SUPPLEMENTAL MATERIAL: Learning Visual Importance for Graphic Designs and Data Visualizations

Zoya Bylinskii[1] **Nam Wook Kim**[2] **Peter O'Donovan**[3] **Sami Alsheikh**[1] **Spandan Madan**[2]
Hanspeter Pfister[2] **Fredo Durand**[1] **Bryan Russell**[4] **Aaron Hertzmann**[4]
[1] MIT CSAIL, Cambridge, MA USA {zoya,alsheikh,fredo}@mit.edu
[2] Harvard SEAS, Cambridge, MA USA {namwkim,smadan,pfister}@seas.harvard.edu
[3] Adobe Systems, Seattle, WA USA {podonova}@adobe.com
[4] Adobe Research, San Francisco, CA USA {hertzman,brussell}@adobe.com

This document contains additional experimental and analysis details that were not included in the main paper. Moreover, we provide a large collection of visual results to demonstrate the various success and failure modes of our models and applications. We include also the interfaces used in our user studies, and discuss results of additional variants of the experiments that were run.

BUBBLEVIEW DATA COLLECTION DETAILS

Participants were shown blurry images of data visualizations, and were instructed to type a text caption describing the image (Fig. 1). Clicking on different parts of the image revealed small regions, or *bubbles*, of the image at full resolution. We posted 476 MTurk HITs (tasks), each consisting of 3 images randomly selected from an original set of 1411 images. An average of 15 participants completed each HIT. To accept one of our HITs, a participant had to have an approval rate of over 95% and live in the United States. A participant was paid \$0.5 for each successfully-completed HIT. We removed data corresponding to participants who provided duplicate or garbage descriptions and who clicked fewer than 10 times. Similar to Komarov et al. [10], we exclude workers whose click rates were more than $3 \times IQR$ (interquartile range) higher than the third quartile, or more than $3 \times IQR$ lower than the first quartile.

MODEL TRAINING DETAILS

The FCN-32s network was initialized with a base learning rate (lr) of $1e-05$, scaled by a factor of 0.1 every 20K iterations. A stochastic gradient descent [3] solver with a momentum of 0.9 and weight decay of 0.0005 was used, and run for 100K iterations. The FCN-16s network was initialized with the weights of the FCN-32s network and a base lr of $1e-11$ (the learning rate used on the last iterations training the FCN-32s network, scaled by 0.001). The rest of the training parameters were the same. The FCN-8s network was similarly initialized with the weights of the FCN-16s network and a base lr of $1e-17$. Our learning rate schedule was similar to the one used for semantic segmentation [12].

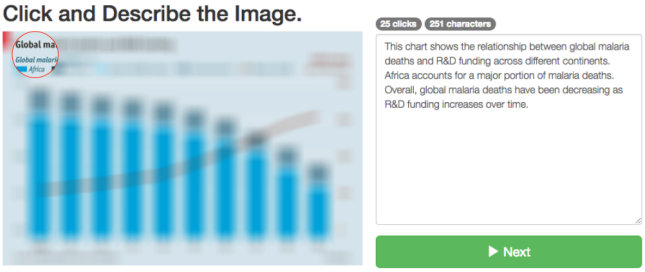


Figure 1. The BubbleView set-up from [9] that we used to collect the ground truth importance data (via BubbleView clicks) for 1.4K data visualizations.

MORE PREDICTION EXAMPLES

Fig. 2 contains more examples of predicted and ground truth importance on graphic designs. We provide a sampling of results with different performance scores. High scoring examples (Spearman's rank correlation close to 1) are ones where design elements are similarly ranked by predicted and ground truth importance. Our model can correctly distribute importance across text and visual elements. Our model can correctly predict the relative importance of different types of text (e.g. titles versus secondary text). We also show cases where model predictions disagree with ground truth. Failure cases include distributing importance across large visual elements (e.g., a face or person taking up a large portion of the image), unusual fonts, and images with many separate elements.

