toward a deep understanding of what makes a scientific visualization memorable," in *Proc. IEEE Sci. Vis. Conf.*, 2018, pp. 26–31.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv: 1804.02767*.

[26] W. Zeng, A. Dong, X. Chen, and Z.-l. Cheng, "VIStory: Interactive storyboard for exploring visual information in scientific publications," *J. Vis.*, vol. 24, no. 1, pp. 69–84, 2021.

[27] J.-D. Fekete, G. Grinstein, and C. Plaisant, "InfoVis 2004 contest: The history of InfoVis," 2004. [Online]. Available: http://www.cs.umd.edu/hcil/iv04contest

[28] L. Xie, "Visualizing citation patterns of computer science conferences," Blog post: 2016. [Online]. Available: http://cm.cecs.anu.edu.au/post/citation_vis/

[29] C. Plaisant, J.-D. Fekete, and G. Grinstein, "Promoting insight-based evaluation of visualizations: From contest to benchmark repository," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 1, pp. 120–134, Jan./Feb. 2008.

[30] K. Cook, G. Grinstein, and M. Whiting, "Introduction: The VAST challenge: History, scope, and outcomes: An introduction to the special issue," *Informat. Vis.*, vol. 13, no. 4, pp. 301–312, 2014.

[31] A. Ponsard, F. Escalona, and T. Munzner, "PaperQuest: A visualization tool to support literature review," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, 2016, pp. 2264–2271.

[32] J. Chuang, S. Gupta, C. Manning, and J. Heer, "Topic model diagnostics: Assessing domain relevance via topical alignment," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 612–620.

[33] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, "Visualization as seen through its research paper keywords," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 771–780, Jan. 2017.

[34] C. Clark and S. Divvala, "PDFFigures 2.0: Mining figures from research papers," in *Proc. ACM/IEEE-CS Joint Conf. Digit. Libraries*, 2016, pp. 143–152.

[35] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011. [Online]. Available: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers

[36] "NLTK library: Nltk.metrics.agreement," Accessed: Oct. 28, 2021. [Online]. Available: https://www.nltk.org/_modules/nltk/metrics/agreement.html

[37] P. Xu *et al.*, "Visual analysis of topic competition on social media," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2012–2021, Dec. 2013.

[38] J. Redmon, S. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[39] D. Han, J. Pan, X. Zhao, and W. Chen, "NetV.js: A web-based library for high-efficiency visualization of large-scale graphs and networks," *Vis. Inform.*, vol. 5, no. 1, pp. 61–66, 2021.

[40] Y. Wang *et al.*, "G6: A web-based library for graph visualization," *Vis. Inform.*, vol. 5, no. 4, pp. 49–55, 2021.

[41] H. Schulz, "Treevis.net: A tree visualization reference," *IEEE Comput. Graph. Appl.*, vol. 31, no. 6, pp. 11–15, Nov./Dec. 2011.

[42] K. Koh, B. Lee, B. Kim, and J. Seo, "ManiWordle: Providing flexible control over wordle," *IEEE Trans. Vis. Comput. Graphics*, vol. 16, no. 6, pp. 1190–1197, Nov./Dec. 2010.

[43] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context preserving dynamic word cloud visualization," in *Proc. IEEE Pacific Vis. Symp.*, 2010, pp. 121–128.

[44] F. Zhou *et al.*, "Reverse-engineering bar charts using neural networks," *J. Visualization*, vol. 24, no. 2, pp. 419–435, 2021.

[45] L. Ying *et al.*, "GlyphCreator: Towards example-based automatic generation of circular glyphs," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 400–410, Jan. 2022.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Informat. Process. Syst.*, 2012, pp. 1097–1105.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[50] D. Haehn, J. Tompkin, and H. Pfister, "Evaluating 'graphical perception' with CNNs," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 641–650, Jan. 2019.

[51] Y. Bengio, I. Goodfellow, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT press, 2017, vol. 1.

[52] M. Everingham, S. M. Eslami, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[53] R. Chen, X. Shu, J. Chen, D. Weng, J. Tang, S. Fu, and Y. Wu, "Nebula: A coordinating grammar of graphics," *IEEE Trans. Vis. Comput. Graphics*, to be published, doi: 10.1109/TVCG.2021.3076222.

