VisImages: A Corpus of Visualizations in the Images of Visualization Publications

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

Abstract—Images in visualization publications contain rich information, e.g., novel visualization designs and common combinations 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. Based on a comprehensive taxonomy for visualizations in publications, the dataset includes 35,096 annotated visualizations, as well as their positions. We demonstrate the usefulness of VisImages through three use cases: 1) exploring and analyzing the evolution of visualizations with VisImages Explorer, 2) training and benchmarking models for visualization classification, and 3) localizing and recognizing visualizations in the images automatically.

Index Terms—visualization, dataset, images, object detection

1 INTRODUCTION

As the saying goes, a picture is worth a thousand words. Images are crucial in publications, especially in the data visualization field, which showcase the visual designs, system framework, model details, experiment results, etc. Building a dataset with images from visualization publications can benefit the visualization community in many aspects. For example, images contain a rich trove of visual information (e.g., color schemes and shape of glyph designs) and semantic information (e.g., different combinations of charts). Such information greatly complements the current literature analysis [1], [2], [3], [4], [5], [6], which mainly focuses on metadata like keywords, citation, and co-authorship [7]. Combining these metadata with images not only provides a comprehensive understanding of the field (e.g., what types of charts are frequently used in different venues across years; how different charts are used under different purposes?), but also inspires new research problems (e.g., how to select different charts and organize them to better illustrate the ideas?).

Moreover, an image dataset from visualization publications affords new opportunities for automated tasks in the visualization field. Existing studies [8], [9], [10] have collected chart images online and trained computer vision models for visualizations, such as chart classification. However, models trained from these datasets, whose images usually contain only a single chart, might fail when dealing with charts with uncommon designs, such as system interfaces with multiple charts arranged in complex layouts. Therefore, an image dataset from visualization publications can serve as a benchmark to test the generalizability and robustness of the models. Moreover, due to the lack of proper labeling, such as positions of charts and textual image descriptions, there is a gap to adopt more frontier and complex computer vision models (e.g., object detection, image generation by natural languages) to visualization tasks (e.g., deconstructing visual analytics systems and automatically generating visualizations). Fortunately, in visualization publications, the figures are naturally accompanied with captions, which have the potential to fill this gap with appropriate annotation.

In this paper, we build and make public an image dataset from visualization publications with comprehensive samples and rich annotations. We attempt to curate and examine the visualizations crafted by the academic community itself, which have a gulf of knowledge from visualization researchers. With the dataset, we hope to open up a wide range of significant applications such as literature analysis and automated tasks in the visualization community.

However, creating such a dataset faces three major challenges. The first challenge is categorizing diverse visualizations. Researchers have proposed various taxonomies to define and distinguish visualizations from the perspectives of data types (e.g., spatial, temporal, and hierarchical) [11], or graphical representations (e.g., point, line, and area) [12], [13]. However, these taxonomies cannot cover the various designs in visualization publications, which might be novel glyphs or the variations of existing visualizations (Fig. 1(A) and (B)). In addition, annotating images with visualization types requires extensive visualization expertise. Second, the layout of visualizations is diverse in the images of visualization publications, including single charts (Fig. 1(D)) and multiple-view visual analytics systems (Fig. 1(C)). Identifying the positions of visualizations would be ambiguous for these images, and needs massive efforts. Third, it is also challenging to ensure the quality of the annotations, given the lack of “the ground truth”.

To address the first challenge, we use the taxonomy proposed by Borkin et al. [12] and revise the categories according to the image data collected. The taxonomy covers most of the visualizations that appeared in visualization publications. Based on the taxonomy, we invite senior visualization researchers and graduate students to specify the visualization types in the images. For the second challenge, we set up a series of criteria for decomposing visualizations in the images and annotating their positions. Based on the criteria, we recruit trained crowd workers to annotate the bounding boxes for each visualization. To tackle the third challenge, we adopt a series of measures for quality control,
Our contributions are threefold. First, we build a dataset named VisImages, which contains 12,267 images collected from visualization publications. Each image is accompanied by a textual caption and a series of bounding boxes representing the positions and types of the visualizations contained. We release our dataset and all related tools for image data collection and processing on https://visimages.org/. Second, we present a panorama of VisImages and compare it with other datasets. From the analysis, we gain insights into the peculiarities of visualizations in academic publications. Third, 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 the state-of-the-art image classification models with VisImages, and 3) localizing and classifying visualizations in the images automatically with VisImages.