Fig. 3 contains more examples of predicted and ground truth importance on data visualizations. Our predicted importance localizes titles well, no matter where they are spatially located in the image. This matches ground truth data, because people also pay a lot of attention to the titles of visualizations [2]. Our

Submission # 4298

model looks for the most important text first. If a title is absent, the description, caption, or legend might be predicted as most important. Our model also learns that the most relevant points on a graph are those at the extremes (e.g., at the top or bottom of a table, left and right on a bar/line plot). This is all learned automatically from the training data, without the need for explicit text detection or a rule-based approach. Despite this, some of the failure modes of our model include assigning too much importance to salient visual regions.

FINE-GRAINED DESIGN VARIATIONS

The Design Improvement Results dataset [13] consists of 11 design templates in multiple variants, produced by MTurk workers. We used the methodology in [13] to gather Explicit Importance annotations for all 393 designs. Fig. 4 contains examples of some of these designs, along with our collected annotations, and our model predictions. Crucially, our model was not trained on systematic design variations, like changes in font, text size, or element location; nevertheless, it can correctly assign relative importance values to different design elements, as they are moved around and resized. This provides evidence that our model can provide meaningful predictions within an interactive tool setting.

COMPARISON TO RELATED WORK

Here we include all baselines from O'Donovan et al. [13], recomputed on our train-test split of the GDI dataset, compared to, and combined with, our predicted importance model (Table 1). To replicate the evaluation in [13], we report root-mean-square error (RMSE) and the R^2 coefficient, where $R^2 = 1$ indicates a perfect predictor, and $R^2 = 0$ is the baseline of predicting the mean importance value. Defining Q as the ground truth importance map and P as the predicted importance map, we iterate over all pixels i to compute:

$$RMSE(P,Q) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_i - P_i)^2} \quad (1)$$

$$R^2(P,Q) = 1 - \frac{\sum_i (Q_i - P_i)^2}{\sum_i (Q_i - \overline{Q})^2} \quad \text{where } \overline{Q} = \frac{1}{N}\sum_{i=1}^{N} Q_i \quad (2)$$

The full O'Donovan model (*OD-Full*) includes human-annotated *text*, *face*, and *person* regions. For a fair comparison, we compare our automatic predicted importance model (*Ours*) to the automatic portion of the O'Donovan model, which does not rely on human annotations (*OD-Automatic*). The addition of our predicted importance model to the *OD-Full* model, *Ours+OD*, improves performance, indicating that our original model captures some features not already captured by the other features in *OD-Full*.

EVALUATION OF RETARGETING APPLICATION

Given a graphic design and a target aspect ratio as input to retargeting, we computed an energy map, and removed image regions with lowest energy, until the desired aspect ratio was achieved (Fig. 6a-b). This is similar to seam carving [1], except instead of removing arbitrary seams, we removed only straight seams from the image. We also tried seam carving, but found that it generated significant visual distortions in graphic designs (Fig. 6c). To compare importance-based retargeting to other approaches, we used 5 variants of energy maps and a random baseline (Fig. 7). For energy maps, we used predicted importance, ground truth importance (GDI annotations [13]), Judd saliency, DeepGaze saliency, and edge energy maps. Judd saliency is a top-performing natural image saliency model [8] often used as a baseline for saliency comparisons. DeepGaze is a more recent saliency model with a neural network architecture [11], and currently a top performer on the MIT Saliency Benchmark [4]. Edge energy maps have pixel values proportional to gradient magnitudes, and were the initial energy maps used in seam carving applications [1].

Model	RMSE↓	R^2 ↑
Saliency	.229	.462
OD-Automatic	.212	.539
Ours	**.203**	**.576**
Annotations	.195	.608
Ours+Annot	.164	.725
OD-Full	.155	.754
Ours+OD	**.150**	**.769**

Table 1. A comparison of our automatic predicted importance model (*Ours*) to the importance model of O'Donovan et al. [13]. Our model outperforms the fully automatic O'Donovan variant (*OD-Automatic*). The *OD-Full* variant includes manual annotations of *text*, *face*, and *person* regions. The performance of these features is also reported separately as *Annotations*. The *Saliency* features include a learned combination of 4 separately-computed saliency models: Itti&Koch [7], Hou&Zhang [6], Judd et al. [8], and Goferman et al. [5]. Note that the first 3 rows of this table correspond to fully automatic models, while the last 4 include manual annotations. The top-performing model is bolded in each case.

