Parsing and Summarizing Infographics with Synthetically Trained Icon Detection

Spandan Madan* Harvard University
Zoya Bylinskii† Adobe Research
Carolina Nobre‡ Harvard University
Matthew Tancik§ UC Berkeley
Adria Recasens§ MIT
Kimberli Zhong§ MIT
Sami Alsheikh§ MIT
Aude Oliva MIT
Fredo Durand MIT
Hanspeter Pfister Harvard University

a) Representative infographic b) Visually29K dataset c) Synthetic training d) Detected icons

Figure 1: Infographics effectively convey messages using a mix of text and pictures (a). Automatically parsing infographics is challenging due to abstract concepts and diverse visual styles. We present an approach to detect and tag icons in infographics (see Fig. 8 for final results). We accomplish this by curating and annotating a dataset of infographics called *Visually29K* (b), and training an icon detection model on synthetic data generated by pasting icons onto patches of infographics (c). We also present three demo applications: topic prediction, multi-modal summarization - outputting text tags and visual hashtags representative of an infographic's topics (d), and multi-modal search - prioritizing infographics containing both visual and textual elements matching a query. *©oceanservice.noaa.gov, Evanmade Graphic Design, Walker Sands, Overson Roofing.*

Abstract

Widely used in news, business, and educational media, infographics are handcrafted to effectively communicate messages about complex and often abstract topics including 'ways to conserve the environment' and 'coronavirus prevention'. The computational understanding of infographics required for future applications like automatic captioning, summarization, search, and question-answering, will depend on being able to parse the visual and textual elements contained within. However, being composed of stylistically and semantically diverse visual and textual elements, infographics pose challenges for current A.I. systems. While automatic text extraction works reasonably well on infographics, standard object detection algorithms fail to identify the stand-alone visual elements in infographics that we refer to as 'icons'. In this paper, we propose a novel approach to train an object detector using synthetically-generated data, and show that it succeeds at generalizing to detecting icons within in-the-wild infographics. We further pair our icon detection approach with an icon classifier and a state-of-the-art text detector to demonstrate three demo applications: topic prediction, multi-modal summarization, and multi-modal search. Parsing the visual and textual elements within infographics provides us with the first steps towards automatic infographic understanding.

Index Terms: Human-centered computing—Visualization—Visualization systems and tools—Visualization toolkits; Computing methodologies—Artificial intelligence—Computer vision—Computer vision problemsObject detection

*e-mail: spandan_madan@g.harvard.edu
†e-mail: bylinski@adobe.com
‡e-mail: cnobre@g.harvard.edu
§work done during time spent at MIT

1 Introduction

Consider the infographic in Figure 1a listing ten things you can do to help protect the planet. The bold and bright icons representing concrete objects (e.g., the bike) and abstract concepts (e.g., recycling) draw you to the infographic and guide you through the text descriptions. Icons (or pictograms) can serve a powerful communicative purpose in informational and promotional media (e.g., visualizations, articles, and presentations), by directing attention to key messages or take-aways and making those messages more memorable and effective [6, 10, 11, 34]. Though unless supporting text (e.g., ALT tags) is present, the information encoded in icons is inaccessible to some human consumers, as well as to computational systems that may need to caption, search, summarize, or categorize such multi-modal media. Indeed, a domain gap prevents current computer vision models (e.g., object detectors) trained on natural images from generalizing to the abstract and diverse styles in graphic designs and infographics [37, 38, 77].

Infographics can enable the communication of complex information through concise, data-driven messages which can be understood both quickly and easily [43, 71]. There is a growing body of research which propose the use of infographics for communicating vital information for better healthcare, education, business and public policy outcomes. This includes increasing public access to medical information [4, 50, 60, 68], summarizing medical literature [49], helping to improve learning outcomes in classrooms [8, 18, 40, 56, 75], and effectively disseminating complex public policy information [7, 55]. However, designing effective infographics is often challenging and time consuming [13, 47, 61]. As a result, techniques that provide

support for infographic designers, such as machine learning driven authoring tools [17, 61, 84], are increasingly important. Our work adds to the current body of work through automatic detection and labeling of icons, which can allow designers to search for semantically meaningful icons in real world infographics.

In this paper, we tackle the challenge of identifying stand-alone visual elements, which we call icons. To adapt to the stylistic, semantic, and scale variations of icons in infographics (Fig. 2), which differentiate them from objects in natural images, we propose a synthetic data generation approach. We paste Internet-scraped icons onto background patches in infographics to create training data for an icon detection model (Fig. 1c). Our resulting icon detections outperform models trained on natural images, achieving 38% precision and 34% recall. In comparison, a popular object detector [62] reaches 14% precision and 7% recall, demonstrating a representation gap between objects in natural images and icons in graphic designs.

For training computational models, we curated a novel dataset of 29K infographics from the *Visual.ly* design website [1] (Fig. 1b), covering diverse topics, including health, technology, and weather. Each infographic is annotated with 1-9 tags, out of a set of 391 tag categories. For 1,400 infographics, we collected a total of 21,288 human-annotated bounding boxes of icon locations. For a subset of 544 infographics, we collected 7,761 tags for icon bounding boxes[1]. We used these annotations to evaluate our automatic approaches.

Finally, using our automatic icon detections and tags in combination with text extraction, we present three proof-of-concept applications: topic prediction, multi-modal summarization, and multi-modal search. Given an infographic as input, we predict the topics depicted in an infographic, and individually tag the automatically detected icons. Our multi-modal summarization demo automatically outputs the text tags and visual hashtags that are most representative of an infographic's topics (Fig. 1d). Our multi-modal search demo re-ranks the infographics in a database based on whether the text and visual elements within an infographic match a query.

These applications present a first step towards combining textual and visual information for a computational understanding of infographics. Paired with text extraction, our automatic icon detection can increase accessibility to information stored in graphical form.

Contributions: In this paper, we introduce: (a) *Visually29K*, a novel dataset of curated and annotated infographics; (b) an automated model for icon detection; (c) demo applications (topic prediction, summarization, and search) that become possible once the visual and textual elements inside an infographic can be parsed. Our dataset annotations and model code are available at:
`https://github.com/diviz-mit/visuallydata`.

2 Related Work

Computer vision in the service of graphic design: Computer vision has traditionally focused on understanding natural images. However, there is a growing interest in graphic designs, which motivates a new set of research questions and technical challenges. Prior work has introduced models that take a graphic design or data visualization as input and produce a saliency/importance map as output, for retargeting and thumbnailing applications [14, 54]. Other work predicts the saliency of mobile user interfaces [31], webpages [82], and comics [5]. Zhao et al. [81] predict the personality of a graphic design (futuristic, cute, romantic, etc.) and produces a map of the regions of the graphic design contributing most to the classification. These approaches make high-level predictions about a graphic design as a whole, but do not parse individual design elements.

Other work has leveraged computer vision tools to parse and transcribe textbook diagrams into structured tables for question answering [39, 69, 70], to parse graphs and charts for retargeting applications [59, 67], and to solve graphical reasoning tasks (e.g., quantity estimation) [32]. Recent work focusing on synthesis has involved training models on graphic designs to learn to modify existing layouts [57, 73] or to generate novel layouts [44, 83]. To the best of our knowledge, there is no work on automated understanding of the elements inside infographics or using computer vision techniques to identify icons in graphic designs. The closest application is that of Liu et al., who produce semantic segmentations of mobile UI screenshots [45]. This involves a detection of mobile application icons, which are much more limited in appearance, location, and scale. In contrast, our icons can be simple or complex, photographic or abstract, large or small (Sec. 4). Detecting and recognizing abstract visual representations across diverse styles has also been tackled by prior work on sketches [79, 80], art [76], and illustrations [25, 26].