2 RELATED WORK

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

2.1 Image Datasets in Visualization

In this subsection, we will introduce existing image datasets in the visualization community and demonstrate that VisImages is, to our best knowledge, the most comprehensive one regarding quantity, chart complexity, chart diversity, and annotations.

The visualization community has built a variety of image datasets that are composed of basic charts (such as single bar charts or scatterplots). The visualizations 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 [18], Vega-Lite [19]). For example, Battle et al. [8] gathered over 41,000 SVG-based charts, manually labeled them under 24 visualization types, and trained classification models to analyze the chart type distribution on the web. Jung et al. [20] collected 5659 images consisting of ten chart types to develop models for chart classification and proposed ChartSense for chart data extraction. Savva et al. [9] delivered a dataset containing 2601 single-chart images in 10 categories. The dataset are used to develop a system called ReVision to redesign the charts for better visual styles. Similarly, Poco et al. [21] collected more than 5000 bitmap images and annotated the corresponding visualization types (area, line, bar, and scatter) and roles of the text content (labels and titles of axes and legend). The data is used for reverse-engineering and reconstructing visualizations with declarative grammars, such as Vega-Lite [19], Borkin et al. [12], [22] developed MassVis for memorability study. They collected more than 2000 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 on detailed information (e.g., annotations, axis, data, etc.).

Differently, the images in visualization publications contain more complex designs created by visualization experts, such as system interfaces, which are the interests many studies lay in. For example, Li et al. [23] collected images from SciVis and conducted user studies to understand the relation between memorability and image characteristics. Chen et al. [24] adapted object detection models (Faster-RCNN [25] and YOLOv3 [26]) to crop the figures and tables from visualization publications and proposed VISTory, a dataset of figures and tables. However, the detailed information of the visual designs, such as chart types, chart positions, and textual descriptions, are not considered in their dataset. Zeng et al. [27] collected figures from IEEE VIS and proposed a technique named VIStory to visualize and analyze the evolution of figures. Chen et al. [28] 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 visualization types and spatial layout of different views.

We compare VisImages with the existing datasets from the perspectives of data source, visualization layout, quantity, taxonomy, and labels. (Table 1). First, VisImages comprises visualizations from visualization academic publications, which targets to visualization experts, while some datasets, such as Beagle [8], ChartSense [20], MassVis [12], [22], and REV [21], mainly collect data from online libraries and news media, thereby contain a majority of basic visualizations and are usually presented to the public. Therefore, VisImages might be a better choice to analyze and study the frontier designs from visualization community.

Second, VisImages is outstanding in the quantity of annotated visualizations, number of annotated visualization types, and diversity of labels compared to existing datasets.
For example, compared to VIS30K and MV Dataset that are built from visualization publications as well, VisImages has smaller granularity on labels and larger numbers of visualizations. Compared to MassVis [12], [22] and REV [21] that mainly annotate the details in single-panel visualizations (e.g., visual marks, legend, and title), VisImages includes rich annotations (i.e., visualization types, locations, and image captions) on multi-panel visualizations (e.g., visual analytics systems). The images and labels can be used to train machine learning models on visualization classification and localization, and discovering new research opportunities of computer vision for visualization (CV4VIS). In all, VisImages is a good complementary to existing datasets, providing spatial layouts of multiple visualizations and semantic information from captions.

2.2 Visualization Literature Analysis & Datasets

Literature analysis is an important research area for indexing and understanding the publications. Current works mainly focus on the following four types of data: text, citations, authors, and metadata [7]. In this paper, we focus on the studies in the visualization research community.

Many datasets (e.g., [29], [30], [31], [32], [33]) propose to support interactive literature analysis for visualization publications. The most up-to-date dataset is vispubdata.org [33], which contains metadata of publications in IEEE VIS sub-conferences, i.e., VIS, InfoVis, VAST, and SciVis. The publication data, including authors, references, keywords, etc., are collected from the electronic proceedings. A series of visual analytics tools, such as CiteVis2, CiteMatrix, and VisList [33], are proposed on the basis of vispubdata.org. To assist researchers in conducting literature reviews, Ponsard et al. [3] proposed PaperQuest, which is a tool to search for relevant papers that are interested in. Several works [2], [6] also attempted to organize publications based on research topics. Chuang et al. [6] introduced a framework using topic modeling to analyze the InfoVis corpus. Isenberg et al. proposed KeyVis [2] by extracting the keywords of visualization papers. However, none of the above studies investigate images in the publications. VisImages serves as a complement to these works and provides a large corpus of images with rich annotations, including visualization types, bounding boxes, and captions.