MTurk details: Fig. 8 is a screenshot of our MTurk experiment and instructions for evaluating retargeting results. Each MTurk participant scored 6 retargeted design variants on a 5-point Likert scale, from 1 = very poor to 5 = very good. Participants were provided with the original design and instructed to highly rate redesigns that include the most important design elements, are legible, and not too distorted. Each participant completed the task for 12 designs, 10 randomly selected from a collection of 216 images, and another 2 validation images for ensuring quality results. The order of images and the placement of validation images in the sequence was randomized. One of the validation images contained identical retargeted designs. If a participant did not assign identical scores for these designs, all of their results were excluded from analyses. Another validation image had significantly distorted designs. If a participant did not assign poor (< 2) scores for these designs, all of their results were excluded from analyses.

We ran three versions of the experiment: (a) retargeting with straight seams, (b) retargeting with crops, and (c) banner retargeting with crops. In (a) and (b), an input image with portrait orientation was retargeted to a landscape with aspect ratio 2:3, while an input image with a landscape orientation was retargeted to a portrait with aspect ratio 3:2. These aspect ratios correspond to the standard mobile screen size, so a motivating application is given a design, to retarget it to a mobile screen.

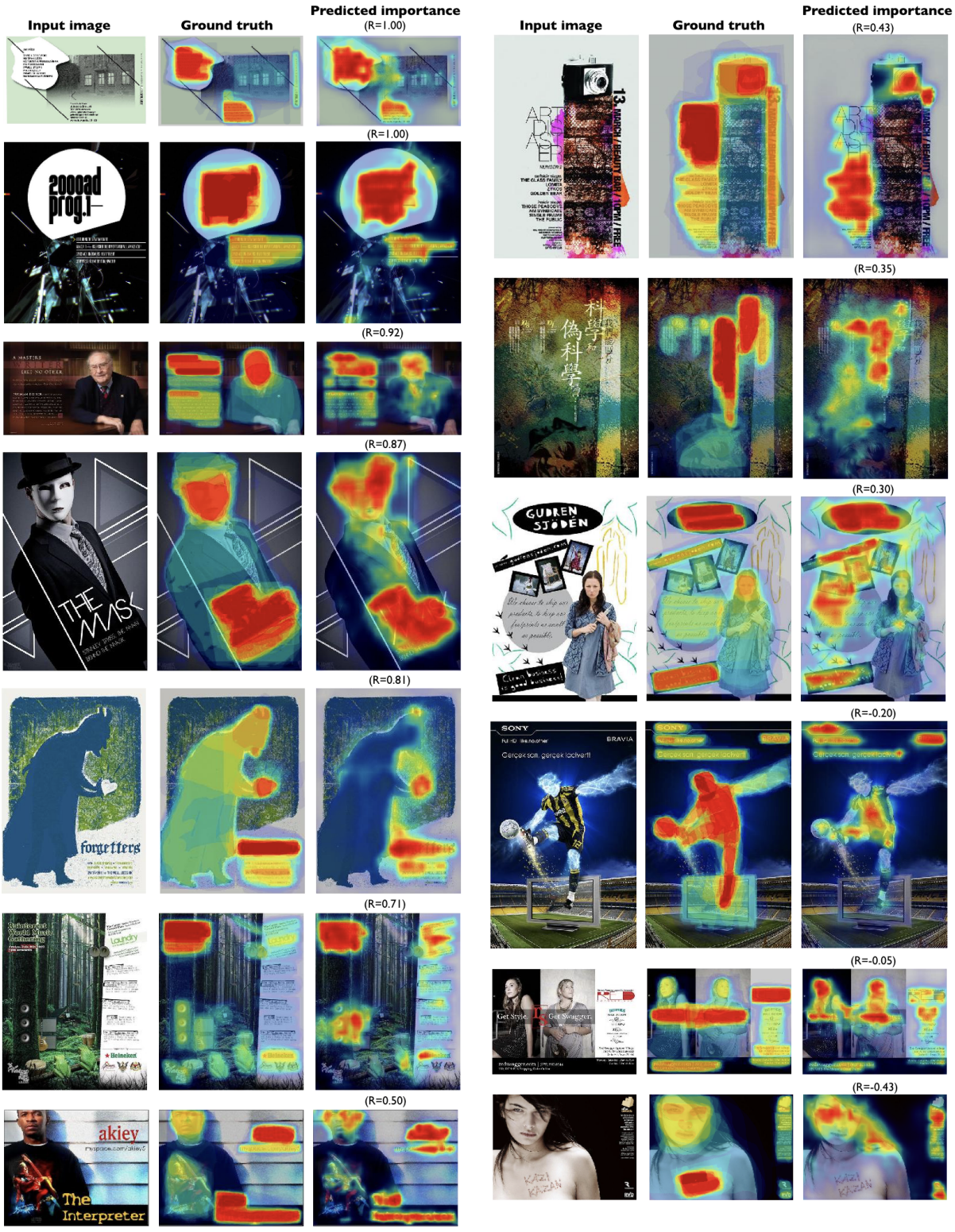


Figure 2. Examples of importance predictions for graphic designs, sorted by performance. We include both successful and unsuccessful predictions. Performance is measured as the Spearman rank correlation (R) between the importance scores assigned to design elements by the ground truth GDI annotations and by the predicted importance maps. The model is most successful when there are a few clear elements. Failures occur in predicting the importance of text written in unusual fonts; when a visual element takes up a large portion of the image (requiring reasoning about the relative importance of object parts); and when there are too many elements in the graphic design. Many of these failures can be ameliorated by training models on larger datasets.