Datasets of graphic designs: In the space of graphic designs, Zitnick et al. introduced abstract scenes to study higher-level image semantics (relationships between objects, storylines, etc.) [85, 86]. Wilber et al. presented an *Artistic Media Dataset* to explore the representation gap between objects in photographs versus in artistic media [77]. Iyyer et al. built a *COMICS* dataset and made predictions about actions and characters using extracted visual and textual elements from comic panels [38]. Hussain et al. presented a dataset of advertisements and described the challenges of parsing symbolism, memes, humor, and physical properties from images [37]. Borkin et al. collected thousands of data visualizations (including infographics) with element annotations (titles, axes, etc.), memory scores, and eye movements on a few hundred visualizations [10, 11]. Saleh et al. collected a large-scale dataset of infographics along with crowdsourced similarity judgments in order to present an application that can group graphic designs by style similarity [65]. Unlike prior datasets, the dataset of infographics presented in this paper contains very rich meta-data with titles, category labels, and curated tags, and a subset of infographics densely annotated with bounding boxes around icon-like elements. Related large-scale data collection efforts include *Webzeitgeist* [42] - a repository of 100K webpages, and *RICO* [20, 45] - a dataset of 9.7K mobile apps covering 72K unique UI screens, both datasets collected to enable statistical analysis of design patterns and design-driven search and machine learning.

Document parsing: An infographic is part image and part document. Related work on document understanding includes classifying documents by type (e.g., email, news article, presentation, scientific publication) [33], separating figures from text in articles [74] and more fine-grained region classification problems [3], where document regions are labeled as text, image, graphic, table, math, etc. There are also vision-based and DOM-based approaches that decompose a website into sub-parts for further analysis [15]. A separate class of approaches transcribe the text from document pages into characters (i.e., optical character recognition methods) [72]. Most document analysis and retrieval methods, however, stop short of processing the semantics of the individual document elements [48].

Synthetic training data: The use of synthetically generated data to train deep neural network models has been gaining popularity, e.g., for learning optical flow [12], action recognition [19], overcoming scattering [66], and object tracking [24]. Simulated environments like video games have been used to collect realistic scene images for semantic segmentation [64]. Our work was inspired by a text recognition system which was trained on a synthetic dataset of images augmented with text [30]. Related to our approach, Dwibedi et al. insert segmented objects into real images to learn to detect natural objects in the wild [22]. We leverage the fact that infographics are digitally-born, so augmenting them with more Internet-scraped design elements is a natural step. Tsutsui and Crandall synthesize compound figures by randomly arranging them on white backgrounds to learn to re-detect them [74]. However, the icons we aim to detect occur on top of complex backgrounds, so we need our synthetic data to capture the visual statistics of in-the-wild infographics.

[1] Our Visually29K dataset is available at: [http://visdata.mit.edu].

Dataset	# of tags	Images per tag	Tags per Image	Aspect ratios
63K (full)	19469	min=1 max=3784 mean=7.8	min=0 max=10 mean=3.7	from 1:20 to 22:1
29K (curated)	391	min=50 max=2331 mean=151	min=1 max=9 mean=2.1	from 1:5 to 5:1

Table 1: Dataset statistics. We curated the original 63K infographics available on *Visual.ly* to produce a representative dataset of 29K infographics with consistent tags and sufficient instances per tag.

Figure 2: Examples of stylistic and semantic variations in our icons. a) Visually similar icons corresponding to different but semantically related tags *medical, doctor, health, hospital, medicine*. b) Icons with varied styles for the tag *dog*. c) Icons with varied semantic representations for the tag *accident*.

3 VISUALLY29K: A LARGE-SCALE CURATED INFOGRAPHICS DATASET

To facilitate the development of automated systems for parsing infographics, we assembled the *Visually29K* dataset. We obtained 63K static infographic images from the *Visual.ly* website, a community platform for human-designed visual content. Each infographic comes categorized, tagged, and described by a designer, making it a rich source of annotated data. We curated this data to obtain a representative subset of 28,973 images, ensuring sufficient instances per tag (Table 1). The existing tags are free-form text, so many of the original tags were either semantically redundant or had too few instances. We reduced the original 19K tags down to 391 tags with at least 50 exemplars each by merging redundant tags manually using WordNet [52] (e.g., equating *web* and *website*; grouping *cellphone* with *mobile phone*; combining *social marketing*, *online marketing*, and *content marketing* under the single tag *marketing*). Tags range from concepts which have concrete visual depictions (e.g., *car, cat, baby*) to abstract concepts (e.g., *search engine optimization, foreclosure, revenue*). Metadata for this dataset also includes labels for 26 categories (for 90% of the infographics), titles (99%) and descriptions (94%), available for future applications.

The infographics in *Visually29K* are very large: up to 5000 pixels per side. Over a third of the infographics are larger than 1000×1500 pixels. Aspect ratios vary between 5:1 and 1:5. Visual and textual elements occur at a broad range of scales. This needs to be taken into account in the design of computational systems that parse the infographics. We filtered out infographics from the original set that had even more extreme aspect ratios, as they are rarer and would create technical challenges during automatic parsing.

4 TRAINING AN ICON DETECTION MECHANISM

We use **icon** to refer to any visual element with a well-defined closed boundary in space and different appearance from the background (i.e., can be segmented as a stand-alone element), similar to how an object is defined by Alexe et al. [2]. While icons can be detected in different graphic designs, we train and test a model for infographics.

Training an object detector often requires a large dataset of annotated instances, which is a costly manual effort. We took a different approach, leveraging the fact that infographics are digitally-born to generate synthetic training data: we augmented existing infographics with Internet-scraped icons. The advantage of this approach is that we can synthesize any amount of training data by re-sampling infographics and selecting appropriate regions to paste new icons.

Collecting icons: Starting with the 391 tags in the *Visually29K* dataset, we queried Google with the search terms 'dog icon', 'health icon', etc. for each tag. The search returned a wide range of stylistically and semantically varied icon images (Fig. 2). We obtained 250K icons with both transparent and non-transparent backgrounds[2]. For instance, the first icon in Fig. 2b has a non-transparent background, while the second icon has a transparent background. Photographic elements are also considered icons as long as they have a well-defined boundary and are not part of the infographic's background. Only transparent-background icons were used to train the icon proposal mechanism described in this section, while all 250K icons were used to train an icon classifier (Section 7).

Synthesizing training data: To generate our synthetic data, we randomly sampled 600×600px windows from the *Visually29K* infographics and pasted icons onto patches free of texture, to avoid overlap with other visual and textual elements (Fig. 3b). For this, Canny edge detection was applied to the patch, followed by Gaussian blur centered on the patch, and finally summed to quantify the local entropy of the patch. Only patches with entropy lower than a threshold were selected. A randomly selected icon from our icon collection (Fig. 3a) was then pasted onto each valid patch, while ensuring that it contrasted sufficiently with the background to be visible. This process was repeated until a desired number of icons were added per window (Fig. 3c). Parameters corresponding to the number and size of icons to add per window, as well as contrast and entropy thresholds, were all tested to find the best setting for generating synthetic data (details in the Supplemental Material). Using this approach, we generated 10K training instances. Each instance corresponds to a window sampled from an infographic with 4 icons pasted into it. The coordinates of each pasted icon are used as ground truth annotations to train the icon detector. Although a patch can contain additional un-annotated icons native to the infographic, this is similar to object detection with partial labels, which has been shown to successfully generalize [22, 27, 78].