3 Dataset Construction

To construct VisImages, we started with collecting images from top-venue visualization publications (Fig. 2(A)). 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 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 our paper. We firstly downloaded PDF files of the papers according to the digital object identifier (DOI) provided in vispubdata.org [33]. Next, we used PDFFigures 2.0 [34] to extract images and the corresponding captions from the PDF files. We extracted the figures and tables indexed by Fig. and Table, and the inline images without a caption. To ensure high quality, we manually checked and revised the image bounding boxes and captions obtained by PDFFigures 2.0. In total, we processed 1397 papers in VAST and InfoVis dated from 1996 to 2018 and collected 12,267 images with 12,057 captions.

Next, to determine the visualization types for annotation (Fig. 2(B) and (C)), we basically used a visualization taxonomy proposed by Borkin et al. [12] (Section 4). Given that recognizing the types of visualizations requires visualization expertise, we recruited qualified participants from the university in the course of annotation (Section 5.1). Specifically, for each image, three participants were asked to identify all possible visualization types that appear (Fig. 2(B)). To address conflicts, we adopted majority voting that a visualization type was accepted only if at least two researchers voted for it (Fig. 2(C)). After this step, we collected 10,289 images containing visualizations types within our taxonomy.

Finally, we employed crowd workers to annotate bounding boxes for each identified visualization (Fig. 2(D)). To ensure high data quality, we carefully designed the tasks and performed cross-validation (Section 5.2). As a result, we obtained a dataset of 35,096 bounding boxes, each corresponding to a specific visualization.

4 Taxonomy

To identify visualization categories for annotation, we used the taxonomy proposed by Borkin et al. [12] that categorizes the visualizations into a two-level structure, i.e., 12

| Dataset | Audience | Layout | #Annotated Visualizations | #Images | #Annotated Categories | Label Types | How to label? |
|---------|----------|--------|--------------------------|---------|-----------------------|-------------|--------------|
| MassVis | general users | single-panel | ~2000 | ~2000 | 12 | ✓ | - | manual annotation |
| REV     | general users | single-panel | ~5000 | ~5000 | 4 | ✓ | - | machine generation + manual refinement |
| Beagle  | general users | single-panel | 33,778 | 33,778 | 24 | ✓ | - | manual annotation |
| ChartSense | general users | single-panel | ~2000 | ~2000 | 10 | ✓ | - | search engine + manual refinement |
| VIS30K  | visualization experts | - | - | ~30,000 | 4 | ✓ | ✓ | object detection models + manual refinement |
| MV Dataset | visualization experts | multi-panel | not reported | 360 | 14 | ✓ | ✓ | manual annotation |
| VisImages | visualization experts | multi-panel | 35,096 | 12,267 | 34 | ✓ | ✓ | ✓ | manual annotation |
categories with sub-types. Their taxonomy categorizes the visualizations used in the public (i.e., infographics, news media, scientific journals, and government & world organization) based on their visual encoding (e.g., bar and area), visual tasks (e.g., statistics), and visual layout (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 corpus consists of 13 categories and 34 sub-types, as shown in Table 2.

### 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                                                           |
| Word           | phrase net, word cloud, word tree                                         |
| Tree & Graph   | graph, tree, treemap, hierarchical edge bundling (HEB), sunburst/icicle plot |
| Special        | glyph-based visualization, unit visualization                            |

### 5.1 Visualization Type Annotation

Distinguishing visualizations and their variations is challenging and requires extensive knowledge of visualizations. Thus, we recruited the researchers and students who were experienced in visualization research to annotate types and adopted the rules to ensure high annotation quality.

**Participants.** We recruited 25 participants, including 1 senior visualization expert who had six-year experience in visualization research, 13 Ph.D. candidates with the research focus on visualizations, 7 master students majoring in information visualization, and 4 undergraduate students who have 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 introduced our taxonomy and the definition of each visualization sub-type with examples, as well as the annotation procedure. Specifically, annotating visualization types for an image was a multi-label task in our study. In such a task, participants were shown an image and asked to select all the visualization sub-types occurring in the image based on our taxonomy. If the participants thought the image contained the visualization types beyond the scope of 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.

The formal study contained 40 rounds of annotation tasks, each round comprised of 40 images and taking about 10 minutes to complete. Each participant would be 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 image.