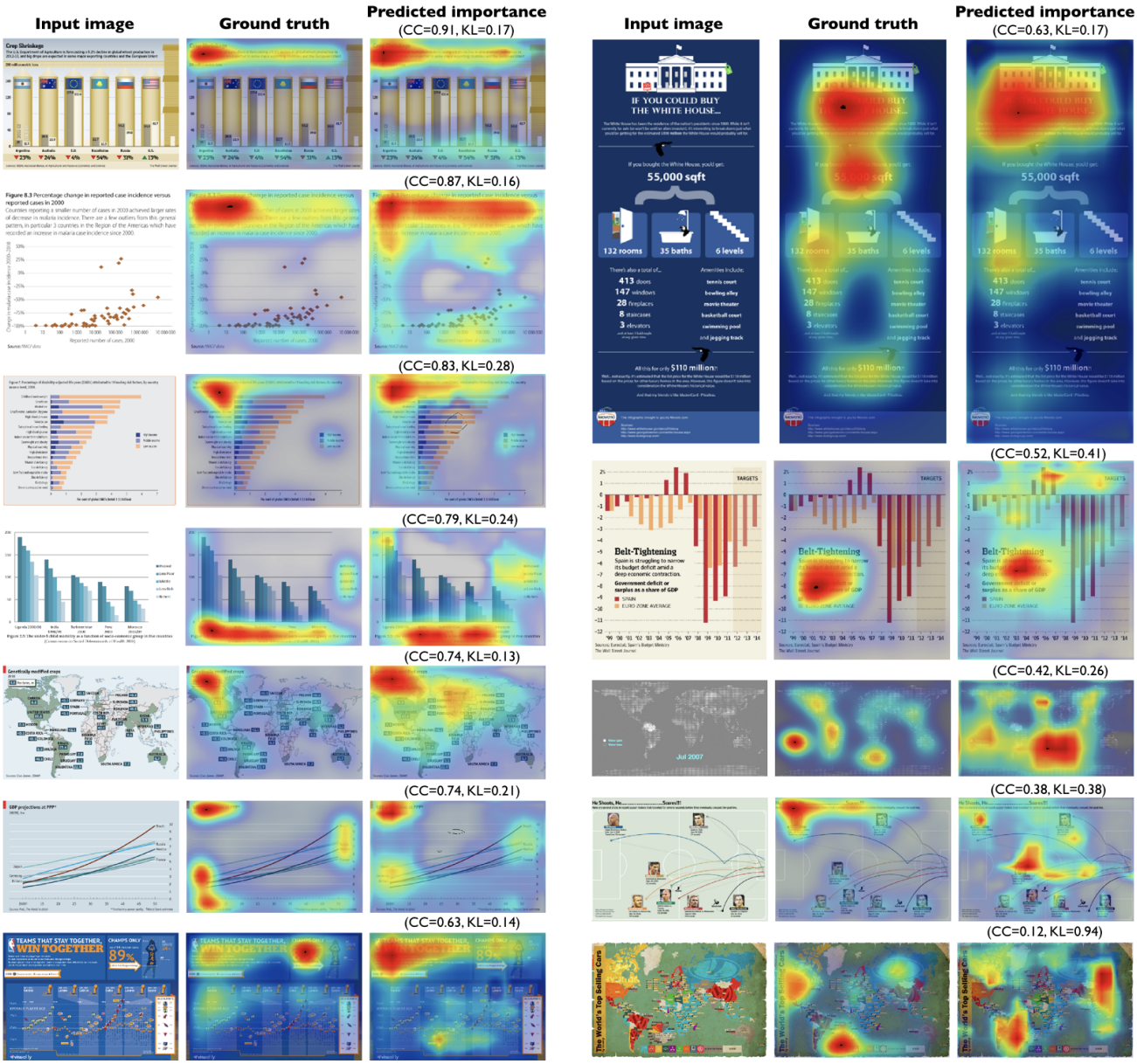


Figure 3. Examples of importance predictions for data visualizations, sorted by performance. We include both successful and unsuccessful predictions. Higher CC scores and lower KL scores are better. Our predicted importance model correctly evaluates the relative importance of different text regions, whether a title, legend, or annotation. This is learned automatically, without the need for decision trees or a rule-based approach. A title need not be located at the top of the visualization to be detected. Our model also learns that the data extremes (top and bottom of tables, left and right of graphs) are more important than the rest of the data. Some visual features continue to confuse the model and lead to some failures of prediction. We include some failures in the right column.