Training an icon detector: We chose the Faster R-CNN neural network architecture [63] for our task, although our methodology of training with synthetic data is agnostic to the choice of architecture. This model puts more emphasis on the local visual appearance of an object rather than the global scene layout [22], which is important given that icons can occur at any location in an infographic. We changed the last layer of the network to classify only two categories: icon versus background. We trained the model for 30K iterations on our 10K synthetic training instances. Training details are provided in the Supplemental Material.

Producing icon detections: To handle the large infographics in the *Visually29K* dataset containing visual features at different scales, we sample windows at 3 different scales as input to our trained icon detector. The first scale spans the entire image. For the two subsequent scales, we sampled 4 and 9 windows, respectively, such that windows at each scale jointly cover the entire image, and neighboring windows overlap by 10%. Before being fed into the detector, every window is rescaled to 600×600px. The predicted detections per window are then thresholded. To address the common problem in detection algorithms whereby the same object may be detected multiple times at different scales, non-maximum suppression (NMS) with a value of 0.3 was used [16, 53]. For detections which overlap above this threshold (as measured by intersection-over-union), NMS

[2]Our icon dataset is available at: `https://github.com/diviz-mit/visuallydata/blob/master/links_to_data_files.md`.

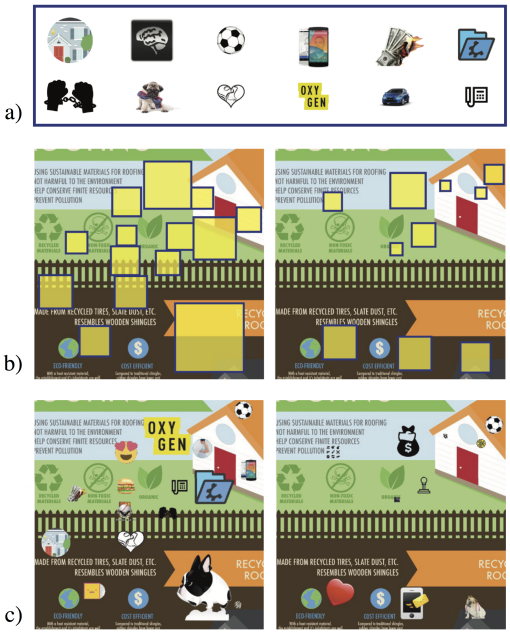

Figure 3: Synthetic data generation pipeline: a) Icons with transparent backgrounds from Google. b) Patch candidates to paste icons into, chosen using different threshold settings (e.g., the approach on the right is more conservative). c) Icon-augmented infographics windows for training. Infographic by ©*Overson Roofing.*

suppresses all detections which do not have the maximum detection score. Further, NMS is applied once again to combine smaller detections (often parts of icons).

5 Humans as Icon Detectors

To produce ground-truth for evaluating computational approaches, we designed two crowdsourcing tasks to collect human annotations of icons for a subset of the infographics from *Visually29K* (Fig. 4). In the first task, we asked participants to annotate all the icons in an infographic. In the second task, we asked participants to annotate only the icons corresponding to a particular tag (e.g., only icons related to *gaming*) in an infographic. We use the annotations from the first task to evaluate our automatic icon detection results, and the second task to evaluate our multi-modal summarization application.

Task 1. Generic icon detection: For a set of 1,400 infographics from *Visually29K*, we asked participants to "put boxes around any elements that look like icons or pictographs" (Fig. 4). No further definitions of "icon" were provided. A total of 45 participants were recruited via student mailing lists, and each spent 0.5-3 hours annotating icons in as many infographics as they wanted. Pay varied between $10-20 per hour as the data collection effort progressed. An average of 15 icon bounding boxes were annotated per infographic. Each infographic was annotated by a single person, resulting in a total of 21,288 bounding boxes across the 1,400 infographics. We refer to these annotations as the "evaluation set". We split this "evaluation set" further into 400 infographics for validation (model tuning) and 1,000 for testing (final model evaluation).

Human consistency: Because no definition of "icon" was provided during the annotation task, we evaluated whether people annotated the same regions in infographics. For instance, people sometimes disagreed about whether a map, embedded graph, or photograph should be counted as an icon. They also occasionally disagreed about the boundaries of the icon (Fig. 5). Each infographic in our "evaluation set" was only annotated by a single person, so to measure human consistency we collected an additional set of

Figure 4: User interface for collecting human ground truth to evaluate icon detection and classification. Participants were asked to either annotate all icons on an infographic (left panel), or only icons corresponding to a specific tag (right panel) (e.g., *gaming*). ©*SonyPS4.com*

annotations on a subset of infographics. Out of the 1,400 annotated infographics, we randomly selected 55 infographics for which we collected annotations from an additional 5 participants each. We then measured the overlap in the bounding boxes generated by each of these 5 participants and the bounding boxes in the "evaluation set". The results of this analysis, averaged across participants and infographics, are reported as "human consistency" in Table 2. We used these scores as an upper bound for computational models, which similarly need to decide what counts as an icon.

Task 2. Topic-specific icon detection: We used a similar annotation set-up as in task 1, but this time asked participants to mark bounding boxes around icons that correspond to a *specific* topic (Fig. 4). Recall that the infographics in our dataset contain an average of 2 tags each (Table 1). We used 544 infographics along with their associated *Visually29K* tags to produce 1,110 separate annotation tasks, each task corresponding to a single image-tag pair. If an image had multiple tags, each image-tag pair would be shown to a different participant so participants would not annotate the same image more than once. For 275 (25%) of these 1,100 annotation tasks, participants indicated that there were no icons on the infographic corresponding to the specified tag. For instance, an infographic with the tag *investing* may not necessarily contain visual elements corresponding to this tag. For remaining 835 image-tag pairs, we collected a total of 7,761 bounding boxes from 45 undergraduate students, averaging 9 bounding boxes per image-tag pair. To compute human consistency for this task, we similarly collected annotations from an additional 5 participants each for 55 infographics (a total of 172 separate image-tag annotation tasks x 5 participants). The results of this analysis, averaged across participants and image-tag pairs, are reported as "human consistency" in Table 3.

6 Evaluation

In this section, we evaluate our trained icon detector compared to alternatives at detecting icons in our annotated test set of 1,000 infographics (Section 5, Task 1). We report standard object detection metrics: precision (*Prec*), recall (*Rec*), F-measure (F_β), and mean Average Precision (*mAP*). We compute the *Intersection Over Union* (IOU) of icon bounding boxes, threshholding the IOU at 0.5 to evaluate precision and recall [23]. F-measure is defined as:

$$F_\beta = \frac{(1+\beta^2)Prec \times Rec}{\beta^2 Prec + Rec}$$

We set $\beta = 0.3$ to weight precision more than recall [9].