[54] C. Su *et al.*, "Natural multimodal interaction in immersive flow visualization," *Vis. Inform.*, vol. 5, no. 4, pp. 56–66, 2021.

[55] C. Tominski *et al.*, "Toward flexible visual analytics augmented through smooth display transitions," *Vis. Inform.*, vol. 5, no. 3, pp. 28–38, 2021.

[56] X. Shu *et al.*, "Dancingwords: Exploring animated word clouds to tell stories," *J. Visualization*, vol. 24, no. 1, pp. 85–100, 2021.

[57] X. Shu, A. Wu, J. Tang, B. Bach, Y. Wu, and H. Qu, "What makes a Data-GIF understandable?," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1492–1502, Feb. 2021.

[58] W. Zhang, Q. Ma, R. Pan, and W. Chen, "Visual storytelling of song ci and the poets in the social–cultural context of song dynasty," *Vis. Inform.*, vol. 5, no. 4, pp. 34–40, 2021.

[59] H. Li, Y. Wang, S. Zhang, Y. Song, and H. Qu, "KG4Vis: A knowledge graph-based approach for visualization recommendation," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 195–205, Jan. 2022.

[60] A. Wu *et al.*, "MultiVision: Designing analytical dashboards with deep learning based recommendation," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 162–172, Jan. 2022.

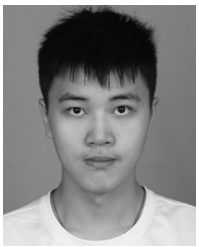
[61] J. Wang, J. Wu, A. Cao, Z. Zhou, H. Zhang, and Y. Wu, "TacMiner: Visual tactic mining for multiple table tennis matches," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 6, pp. 2770–2782, Jun. 2021.

[62] Z. Deng *et al.*, "Compass: Towards better causal analysis of urban time series," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 1051–1061, Jan. 2022.

[63] B. Lee *et al.*, "Broadening intellectual diversity in visualization research papers," *IEEE Comput. Graph. Appl.*, vol. 39, no. 4, pp. 78–85, Jul./Aug. 2019.

[64] D. Deng *et al.*, "EventAnchor: Reducing human interactions in event annotation of racket sports videos," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–13.

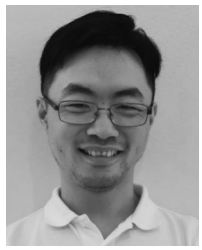
Dazhen Deng received the BE degree in applied mathematics from Zhejiang University in 2018. He is currently working toward the PhD degree with the State Key Lab of CAD&CG, Zhejiang University. His research interests mainly lie in sports visualization and machine learning for visual analytics. For more information, please visit https://dazhendeng.github.io/.



Yihong Wu received the BE degree from Zhejiang University in 2020. He is currently working toward the PhD degree with the State Key Lab of CAD&CG, Zhejiang University. His research interests mainly lie in computer vision and visual analytics.

Xinhuan Shu received the BE degree in computer science and technology from Zhejiang University, China and the PhD degree from HKUST. Her research interests include visual data communication, animated visualization, and visual analytics.

Jiang Wu received the BE degree in computer science and technology from Zhejiang University, China. He is currently working toward the PhD degree with the State Key Lab of CAD&CG, Zhejiang University. His research interests mainly lie in sports visualizations and event sequence analysis.



Siwei Fu received the PhD degree in computer science and engineering from the Hong Kong University of Science and Technology. He is an associate research scientist with Zhejiang Lab. His main research interests include visual analytics, intelligent user interface, and natural language interface. For more information, please visit https://fusiwei339.bitbucket.io/

Weiwei Cui received the BS degree in computer science and technology from Tsinghua University, China and the PhD degree in computer science and engineering from the Hong Kong University of Science and Technology. He is currently a principal researcher with Microsoft Research Asia, China. His primary research interests lie in visualization, with focuses on text, graph, and social media. For more information, please visit http://research.microsoft.com/en-us/um/people/weiweicu/

Yingcai Wu received the PhD degree in computer science from the Hong Kong University of Science and Technology. He is currently a professor with the State Key Lab of CAD&CG, Zhejiang University. His main research interests are in information visualization and visual analytics, with focuses on sports science and urban computing. Prior to his current position, he was a postdoctoral researcher with the University of California, Davis from 2010 to 2012, and a researcher with Microsoft Research Asia from 2012 to 2015. For more information, please visit http://www.ycwu.org.