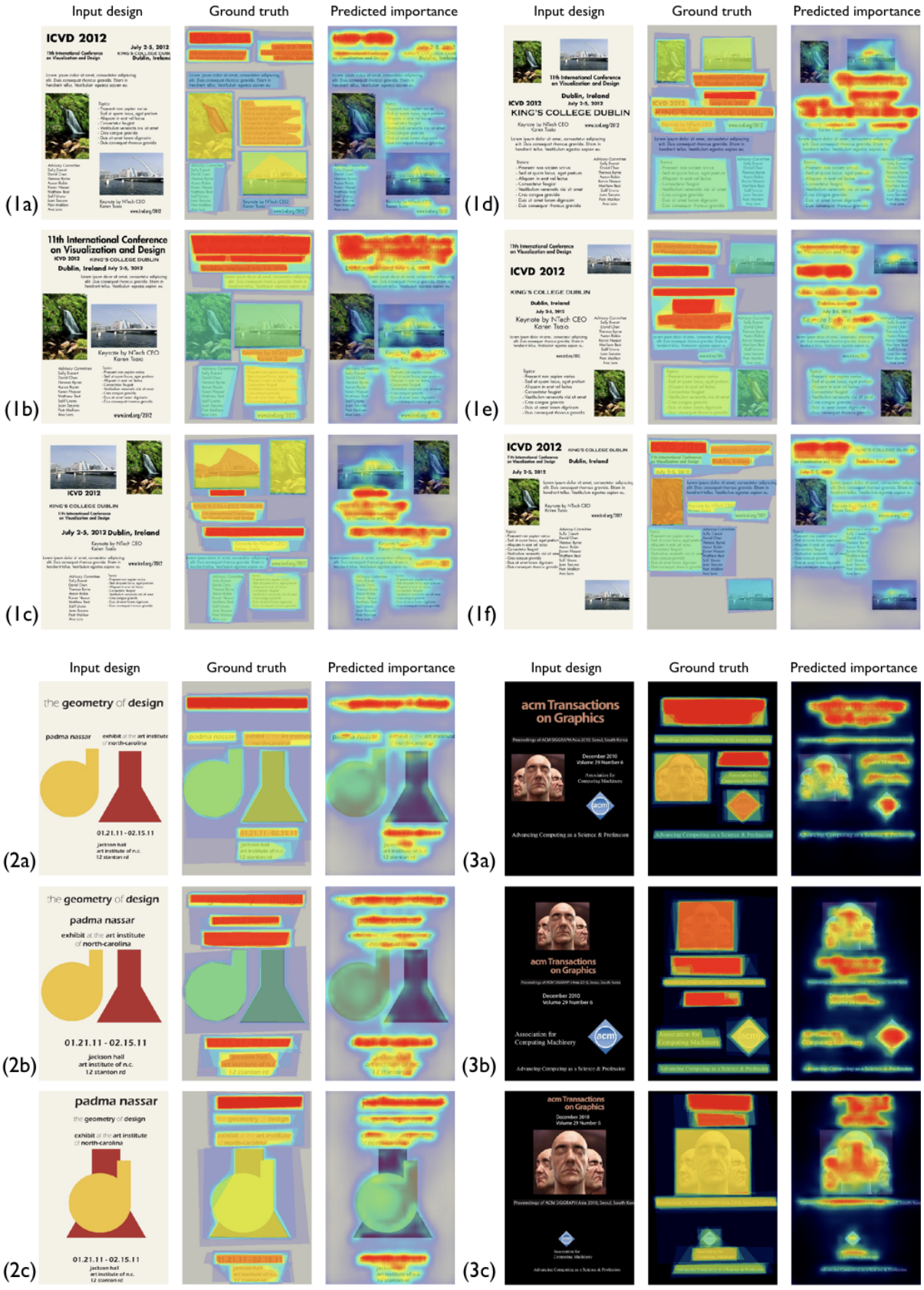


Figure 4. Examples of fine-grained design variations from [13], importance annotations we collected from MTurk participants as ground truth, and our automatic model predictions. Our model was not trained on systematic design variations; nevertheless, it can correctly assign relative importance values to different design elements, as they are moved around and resized. Example 3 shows that the model is not perfect, and can under or over-estimate the importance of various design elements, like the salient logo and the human faces.

The difference is that in (a) we carved away straight seams, in (b) we extracted crops, and in (c) we cropped all visualizations to an aspect ratio 1:4, akin to a banner for a webpage.

Results: We present the aggregate ratings for all design variants in Fig. 5. We include the total counts for each retargeted design in the 3 experiments described above. provided the mean ratings.

Experiment (a) retargeting with straight seams: A total of 143 MTurk HITs were completed, resulting in 92 HITs after filtering. Retargeting by ground truth importance achieves the highest score (Mean: 2.83), but the scores of the other 4 variants: DeepGaze, predicted importance, Judd saliency, and edge maps, were not statistically significantly different from each other at the $p = 0.05$ level. All comparisons were made using Bonferonni-corrected t-tests. Most design variants achieved relatively low scores, and upon inspection, the MTurk workers could not differentiate between design quality when straight seam carving was used. Pilot experiments with standard seam carving showed even lower scores, and so retargeting by cropping was found to be more suitable for this task.