**Quality Control.** We adopted the gold standards and majority voting methods to ensure high quality. The gold standards were the images manually selected and inserted into each round to test whether the participants were focus-
ing on the tasks. The gold standard images contained simple charts placed at the prominent positions of the image, and the participants should correctly identify these visualization types. Each round included eight gold standard images. If a participant failed in two or more gold standards in a round, all results from this round would be rejected, and we would reassign these images to other participants. In addition, we used majority voting to address ambiguities in the annotation. To ensure validity, each image would be annotated by three participants, and 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 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 × 3 repetitions = 36,801 annotations.

Finally, we found 10,289 out of 12,267 images were assigned visualization sub-types from the taxonomy, and the rest were assigned the label “others”. The frequency distribution of each sub-type is shown in Table 3.

### 5.2 Bounding Box Annotation

With the specified visualization sub-types in each image, we further focused on annotating bounding boxes (i.e., the positions in the images) for these visualizations. To improve efficiency, we employed the crowd from a data annotation company whose workers are well-trained for similar tasks.

**Criteria.** Our criteria are based on the composition of the visualizations, i.e., visualization with coordinates and without coordinates. For a visualization with coordinates, the bounding box should cover all the components of the coordinate, e.g., axis name, axis labels, chart title, and legends if they are close to the visual representations (Fig. 3B). If more than one sub-types are included in the same coordinate (e.g., error bar & bar chart in Fig. 3B), the area of their bounding boxes was the same. For the visualizations without coordinates, we distinguish 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. 3A1)). For the second case, we only focus on the contents of the requested sub-type. For example, the tree in Fig. 3C is connected to the sankey diagram, and the word cloud in Fig. 3D overlays on the area charts. The bounding boxes only cover the contents of tree and word cloud, respectively. However, in addition to the aforementioned criteria, 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. 3A2)). In this case, we annotate them integrally with a single box with specific.

**Procedure.** The annotation procedure consisted of a training session and the 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 a 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-types and requirements. Only crowd workers who passed the test proceeded to the formal annotation. They were then assigned images with 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 defined bounding box correctness and task correctness to measure the quality of annotation. The correctness of a bounding box was measured by intersection over union (IoU, illustrated in Fig. 4A) 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 F1 score, a metric balancing the recall and precision. The calculation of recall, precision, and F1 score is presented in Fig. 4B).

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 F1 score. If the F1 was lower than 95%, the whole batch of annotation would be rejected. The rejected batch would be annotated again until the F1 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 F1 score evaluation. If the F1 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$.

### 6 VISIMAGES

In this section, we first present an overview of VisImages. Next, we compare the distribution of visualizations in VisImages with that from different sources [12]. Finally, we revisit the taxonomy and gain insights into the sub-types.

### 6.1 Overview of the Data

VisImages contains 12,267 images from 22-year VAST and InfoVis publications with 12,057 captions and
Fig. 3: Criteria for bounding box drawing. (A) shows how to draw bounding boxes for an independent visualization (A1) and multiple identical visualization sub-types (A2) [16]. (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 [17]. (D) shows the bounding box of the target visualization (word clouds) which is overlaid on another visualization [35].

35,096 bounding boxes. Table 3 shows the number of images (#img) and bounding boxes (#bbox) of each sub-type. For the frequent types, we observed that the number of bounding boxes of some charts (e.g., bar chart, scatterplot, etc.) 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 chart and scatterplot, 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. 5 depicts the distribution of each sub-type from 1996 to 2018 using horizon charts, with color mapped to the number of bounding boxes. Many sub-types are becoming increasingly popular, such as bar chart, area chart, scatterplot, matrix, line chart and heatmap (Fig. 5A). We notice that the dark area in graph distributes evenly across years (Fig. 5B), indicating that graph visualization has long been a hot research topic in the VIS community. Similarly, tables have always been a common visualization type in publications (Fig. 5B). 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. 5D). We infer that the increase of tree and treemap might be stimulated by the phenomenon that many researchers transform the data into hierarchical structures for exploiting [37]. Moreover, the number of error bars continuously increases in recent years (Fig. 5C), indicating that statistical analysis on the error is increasing, such as the experiment results of user study or models.