Icon detection is related to objectness, general object detection, and object segmentation. We ran 5 methods spanning these different tasks, originally trained on natural images, to evaluate the representation gap when applied to infographics. We used objectness [2], state-of-the-art object detectors YOLO9000 [62], SSD [46], and Faster R-CNN [63], and class-agnostic object masks [58]. To account for the fact that some methods are trained to detect multiple object classes, we considered a detection of any object class with score above a threshold as a detected icon. Default parameters were

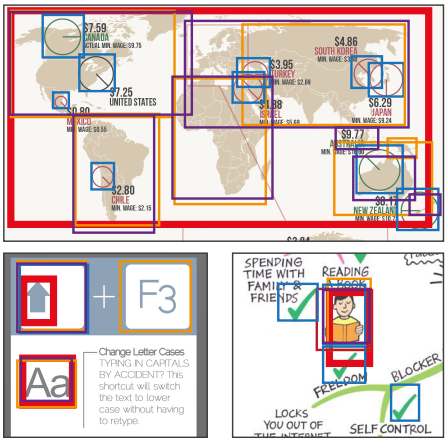

Figure 5: Human consistency in annotating icons is not perfect because people have different notions of what should be counted as an "icon". Here we include 3 crops from annotated infographics. For instance, in the world map we notice three strategies: labeling the entire map as an icon, labeling individual continents, and labeling the circle graphics superimposed on the map. The set of participant annotations belonging to the "evaluation set" are indicated in red. The other colored boxes depict annotations from additional participants recruited for consistency analyses. Crops from infographics by ©*Lisa Mahapatra, Computeach.co.uk, Learning Fundamentals.*

used for Faster R-CNN and YOLO9000. We tried SSD with three thresholds: 0.01, 0.1 and 0.6 (default), and report results on the best setting (0.01). From Table 2, we find that the icon detector trained on our synthetic data significantly outperformed all other models trained on natural images. These results confirm that a representation gap exists between objects in natural images and icons in graphic designs. Our contribution was to show how synthetically-generated data can be used to retrain an object detection model originally designed for natural images, and adapt it to detecting icons in infographics.

Discussion: There are several attributes of icons which make their detection challenging for existing algorithms trained on natural image datasets like ImageNet [21] and CIFAR [41]. Firstly, most natural image datasets contain smaller images with average size under 500x500px, and networks are often designed to be trained by downsizing these images to around 250x250px. Infographics are frequently very large, and downsizing them to these resolutions makes visual elements too small to be identifiable. Secondly, icons occur at a large range of scales. Modern convolutional neural networks are not scale invariant, in that they struggle to generalize to scales beyond those seen during training. Thirdly, the wide range of stylistic and semantic variation among icons present a major challenge. For instance, Fig. 2—row (a) shows visually similar icons corresponding to different but related tags like medical, doctor, health, hospital, medicine; row (b) shows icons with varied styles for the tag dog; row (c) shows icons with varied semantic representations for the tag accident. By augmenting infographics with internet scraped icons at multiple scales during training, we're able to mimic these attributes of icons in the training data. This allows our models to detect icons across these variations which is not possible for models trained on natural image datasets.

Ablation experiments: In generating our synthetic training data, we made three design choices: (i) pasting icons into existing infographics, (ii) using icons with transparent backgrounds, (iii) pasting icons onto regions of infographics where they do not overlap with other elements. To evaluate the contributions of these respective design choices, we trained three alternative variants of our model (Fig. 7): (i) "blank background": pasting icons onto plain white backgrounds, instead of existing infographics (similar to [74]); (ii) "non transparent icons": using icons with their original backgrounds, so that when pasted into an infographic there are visible boundaries; (iii) "random locations": adding icons at random locations inside the infographics, disregarding any overlap. All three model variants performed significantly worse than our final model (Table 2). The worst-performing variant was the one trained on blank background images, demonstrating a failure to generalize to infographics.

7 Applications

To facilitate automated applications for infographics like topic prediction, summarization, and search, the visual and textual components inside the infographics need to first be detected and recognized. Here, we put together a number of automatic modules (in bold, below; visualized in Fig. 6) to make these applications possible. Given an infographic, we use our **icon detector** to locate icons, which we then classify into one of 391 tag categories using a separately-trained **icon classifier**. We simultaneously run a **text detector** and feed its output to a **topic prediction module**. The result is a full annotation of the infographic as in Fig. 8 that can be used as input to different applications. In this section, we provide proof-of-concept demonstrations of topic prediction, multi-modal summarization, and multi-modal search.

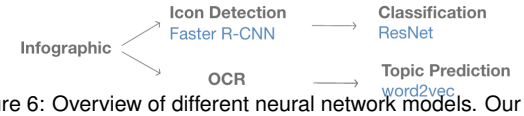

Figure 6: Overview of different neural network models. Our icon detector is used to locate icons, which are then classified into one of 391 tag categories using a separately-trained icon classifier. We simultaneously run a text detector and feed its output to a topic prediction module. Together, these models open new avenues for summarizing and parsing infographics as presented in our applications.

7.1 Topic prediction

Automatically predicting topics that an infographic depicts would facilitate applications that categorize, search through, and caption infographics. The *Visually29K* dataset contains multiple tags per infographic, and provides valuable data for training a topic prediction model. We use the text detected inside an infographic to predict the topics depicted in it, and then automatically tag all detected icons inside the infographic to provide finer-grained annotations.

Predicting topics from text: Given an infographic as input, text extracted from the infographic is used to predict topics depicted in the infographic. We used Google's Cloud Vision optical character recognition [28] as our **text detector**, as it is one of the best publicly-available text detection and parsing systems, capable of generalizing to different fonts and text sizes. We extracted on average 236 words per infographic. Individual words were then converted into their *word2vec* feature representations [29] (for all words for which such a representation was available). This feature space commonly used for natural language processing has been trained to embed semantically related words close together [51].

For our **topic prediction module**, we compute the average *word2vec* representation across all the extracted words, and use it as input to a small neural network to predict tags (details in the Supplemental Material). We used 26K infographics from our *Visually29K* dataset for training this model. Each infographic comes with 1-9 tags (2 on average). The output of our model is a 391-dimensional vector of probabilities, corresponding to the 391 unique tags in our dataset. The trained model can then output the top N tags predicted most likely for a given infographic. Evaluating the top-1 predicted tag, we achieve 42.6% average precision and 24.6% average recall at predicting tags for the infographics in our test set. Fig. 10 illustrates the top 3 tags predicted for 5 sample infographics (the original infographics can be found on Visual.ly by their titles).

a) blank background b) non transparent icons c) random locations

Figure 7: Three alternative icon detection models were trained by modifying synthetic data generation: a) pasting icons onto plain white backgrounds instead of infographics; b) using icons without transparent backgrounds, i.e., with visible boundaries; c) pasting icons in random locations, disregarding overlap with other infographic elements. These variants performed significantly worse than our model.