Experiment (b) retargeting with crops: A total of 147 MTurk HITs were completed, resulting in 96 HITs after filtering. Retargeting by ground truth importance achieves the highest score (Mean: 3.27), followed by DeepGaze saliency (Mean: 3.19), and predicted importance (Mean: 3.06). However the differences between DeepGaze, predicted importance, and edge energy (Mean: 2.95) were not statistically significant at the $p = 0.05$ level. All were significantly better scoring than Judd saliency (Mean: 2.78) and the random jumbled baseline (Mean: 1.24).

Experiment (c) banner retargeting with seams: A total of 146 MTurk HITs were completed, resulting in 90 HITs after filtering. Differences between the retargeting variants are larger compared to experiments (a-b) because the cropping is more aggressive, requiring a more careful selection of the important design regions to include in the retargeted result. As reported in the main paper, retargets obtained using ground truth importance had the highest score (Mean: 3.19), followed by DeepGaze (Mean: 2.95) and predicted importance (Mean: 2.92). However, the difference between the latter two models was not statistically significant. Edge energy maps (Mean: 2.66) and Judd saliency (Mean: 2.47) were significantly worse, but not statistically different from each other. The random crop baseline (Mean: 2.23) was significantly worse than all other methods. One notable difference in the random baselines between (a-b) and (c) is that in the case of the first two, a jumbled image was used (broken up into 6 rectangular blocks), as in the bottom left of Fig. 8. In the case of (c), a random crop was taken, by selecting a random image coordinate. As a result, the random jumbled baseline in (a-b) tends to be significantly worse than the random crop baseline in (c).

Summary: Across all three experiments, retargeting based on ground truth importance consistently received the highest scores, indicating it can capture the relevant regions of a graphic design. Both our predicted importance and DeepGaze saliency performed similarly, but worse than the ground truth importance. Both models are neural network models that attempt to capture observer attention patterns on images. While predicted importance was trained on graphic designs and DeepGaze on natural images, the latter model was trained on 10x more data, and may be able to generalize to graphic designs as a result of learning image statistics from a larger collection of images.

EVALUATION OF THUMBNAILING APPLICATION

Given a visualization as input, we generate thumbnails using straight seam carving and blending. Qualitatively, we found that the straight seam carving worked for the structured visualization images - e.g., removing the middle rows of tables, removing the clusters of data points near the middle of the plots, etc. - while preserving the spatial relationships between visualization elements like axes. Some examples of automatically-generated thumbnails are in Fig. 9.

MTurk details: We ran an MTurk task where given a description and a grid of thumbnails, the goal of participants was to find the visualization corresponding to the description (Fig. 10). Clicking on a thumbnail displayed a pop-up window with an enlarged version of the visualization (this provided additional disinsentive to click around randomly, as it would slow down task completion). Only when the correct image was clicked on, would the task end.

This task was intended to imitate a search through a database of visuals to determine if our thumbnails can facilitate this search. A single MTurk HIT consisted of finding the matching visualization for a specific description. We selected a total of 13 user-generated descriptions from our BubbleView data collection. For a given HIT, we randomly selected a description and a set of 60 images to show on the screen, in a 20x3 grid. These images were randomly sampled from our 202 test set of visualizations (except for the 1 image matching the description). We ran two versions of the study: (a) with the original visualizations resized to thumbnails, and (b) with our automatically-computed importance-based thumbnails. In both cases, we measured how many clicks it took for participants to find the visualization matching the description.

We employed the interquartile range (IQR)-based outlier removal procedure from Komarov et al. [10] in order to exclude experimental runs where the number of clicks generated was more than 3xIQR higher than the third quartile, or more than 3xIQR lower than the first quartile. A total of 223 MTurk HITs were completed for experiment version (a), which after the outlier removal procedure, produced 200 HITs for analysis. A total of 182 HITs were completed for version (b), resulting in 169 HITs after outlier removal.