6.2 Visualizations in VisImages vs. in public

Furthermore, we compare the distribution of visualizations that appeared in academic visualization publications and that targets to the 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 media, and government & world organization. The comparison can be conducted directly because we use a same taxonomy. We use “Special” to categorize the visualizations that beyond the original taxonomy. From Table 4 we notice that the distribution of visualization publications is more balanced compared to the others. Tree and Networks 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 presenting data with complex relationships, trees and networks are frequently employed. On the other hand, news media and government & world organizations prefer basic visual representations such as Bars, Table, and Lines because the data they mostly present is relatively simple and in the form of tabular. Scientific papers prefer Diagrams, Lines, and Points 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 ManiWor-dle [38] and dynamic word cloud [39]. However, in public, given that the most commonly used media might be text,
Table 4: The distribution of visualizations in VisImages and MassVis.

| Source | VisImages | MassVis |
|--------|-----------|---------|
|        | Scientific | Infographics | News | Government |
| Text   | 1.1% | 0% | 0% | 0.5% | 0% |
| Grid   | 2.2% | 2.5% | 1.9% | 0% | 0% |
| Dist.  | 3.7% | 3.2% | 0.3% | 0% | 0% |
| Network | 8.7% | 8.1% | 0.9% | 0.3% | 1.3% |
| Circle | 3.8% | 0.3% | 0.0% | 0% | 0% |
| Area   | 4.0% | 1.9% | 0.0% | 4.4% | 3.5% |
| Points | 10.6% | 16.6% | 2.8% | 0% | 0.5% |
| Map    | 6.0% | 9.2% | 9.1% | 13.3% | 7.3% |
| Line   | 11.2% | 19.1% | 1.6% | 19.1% | 12.9% |
| Diag.  | 6.3% | 20.5% | 3.0% | 7.2% | 5.0% |
| Table  | 9.8% | 9.3% | 32.2% | 8.2% | 28.5% |
| Bar    | 12.1% | 6.4% | 5.9% | 40.2% | 36.9% |
| Special| 9.5% | 0% | 0% | 0% | 0% |

#Images 10289 348 490 704 528

Fig. 5: The average number of bounding boxes of visualizations in a paper over time.

The average number of bounding boxes of visualizations in a paper over time.

the authors might expect to use a different media other than text to improve the expressiveness.

7 USE CASES

We present three use cases in this section to demonstrate the usability of VisImages.

7.1 Investigating the Use of Visualizations

VisImages contains rich information from IEEE VIS publications, such as raw images, visualization types and their bounding boxes, captions, and publication metadata. For the rapid exploration of the dataset, we develop VisImage Explorer for users, especially for visualization researchers, to investigate and examine the visual designs. The explorer consists of a search panel, an image gallery, a visualization stream, and a caption cloud.

The search panel (Fig. 6(A)) allows users to search papers by the text, publication year, venue, and authors. A histogram is displayed to show the number of the retrieved publications across years, and the corresponding papers are listed below. Users can further filter the search results according to the venue, authors, and publication years. The image gallery (Fig. 6(B)) then exhibits all images based on the search results. Users are allowed to further filter images by the visualization types (Fig. 6(B1)) and also examine the annotation of each image with a zoom-in view by clicking the raw image. Moreover, the visualization stream (Fig. 6(C)) and caption cloud (Fig. 6(D)) provide the statistical information of the visualizations and captions in these images filtered by the search panel and image gallery. The visualization stream presents the number of bounding boxes of each visualization over time. The caption cloud visualizes the frequent words that occurred in the captions.

To demonstrate the usefulness of VisImages Explorer, we present a use case which investigates graph visualizations in InfoVis and VAST. We begin by searching the related terms in the search panel (Fig. 6(A)), including “graph,” “network,” and “node-link,” and then drill down into the graph-related papers. From the histogram (Fig. 6(A)), we notice that there is a peak in the last five years. Thus, we brush the interval from 2014 to 2018 and first selects “InfoVis” in the conference field. The other three views then update accordingly. After this step, VisImages Explorer finds 16 papers and 219 images.

In the image gallery (Fig. 6(B)), we select the “graph” in the multi-selection list (Fig. 6(B1)), and only the images containing graph visualization are left (106 images) for further exploration. From the visualization stream (Fig. 6(C)), we discover that the number of graph visualizations reaches a peak in 2016. By examining the publications, we find that there are eight papers related to the graph in 2016, which is the most among the selected years. In addition to the dominant area of the graph, there are several other streams representing other visualizations. One obvious stream is for hierarchical edge bundling, which indicates that it tends to co-occur with the graph. By exploring the images, we discover that in 2016, two papers [40], [41] adopt this technique to reduce the visual clutter of the graph, contributing many images to the dataset (Fig. 6(B2, B3, B4)). From the caption cloud (Fig. 6(D)), we discover that the words with the largest size are “network,” “graph,” “edge,” and “node”, which are closely relevant to the definition of graph. In addition, other large words like “constrain” and “layout” that are related to the shape of the graph indicate that researchers in InfoVis focus on how to improve the graph layout for better performance, such as reducing clutter.