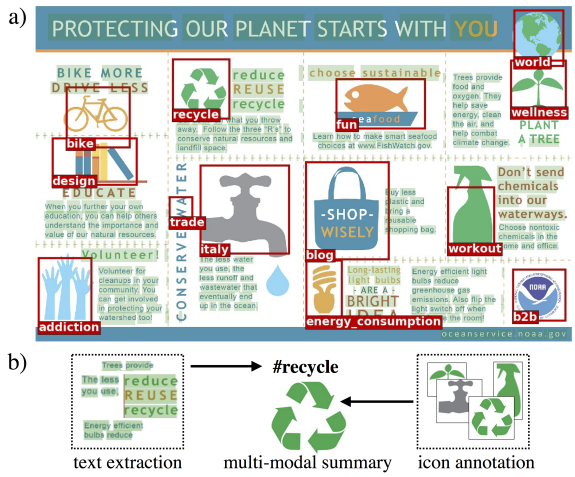

Figure 8: a) The output of our fully-automatic annotation system, running text detection and OCR [28] (semi-transparent green boxes), with our own icon detection and classification (red outlines). We trained an icon detector with synthetic data to make this system possible. The underlying infographic has been faded for visualization. b) Our multi-modal summarization application uses the detected text and icons to produce the text tags and visual hashtags most representative of the infographic's topics. ©*oceanservice.noaa.gov*

Tagging icons: Given an infographic as input, we predict tags for all the automatically-detected icons inside it. We used the same icon dataset that we trained our icon detector on to additionally train an **icon classifier** that takes a detected icon as input, and predicts the most likely tags. Since the icon dataset was collected by searching on Google using tags as queries, we use these queried tags as the ground truth labels to train our icon classifier. Of the total 250K icons, with a few hundred icons collected on average for each of the 391 unique tags, we used 80% icons for training and 20% for validation. Our icon classifier is a ResNet18 architecture [35] re-trained on these 200K icons (training details in Supplemental Material).

On the validation set of 50K icons, our icon classification network achieved 19.1% top-1 accuracy at predicting the correct tag, where chance performance is 1 in 391, or 0.2%. We also evaluated the joint performance of the icon detector and classifier at retrieving relevant icons across infographics. Fig. 9 contains top-ranking icons for a few tags automatically extracted from our infographics dataset.

Discussion: In this section, we showed how the icon detector introduced in this paper can be bundled together with an icon classifier and a text detection system to predict the topics that an infographic is about, as well as to automatically annotate the individual elements, the text and icons, within an infographic. A sample output of this fully-automated system is visualized in Fig. 8a. The output of such a system can now serve as input for future applications seeking to caption, answer questions about, or extract information from infographics. This can increase availability to information that may have otherwise been previously stored in inaccessible, graphical form. In the next two sections, we present two proof-of-concept applications that make use of the text and icons in infographics for respectively summarizing and searching infographic images.

Model	Prec.	Rec.	$F_{0.3}$	mAP
Final model (ours)	**38.8**	**34.3**	**43.2**	**44.2**
Random locations	29.1	15.1	29.6	32.5
Non transparent icons	24.6	17.1	25	26.1
Blank background	7.9	24.3	10.1	10.3
YOLO9000 [62]	13.6	7.1	12.6	13.7
Faster R-CNN [63]	11.0	6.0	10.2	11.4
SSD [46]	9.3	34.2	10.0	11.4
Objectness [2]	2.9	5.6	3.1	3.0
Sharpmask [58]	1.1	1.4	1.2	1.1
Human consistency	63.1	64.7	61.8	66.3

Table 2: Model performance at detecting icons in infographics. First 4 models were trained with synthetic data containing icons. The next 5 models were trained to detect objects in natural images. Human consistency was computed by comparing icon annotations of multiple annotators. All listed values are percentages. Top scores are bolded.

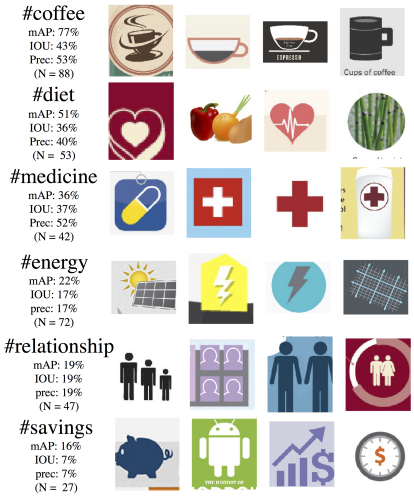

Figure 9: Visual hashtags for different concepts. We include 6 different topic tags, sorted by mAP scores. For each tag, depicted are the top 4 icons classified most confidently as belonging to the tag, sampled out of automatically detected icons from our infographics test dataset (the total number (N) of icon detections per tag is also listed).

7.2 Multi-modal summarization

Just as video thumbnails facilitate the sharing, retrieval, and organization of complex media files, we propose to create multi-modal summaries that can be used for effectively capturing a visual digest of complex infographics. Given an infographic as input, our multi-modal summary consists of textual and visual hashtags representative of an infographic's topics. We define "visual hashtags" as icons that are most representative of a particular text tag.

Predicting visual hashtags: Given an input infographic, we use topic prediction from the previous section to output the set of most representative topics (text tags). We then use our icon classifier to identify the most representative icon for each predicted text tag. We do this by passing all the detected icons to the icon classifier and automatically selecting the icon with highest probability for the tag - this is the visual hashtag. The automatic output of our system is visualized in Fig. 10: for each infographic, we output the top 3

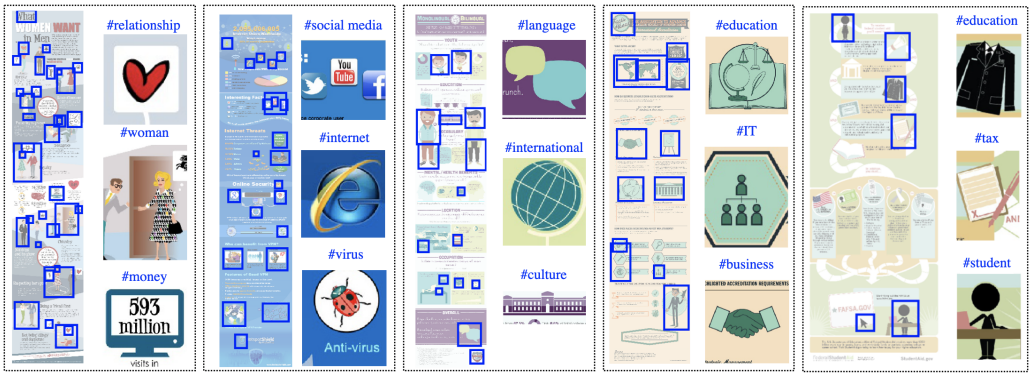

Figure 10: Examples of our automated multi-modal summarization pipeline, which given an infographic as input, predicts text tags and corresponding visual hashtags. Sources: a) *"What women want in men" (Parship.de), b) "Interesting facts about internet threats and security" (Hotspot Shield), c) "Monolingual VS Bilingual" (Technovate Translations), d) "Facts About AACSB International Accreditation: A Closer Look" (Howard University), e) "Eligibility for Federal Student Aid" (StudentAid.gov).*

Model	Top-1 Prec.	mAP
Icon detections + classification	**27.2**	**18.0**
Random locations + classification	16.7	14.2
Non transparent icons + classification	15.9	14.5
Blank background + classification	16.2	14.5
Icon detections only	16.2	14.5
Human consistency	55.4	57.2

Table 3: Given an infographic and text tag as input, we evaluate the visual hashtags returned. For each image-tag pair, we compute IOU with the ground truth bounding box annotations. Precision is measured as the percent of instances with IOU > 0.5. Human consistency was evaluated by comparing the annotations of multiple annotators. All listed values are percentages. Top scores are bolded.

predicted topics and their corresponding visual hashtags.