Next, we select “VAST” in the conference field in the search panel (Fig. 6(A)) for comparison. The visualization stream of VAST (Fig. 6(E)) exhibits a different distribution compared to that of InfoVis (Fig. 6(C)). The co-occurring visualizations are more diverse in types, given the more streams, such as matrix, map, and heatmap. From the image gallery, we discover that the graphs are coordinated with other visualizations in visual analytic systems to facilitate better analysis. For example, the maps are combined with
graphs to visualize the spatial relationship among multiple places, while the matrices commonly serve as a complement to the graphs showing the correlation of the nodes. Moreover, in the caption cloud, the words “pattern,” “overview,” and “cluster” occupy a large size, while they are invisible in Fig. 6(D). The reason might be that VAST systems tend to use graph visualizations as an overview and analyze the patterns. Therefore, we can see that InfoVis and VAST researchers have different research focuses on graph visualizations.

7.2 Classification Benchmarking with VisImages

Object classification has been adopted in many visualization scenarios, such as visualization reconstruction [21], visualization demographic analysis [8], and chart data extraction [22]. In this case, we show how VisImages can serve as a benchmark for visualization classification models trained on other datasets. We select Beagle as the baseline to train classification models because Beagle has the most categories in common with VisImages and the largest number and categories of samples among existing datasets. Please note that since the images in Beagle are in SVG format, we convert them into raster images for model training and evaluation. For the images in VisImages, we crop the visualizations from the images and categorized them by sub-types. We select 17 common categories from VisImages and Beagle.

In the experiments, we select two widely-used object classification models with different numbers of layers, i.e., ResNet-50, ResNet-101 [42], VGG-16, and VGG-19 [43]. Before the experiment, we randomly divide Beagle into training (60%), validation (15%), and test (25%) sets. We follow a similar training process described by Krizhevsky et al. [44] that all models are trained in two stages with weights pre-trained on ImageNet [45]. In the first stage, we freeze the weights of convolutional layers and train the weights of classification heads. In the second stage, we unfreeze the convolutional layers and fine tune the overall weights. In each stage, the models are trained with stochastic gradient descent (SGD) with 100,000 steps. The initial learning rates of ResNets and the VGGNets are 1e-3 and 1e-5, respectively.

Since a visualization can have multiple sub-types, we use both top-1 accuracy and top-3 accuracy as evaluation metrics. Table 5 shows our results with top-1 and top-3 accuracy. We discover that ResNet-101 achieves the best performance on both levels. This result is consistent with the experiments in ResNets [42] that the network structure of ResNets enables deeper networks and makes the models achieve better results than VGGNets in object classification.

We evaluate the models trained from Beagle on VisImages where charts are collected from visualization publications instead of visualization tool or repository. The models encounter a steep decrease of over 25% (e.g., ↓26.0=80.6-54.6) on both top-1 and top-3 accuracy when tested on VisImages. The setbacks of models trained from the Beagle might be caused by two reasons. First, the imbalanced class distribution in Beagle might result in a limited ability to recognize the minor visualizations of the models. In Beagle, most samples (over 90%) are bar charts, line charts, pie charts, and scatterplots. Second, in many cases, the appearance of charts in Beagle is similar, because the charts are generated by similar code snippets (e.g., D3 and Ploly) with different data or parameters. Compared to Beagles, in addition to charts created from popular tools, the charts in VisImages can also be hand-crafted using design tools, such as Adobe Illustrator, or programming languages that do not rely on specific visualization libraries. Therefore, due to the higher diversity in layout and design of the samples, VisImages can be a good benchmark and complementary for existing datasets.
TABLE 5: The performance of different models on visualization classification with top-1 and top-3 accuracy. The bold numbers show the highest accuracy among models.

| Sub-type            | AP_{IoU=0.50} | AP_{IoU=0.75} | AP_{IoU=0.90} |
|---------------------|---------------|---------------|---------------|
| graph               | 0.63          | 0.73          | 0.76          |
| table               | 0.78          | 0.89          | 0.91          |
| scatterplot         | 0.77          | 0.87          | 0.89          |
| line chart          | 0.81          | 0.90          | 0.92          |
| heatmap             | 0.80          | 0.90          | 0.92          |
| flow diagram        | 0.75          | 0.85          | 0.87          |
| bar chart           | 0.69          | 0.79          | 0.81          |
| map                 | 0.68          | 0.78          | 0.80          |
| parallel coordinate | 0.67          | 0.77          | 0.79          |
| mAP                 | 0.78          | 0.88          | 0.90          |

TABLE 6: The average precision (AP) of different sub-types 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          | 0.95          | 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          |

7.3 Visualization Detection with VisImages

This case exhibits the usability of VisImages on training object detection models, which targets localizing and classifying visualizations from the images. For VisImages, we conduct experiments with Faster R-CNN [25], one of the most popular object detection models, to predict all visualizations and their bounding boxes.