To evaluate the ability of our computational system to output a relevant visual hashtag for a given infographic and tag, we compare against the human annotations from Section 5, Task 2. Similar to the task that our computational system receives, participants were asked to annotate all icons corresponding to a particular text tag on an infographic. Note that we excluded instances where no icon could be found to correspond to the text tag. We used the remaining 835 image-tag pairs with human annotations. For each image-tag pair, we passed the image to our icon detector, and used our icon classifier to select the detected icon most representative of the tag. We computed the intersection-over-union (IOU) of each of our predicted hashtags with ground-truth human annotations. We report precision as the percent of predicted visual hashtags that have an IOU > 0.5 with at least one of the ground truth annotations (Table 3). We also include the mAP score by considering all our icon detections per image-tag pair, sorted by the icon classifier's confidence. From Table 3 we see that sorting the icon proposals using our icon classifier produces more relevant results (mAP = 18.0%) for a given tag than just returning the most confident icon detections (mAP = 14.5%). We compare to our other baseline models, to once again validate our synthetic training design choices.

Compared to icon detection (Table 2), the performance is worse for icon classification (precision of 27.2% versus 38.8%). Correctly predicting what an icon depicts, rather than just locating the icon, is a significantly harder task, compounded by the abstract nature of some of the icons and their diverse styles (Fig. 2). We leave improving the performance of icon classification to future work.

Browsing infographic collections: We hypothesize that multi-modal summaries, containing both text tags and visual hashtags, may be able to facilitate the sharing, browsing, or organizing of large collections of infographics. To evaluate the potential utility of multi-modal summaries we also ran a small pilot study. We asked 10 participants to browse through an online collection of 138 infographics from *Visually29K*, presented as thumbnails. For half the infographics, hovering over the thumbnails showed the infographic's title, while for the other half, hovering over showed both the title and our automatically-computed multi-modal summary with two text tags and corresponding visual hashtags. The next day, we showed participants another page with 138 infographics, half of them were from the previous day, and the other half were new. Participants were asked to select all infographics they remembered seeing before. Our preliminary results show that the multi-modal summaries increased recall of infographics previously seen by 19.6%, over just seeing the thumbnails and titles of the original infographics (more details in the Supplemental Material). This supports prior work showing that icons can improve the memorability of content [6, 10, 11, 34].

Discussion: In this section, we introduced a potential application of using icons extracted from within infographics to create multi-modal summaries, containing text and visual hashtags to represent the infographics' topics. We evaluated the quality of the visual hashtags retrieved, and discussed the challenges of icon classification over icon detection. We also provided some very preliminary results to support the hypothesis that multi-modal summaries may be able to facilitate the browsing of infographics collections, by increasing recall. It remains up to future work to more rigorously evaluate the benefits and shortcomings of multi-modal summaries for facilitating the sharing, browsing, and organizing of infographic collections.

7.3 Multi-modal search

Search engines typically use text-based metadata (e.g., captions, titles, ALT tags) to determine which images to retrieve for a particular search query. Given the infrastructure we have developed to detect the text and icons in infographics, we propose to retrieve results relevant to a search query by looking *inside the image*.

As a proof-of-concept, we developed a small demo at `http://visdata.mit.edu/explore.html` which retrieves the top 30 infographics for 344 tags in our dataset (removing tags for which we did not find high-confidence icon predictions, or matching text). We sort infographics by total relative area covered by tag-related icons or text. In other words, search results high on the list would correspond to infographics that have text and icons matching the query take up a larger portion of the design. For this demo, we only include text exactly matching the tag, though extensions could include related terms and matching word stems. We pass all icons detected in the infographic to our icon classifier and return those for

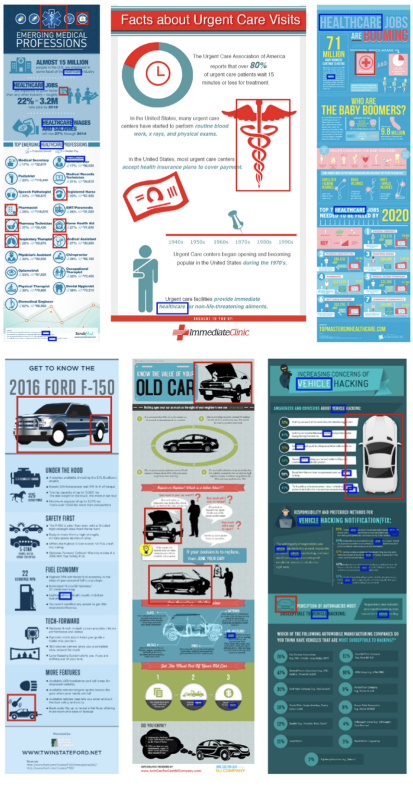

Figure 11: Automatically-retrieved results for tags a) healthcare and b) vehicle, chosen as infographics with the largest area covered by tag-related icons and text matching the tag. Red bounding boxes represent icons automatically detected and classified as the queried tag by our synthetically-trained model. Blue bounding boxes localize text regions where the OCR results match the tag text. More results available at: `http://visdata.mit.edu/`. *©ScrubMed.com, ImmediateClinic.com, TopMastersInHealthcare.com, CureMD.com, TwinStateFord.net, Diana L. Lyons, Andrea Davis, mrmatt.*

which the query tag occurs in the top-5 most confident predictions. In these cases, the icon is considered relevant to the search query.

Fig. 11 visualizes the top 3 infographics returned by our demo search application for two tags: *healthcare* and *vehicle*. Note how the query text might only occur sparsely in the infographic's text, while a large icon or multiple related icons might be an indication that the infographic contains relevant content. In comparison, the Visual.ly search[3] returns infographics where the search term is part of the title. Looking *inside the image* for retrieval is especially useful if the infographic contains terms that are associated but not exactly matching the search phrase (e.g., *urgent care*, *HIPAA*, *medical professions* for the query *healthcare*), for technical terms that might not be in a lexical database (e.g., *FORD F-150* for the query *vehicle*), or if the infographic contains illegible text (due to automatic text parsing failures or an unfamiliar language). In such cases, an icon detector can help retrieve relevant content, independently of the text.

8 Discussion

From the results of the evaluation in Table 2, we find that our trained icon detector performs over three times better as compared to existing object detection approaches trained on natural images according to all four metrics. We were able to achieve these gains by re-training an existing Faster R-CNN model using our synthetic training methodology, with some modifications to the architecture, training, and testing steps. While our methodology has significantly improved performance (e.g., precision increased from 11.0% to 38.8%), there is still room to grow. As presented in Fig. 2, the stylistic and semantic variations depictions of icons make icon detection an extremely challenging problem for neural networks. Here, we have presented a first stab at how this problem can be tackled using data augmentation via internet scraped icons. Our findings support related work showing that there is a big domain gap that prevents models designed for natural images from being usable, out-of-the-box, on graphic designs [37, 38, 77]. We have also demonstrated, via our ablation experiments (and further tests described in the Supplemental Material), that the design of the synthetic training data requires great care for the final trained model to generalize properly to real-world data. While our top precision score on icon detection is 38.8%, we note that humans also occasionally disagree about what constitutes an icon and where its boundaries lie, leading to an upper-bound precision score of 63.1% (when humans are compared to other humans). Our model thus captures 61.4% of this upper bound. We hope that future work building on top of our findings can help close this performance gap.

9 Limitations and Next Steps

This paper contributes an approach to detecting and classifying icons within infographics. We demonstrate the importance of training a model specifically on icons in infographics, since object detectors trained on natural images were shown not to generalize to infographics. While we succeed at detecting many of the icons in infographics, classifying them correctly is more challenging (Fig. 8). There are several aspects which make icon detection and classification challenging for object detectors trained on natural image datasets like ImageNet [21] and CIFAR [41]. These include variation in scale, stylistic and semantic variations as depicted in Fig. 2, and the variation in number of instances per infographic. While a *pictogram* like the bike icon represents the object it refers to, an *ideogram* like the recycling symbol is not a depiction of any one object, but rather of an idea [36]; and while the recycling symbol is a commonly-encountered ideogram, the icon of the hands, to represent the idea of volunteering, is not. Moreover, each infographic comes in its own unique visual style, adding to the complexity of the task.

At the same time, there is quite a bit of context available within infographics that we have not capitalized on in this work. While we take the approach of individually classifying each icon, knowing the identities of the other icons in an infographic can constrain the possible labels for the remaining icons. The text inside the infographics can also be used to disambiguate interpretations. The spatial location of the elements inside an infographic, and the proximity of text to icons, can signal a relationship between the text and visuals. Further, not all text concepts can be visually represented - an aspect that poses a limitation for our multi-modal summarization. However, when considered together, the text and icons in an infographic can paint a more complete picture about the topics of an infographic.

The three demo applications in this paper serve in support of our main contributions, namely a dataset and tools to support the parsing of infographics. We provided proof-of-concepts for how icon detection, classification, and text parsing can be used for the topic prediction, summarization, and search of infographics. More comprehensive applications like captioning, question-answering, and knowledge extraction will require additional computational modules which are beyond the scope of the present paper. To facilitate future developments, we make all resources developed in this paper available to the broader research community, including our *Visually29K* dataset, dataset of 250K icon images with tags, as well as the code and technical details of our icon detection, icon classification, and text-based topic prediction models. More results and demos of our applications are provided at `http://visdata.mit.edu/`.

[3] e.g., `https://visual.ly/search/node?keys=healthcare`

10 Conclusion

The space of complex visual information beyond natural images has received less attention in computer vision, but this is changing with the increasing popularity of work on graphic designs [37–39, 77, 85, 86]. Within this space, we presented a novel dataset of infographics, *Visually29K*, containing a rich mix of textual and visual elements. We developed a synthetic data generation methodology for training an icon detector. We showed that key design decisions for this methodology included augmenting icons with transparent backgrounds onto appropriate regions of infographics. Our trained icon detector successfully generalized to real-world infographics, and together with a text parsing system [28] and an icon classifier, was used for annotating infographics. Facilitated by these computational tools, we presented three demo applications: topic prediction, multi-modal summarization, and multi-modal search.

Infographics are specifically designed with a human viewer in mind, characterized by higher-level semantics, such as a story or a message. Beyond detecting and classifying the elements contained within, an understanding of these infographics involves understanding the included text, the layout and spatial relationships between the elements, as well as the intent of the designer. Human designers are experts at piecing together elements that are cognitively salient and memorable, to maximize the utility of the presented information. This new space of multi-modal data can give researchers the opportunity to model and understand the higher-level properties of textual and visual elements in the story being told. At the same time, our automated tools allow this story, previously stored in graphical form, to start to become more broadly accessible.

References

[1] Visually community. `https://visual.ly/view`, Accessed: 2019-05-01.

[2] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE TPAMI*, 34(11):2189–2202, 2012.

[3] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher. A realistic dataset for performance evaluation of document layout analysis. In *ICDAR*, pp. 296–300. IEEE, 2009.

[4] M. Balkac and E. Ergun. Role of infographics in healthcare. *Chinese medical journal*, 131(20):2514, 2018.

[5] K. Bannier, E. Jain, and O. L. Meur. Deepcomics: Saliency estimation for comics. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, p. 49. ACM, 2018.

[6] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2573–2582. ACM, 2010.

[7] T. Bhasin and C. Butcher. Teaching effective policy memo writing and infographics in a policy program. 2020.

[8] H. Bicen and M. Beheshti. The psychological impact of infographics in education. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 8(4):99–108, 2017.

[9] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE TIP*, 24(12):5706–5722, 2015.

[10] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics*, 22(1):519–528, 2015.

[11] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013.

[12] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

[13] Z. Bylinskii, S. Alsheikh, S. Madan, A. Recasens, K. Zhong, H. Pfister, F. Durand, and A. Oliva. Understanding infographics through textual and visual tag prediction. *arXiv preprint arXiv:1709.09215*, 2017.

[14] Z. Bylinskii, N. W. Kim, P. O'Donovan, S. Alsheikh, S. Madan, H. Pfister, F. Durand, B. Russell, and A. Hertzmann. Learning visual importance for graphic designs and data visualizations. In *UIST*, 2017.

[15] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Vips: a vision-based page segmentation algorithm. *Microsoft Research Technical Report*, 2003.

[16] J. Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[17] W. Cui, X. Zhang, Y. Wang, H. Huang, B. Chen, L. Fang, H. Zhang, J.-G. Lou, and D. Zhang. Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE transactions on visualization and computer graphics*, 26(1):906–916, 2019.

[18] I. Damyanov and N. Tsankov. The role of infographics for the development of skills for cognitive modeling in education. *International Journal of Emerging Technologies in Learning (iJET)*, 13(1):82–92, 2018.

[19] C. De Souza, A. Gaidon, Y. Cabon, and A. Lopez Pena. Procedural generation of videos to train deep action recognition networks. In *CVPR*, 2017.

[20] B. Deka, Z. Huang, C. Franzen, J. Hibschman, D. Afergan, Y. Li, J. Nichols, and R. Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th Annual Symposium on User Interface Software and Technology*, UIST '17, 2017.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[22] D. Dwibedi, I. Misra, and M. Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. *ICCV*, 2017.

[23] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010.

[24] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016.

[25] E. Garces, A. Agarwala, D. Gutierrez, and A. Hertzmann. A similarity measure for illustration style. *ACM Transactions on Graphics (TOG)*, 33(4):93, 2014.

[26] E. Garces, A. Agarwala, A. Hertzmann, and D. Gutierrez. Style-based exploration of illustration datasets. *Multimedia Tools and Applications*, 76(11):13067–13086, 2017.

[27] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2020.

[28] Google. Cloud Vision API: Optical Character Recogition. `https://cloud.google.com/vision/`, accessed in October 2017.

[29] Google. Word2vec model. `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing`, accessed in October 2017.

[30] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, 2016.

[31] P. Gupta, S. Gupta, A. Jayagopal, S. Pal, and R. Sinha. Saliency prediction for mobile user interfaces. In *WACV*, pp. 1529–1538. IEEE, 2018.

[32] D. Haehn, J. Tompkin, and H. Pfister. Evaluating 'graphical perception'with cnns. *IEEE transactions on visualization and computer graphics*, 25(1):641–650, 2018.

[33] A. W. Harley, A. Ufkes, and K. G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *ICDAR*, pp. 991–995. IEEE, 2015.

[34] S. Haroz, R. Kosara, and S. L. Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1191–1200. ACM, 2015.

[35] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[36] J. Hicks. *The Icon Handbook*. Five Simple Steps, 2011.

[37] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic understanding of image and video adver-

tisements. In *CVPR*, 2017.

[38] M. Iyyer, V. Manjunatha, A. Guha, Y. Vyas, J. Boyd-Graber, H. Daumé, III, and L. Davis. The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives. In *CVPR*, 2017.

[39] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A digram is worth a dozen images. In *ECCV*, 2016.