Before the experiment, we first prepared the data for training and testing. We used 80% of images for training and validation, and 20% for testing. Following a similar pipeline in original Faster R-CNN paper [25], we train the model using SGD optimizer with a learning rate of 0.003 for 15k mini-batches. A momentum of 0.9 and a weight decay of $1e^{-4}$ are adopted.

We evaluated the results using average precision (AP), a commonly used metric for model performance in object detection [49]. The results of AP are related to confidence score threshold and intersection-over-union (IoU) threshold. The confidence score indicates the probability that a predicted box correctly represents the class and position of an object. The IoU is used to measure the overlap between a predicted box and a ground-truth box. For detailed definition of IoU, please refer to Sec. 5.2 and Fig. 4(A). Table 6 shows the AP of different sub-types, as well as the mean average precision (mAP) across sub-types under different IoU thresholds. During the experiments, we set the confidence score threshold as 0.80 and evaluate the AP under different IoU thresholds.

Experiment results show that the mAPs on VisImages are 0.78 (IoU=0.50), 0.78 (IoU=0.75), and 0.52 (IoU=0.90). Furthermore, we discover that the graph and table achieves the highest AP. By exploring the samples, we discover that the graphs can be correctly recognized even the instances have diverse layout, such as Fig. 7(C1) and Fig. 7(C2). We infer that this might because graphs have specific shapes with links connecting discrete points or point clusters, making them distinguishable compared to other sub-types. In addition, tables commonly occupy the whole images with similar styles in the publications (Fig. 7(A)). Therefore, detecting the tables from the images is relatively easy for the model with relatively homogeneous samples. The model also achieves good performance on bar chart, heatmap, and map. For example, the map in Fig. 7(D2) is a choropleth whose color encodes the data value of each region. The model recognizes it as a combination of map and heatmap, which is consistent with our taxonomy.

On the other side, we notice that Faster R-CNN may still face challenges in some scenarios when detecting visualizations. For example, the model may not work robustly on the bar charts with imbalanced distribution. In Fig. 7(D), the model can correctly localize and recognize some of the bar charts (Fig. 7(D3)), but fails in the cases in Fig. 7(D4) which contain a prominent bar and other barely visible bars. This cases indicates that Faster R-CNN might only recognize the overall shape of the charts, instead of visual encoding.

According to these cases, the model seems to focus on low-level visual features (e.g., the overall distribution of the bars, the colors of the area, the shape of the grids), and cannot clearly learn the differences in the semantics. However, visualizations are not only defined by visual appearance, but also their visual encodings (i.e., the mapping between data and visual properties), which lacks consideration in the current model development process. Therefore, we encourage to improve the detection models for the visualization scenarios, such as adding specific modules to recognize visual encoding of the charts.

8 DISCUSSION

We see VisImages as an exciting start point for leveraging the intelligence of the visualization community itself, by forging a path to the high-quality, fine-grained, and large-scale visualization dataset. We envision that VisImages can inform opportunities for automated visualization and advance our knowledge of the field.

Opportunities for automated visualizations. Our corpus can serve as a useful resource and a reference for a wide range of automated tasks, such as visualization classification, detection, recommendation, and captioning, inspiring the needs to develop and improve machine learning models in the visualization scenarios. For example, our use case benchmarks a set of visualization classification models. Future research can compare with our results for evaluation. Beside, visualization captioning aims to generate a textual description for a visualization. Images with elaborate captions created by paper authors in VisImages can naturally be a qualified resource for training caption models. Moreover, a line of research puts efforts on visualization recommendation and automated visual analysis. VisImages contains screenshots of well-crafted visual analytics systems, along with annotation of specific visualization types and positions, which provides a valuable resource for training models on such tasks.