[40] P. N. Kibar and B. Akkoyunlu. A new approach to equip students with visual literacy skills: Use of infographics in education. In *European Conference on Information Literacy*, pp. 456–465. Springer, 2014.

[41] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[42] R. Kumar, A. Satyanarayan, C. Torres, M. Lim, S. Ahmad, S. R. Klemmer, and J. O. Talton. Webzeitgeist: Design mining the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pp. 3083–3092. ACM, New York, NY, USA, 2013. doi: 10.1145/2470654.2466420

[43] J. Lankow, J. Ritchie, and R. Crooks. *Infographics: The power of visual storytelling*. John Wiley & Sons, 2012.

[44] J. Li, T. Xu, J. Zhang, A. Hertzmann, and J. Yang. LayoutGAN: Generating graphic layouts with wireframe discriminator. In *ICLR*, 2019.

[45] T. F. Liu, M. Craft, J. Situ, E. Yumer, R. Mech, and R. Kumar. Learning design semantics for mobile apps. In *The 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 569–579. ACM, 2018.

[46] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single Shot MultiBox Detector. In *ECCV*, 2016.

[47] S. Madan, Z. Bylinskii, M. Tancik, A. Recasens, K. Zhong, S. Alsheikh, H. Pfister, A. Oliva, and F. Durand. Synthetically trained icon proposals for parsing and summarizing infographics. *arXiv preprint arXiv:1807.10441*, 2018.

[48] S. Marinai, B. Miotti, and G. Soda. Digital libraries and document image retrieval techniques: A survey. In *Learning Structure and Schemas from Documents*, pp. 181–204. Springer, 2011.

[49] L. J. Martin, A. Turnquist, B. Groot, S. Y. Huang, E. Kok, B. Thoma, and J. J. van Merriënboer. Exploring the role of infographics for summarizing medical literature. *Health Professions Education*, 5(1):48–57, 2019.

[50] A. McCrorie, C. Donnelly, and K. McGlade. Infographics: healthcare communication for the digital age. *The Ulster medical journal*, 85(2):71, 2016.

[51] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[52] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[53] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3, pp. 850–855. IEEE, 2006.

[54] P. O'Donovan, A. Agarwala, and A. Hertzmann. Learning layouts for single-page graphic designs. *IEEE Transactions on Visualization and Computer Graphics*, 20(8):1200–1213, Aug 2014. doi: 10.1109/TVCG.2014.48

[55] J. J. Otten, K. Cheng, and A. Drewnowski. Infographics and public policy: using data visualization to convey complex information. *Health Affairs*, 34(11):1901–1907, 2015.

[56] F. Ozdamlı, S. Kocakoyun, T. Sahin, and S. Akdag. Statistical reasoning of impact of infographics on education. *Procedia Computer Science*, 102:370–377, 2016.

[57] X. Pang, Y. Cao, R. W. Lau, and A. B. Chan. Directing user attention via visual flow on web designs. *ACM Transactions on Graphics (TOG)*, 35(6):240, 2016.

[58] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollàr. Learning to Refine Object Segments. In *ECCV*, 2016.

[59] J. Poco and J. Heer. Reverse-engineering visualizations: Recovering visual encodings from chart images. *Computer Graphics Forum (Proc. EuroVis)*, 2017.

[60] C. F. Provvidenza, L. R. Hartman, J. Carmichael, and N. Reed. Does a picture speak louder than words? the role of infographics as a concussion education strategy. *Journal of visual communication in medicine*, 42(3):102–113, 2019.

[61] C. Qian, S. Sun, W. Cui, J.-G. Lou, H. Zhang, and D. Zhang. Retrieve-then-adapt: Example-based automatic generation for proportion-related infographics. *arXiv preprint arXiv:2008.01177*, 2020.

[62] J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. In *CVPR*, 2017.

[63] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *ICCV*, 2015.

[64] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

[65] B. Saleh, M. Dontcheva, A. Hertzmann, and Z. Liu. Learning style similarity for searching infographics. In *Proceedings of the 41st Graphics Interface Conference*, pp. 59–64, 2015.

[66] G. Satat, M. Tancik, O. Gupta, B. Heshmat, and R. Raskar. Object classification through scattering media with deep learning on time resolved measurement. *Optics Express*, 25(15):17466–17479, 2017.

[67] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. Revision: Automated classification, analysis and redesign of chart images. In *UIST*, 2011.

[68] H. Scott, S. Fawkner, C. Oliver, and A. Murray. Why healthcare professionals should know a little about infographics, 2016.

[69] M. J. Seo, H. Hajishirzi, A. Farhadi, and O. Etzioni. Diagram understanding in geometry questions. In *AAAI*, pp. 2831–2838, 2014.

[70] N. Siegel, Z. Horvitz, R. Levin, S. Divvala, and A. Farhadi. Figureseer: Parsing result-figures in research papers. In *European Conference on Computer Vision*, pp. 664–680. Springer, 2016.

[71] M. Smiciklas. *The power of infographics: Using pictures to communicate and connect with your audiences*. Que Publishing, 2012.

[72] R. Smith. An overview of the tesseract ocr engine. In *ICDAR*, vol. 2, pp. 629–633. IEEE, 2007.

[73] K. Todi, J. Jokinen, K. Luyten, and A. Oulasvirta. Familiarisation: Restructuring layouts with visual learning models. In *IUI*, pp. 547–558. ACM, 2018.

[74] S. Tsutsui and D. Crandall. A data driven approach for compound figure separation using convolutional neural networks. *ICDAR*, 2017.

[75] P. Vanichvasin. Enhancing the quality of learning through the use of infographics as visual communication tool and learning tool. In *Proceedings ICQA 2013 international conference on QA culture: Cooperation or competition*, p. 135, 2013.

[76] N. Westlake, H. Cai, and P. Hall. Detecting people in artwork with cnns. In *European Conference on Computer Vision*, pp. 825–841. Springer, 2016.

[77] M. J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, and S. Belongie. BAM! The Behance Artistic Media Dataset for Recognition Beyond Photography. *ICCV*, 2017.

[78] M. Xu, Y. Bai, B. Ghanem, B. Liu, Y. Gao, N. Guo, X. Ye, F. Wan, H. You, D. Fan, et al. Missing labels in object detection. In *CVPR Workshops*, 2019.

[79] Y. Yang and T. M. Hospedales. Deep neural networks for sketch recognition. *arXiv preprint arXiv:1501.07873*, 1(2):3, 2015.

[80] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao. Sketchnet: Sketch classification with web images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1105–1113, 2016.

[81] N. Zhao, Y. Cao, and R. W. Lau. What characterizes personalities of graphic designs? *ACM Transactions on Graphics (TOG)*, 37(4):116, 2018.

[82] Q. Zheng, J. Jiao, Y. Cao, and R. W. Lau. Task-driven webpage saliency. In *ECCV*, pp. 287–302, 2018.

[83] X. Zheng, X. Qiao, Y. Cao, and R. W. Lau. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics (TOG)*, 38(4):133, 2019.

[84] S. Zhu, G. Sun, Q. Jiang, M. Zha, and R. Liang. A survey on automatic infographics and visualization recommendations. *Visual Informatics*, 4(3):24–40, 2020.

[85] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013.

[86] C. L. Zitnick, R. Vedantam, and D. Parikh. Adopting abstract images for semantic scene understanding. *IEEE TPAMI*, 38(4):627–638, 2016.