Benefits to literature analysis. The image corpus from visualization publications offers new possibilities to conduct literature analysis in visualization and help understand the evolution of the field. First, we develop VisImages Explorer that allows users to explore the image gallery freely. Our use case reviews the graph visualizations by analyzing the common combination with other visualizations and their
research focuses in InfoVis and VAST. We make the explorer available for the community to discover more new insights. Furthermore, researchers can investigate VisImages combined with existing publication metadata collections (e.g., Vispubdata.org [33] and KeyVis [2]). The integration of images and text supports the multi-modal analysis and is supposed to reveal unexpected insights. For example, given a task claimed in the paper, what visualization designs are mostly proposed?

**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 “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. Users can directly report problems on the VisImages homepage. Second, with the rapid development of the field and the emergence of new visual designs, the current taxonomy would become outdated quickly. Thus, the taxonomy should be improved continually to cover novel and nuanced designs. Third, we plan to continually refine and improve our taxonomies to meet the growing diversity of visualization designs. 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 pipeline to semi-automate the annotation process, leveraging both human and machine intelligence. Third, we plan to continually refine and improve our taxonomies to meet the growing diversity of visualization designs.

**9 CONCLUSION AND FUTURE WORK**

In this paper, we create and make available VisImages, a corpus of images from the top-venue visualization publications. VisImages contains 12,267 images with captions from 1397 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 in the dataset. We further investigate VisImages with an overview of visualization distribution across years and types, and revisit our taxonomy. VisImages present 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 by three use cases including visual literature analysis and automated tasks like classification and detection. We envision that VisImages can broaden the diversity of visualization research [50] and bring new research opportunities.

VisImages takes the first step to curate and explores visualization images from 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 pipeline to semi-automate the annotation process, leveraging both human and machine intelligence. Third, we plan to continually refine and improve our taxonomies to meet the growing diversity of visualization designs.

**REFERENCES**

[1] J. Stasko, J. Choo, Y. Han, M. Hu, H. Pileggi, R. Sadana, and C. D. Stolper, “CiteVis: Exploring conference paper citation data visually,” in Proceedings of IEEE InfoVis(Posters), vol. 2, 2013.

[2] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, “Visualization as seen through its research paper keywords,” IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 1, pp. 771–780, 2017.

[3] A. Ponsard, F. Escalona, and T. Munzner, “PaperQuest: A visualization tool to support literature review,” in Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2016, pp. 2264–2271.

[4] M. Delest, T. Munzner, D. Auber, and J.-P. Domenger, “Exploring InfoVis publication history with Tulip,” in Proceedings of IEEE Symposium on Information Visualization, 2004, pp. r10–r10.

[5] Z. Li, C. Zhang, S. Jia, and J. Zhang, “Galex: Exploring the evolution and intersection of disciplines,” IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 1, pp. 1182–1192, 2020.

[6] J. Chuang, S. Gupta, C. Manning, and J. Heer, “Topic model diagnostics: Assessing domain relevance via topical alignment,” in Proceedings of International Conference on Machine Learning, 2013, pp. 612–620.

[7] P. Federico, F. Heimerl, S. Koch, and S. Miksch, “A survey on visual approaches for analyzing scientific literature and patents,” IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 9, pp. 2179–2198, 2017.
Dazhen Deng is currently a Ph.D. student in the State Key Lab of CAD&CG, Zhejiang University. He received the B.E. degree from Zhejiang University in 2018. His research interests mainly lie in sports visualizations and machine learning for visualization design.

Yihong Wu is currently a Ph.D. student in the State Key Lab of CAD&CG, Zhejiang University. He received the B.E. degree from Zhejiang University in 2020. His research interests mainly lie in computer vision and visual analytics.

Xinhuan Shu is currently a Ph.D. candidate in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). She received her B.E. degree in Computer Science and Technology from Zhejiang University, China in 2017. Her research interests include data-driven storytelling, animated visualization, and visual analytics.

Jiang Wu is currently a Ph.D. student in the State Key Lab of CAD&CG, Zhejiang University. He received the B.E. degree from Zhejiang University in 2019. His research interests mainly lie in sports visualizations and event sequence analysis.

Mengye Xu is currently pursuing a B.S. degree in Digital Media Technology, Zhejiang University.

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

Weiwei Cui is a Lead Researcher at Microsoft Research Asia, China. He received his PhD in Computer Science and Engineering from the Hong Kong University of Science and Technology and his BS in Computer Science and Technology from Tsinghua University, 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 is a Professor at 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. He received his Ph.D. degree in Computer Science from the Hong Kong University of Science and Technology. Prior to his current position, Dr. Wu was a postdoctoral researcher at the University of California, Davis from 2010 to 2012, and a researcher in Microsoft Research Asia from 2012 to 2015. For more information, please visit http://www.ycwu.org.