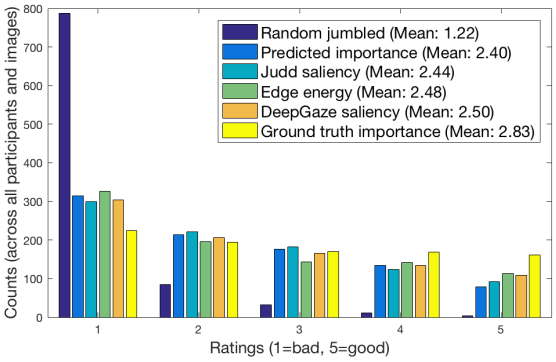
Results: We measured how many clicks it took for participants to find the right visualization with the resized visualizations (Mean: 3.25 clicks, Median: 2 clicks), and the importance-based thumbnails (Mean: 1.96 clicks, Median: 1 click). Each MTurk assignment, containing a single description search task assigned to a single participant, was treated as a repeated observation. The difference in the mean number of clicks was statistically significant at the $p = 0.001$ level.

We repeated this task with thumbnails computed using ground truth importance. We again ran two versions of the study: (a) with the original resized visualizations (191 total HITs, 178 after filtering), and (b) with importance-based thumbnails (209 total HITs, 201 after filtering). The total clicks required to find the visualization corresponding to the description was again higher for the resized visualizations (Mean: 3.38 clicks, Median: 2 clicks) than the importance-based thumbnails (Mean: 1.90 clicks, Median: 1 click), statistically significant at the $p = 0.001$ level.

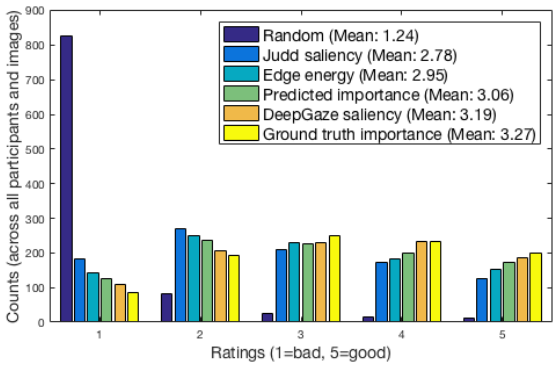
These results demonstrate that our importance-based thumbnails captured visualization content that was relevant for retrieval. Moreover, thumbnails generated using predicted importance were sufficiently effective for this task, not far from ground truth importance.

REFERENCES

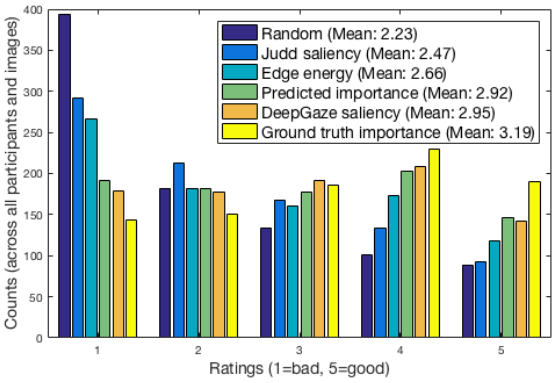
1. Shai Avidan and Ariel Shamir. 2007. Seam Carving for Content-aware Image Resizing. *ACM Trans. Graph.* 26, 3, Article 10 (July 2007). DOI: http://dx.doi.org/10.1145/1276377.1276390
2. Michelle A. Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S. Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2016. Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan 2016), 519–528. DOI: http://dx.doi.org/10.1109/TVCG.2015.2467732
3. Léon Bottou. 2004. Stochastic learning. In *Advanced lectures on machine learning*. Springer, 146–168.
4. Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. 2012. MIT Saliency Benchmark. (2012).
5. Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. 2012. Context-Aware Saliency Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 10 (Oct 2012), 1915–1926. DOI: http://dx.doi.org/10.1109/TPAMI.2011.272
6. Xiaodi Hou and Liqing Zhang. 2007. Saliency Detection: A Spectral Residual Approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8. DOI: http://dx.doi.org/10.1109/CVPR.2007.383267
7. Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (Nov 1998), 1254–1259. DOI: http://dx.doi.org/10.1109/34.730558
8. Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*. 2106–2113. DOI: http://dx.doi.org/10.1109/ICCV.2009.5459462
9. Nam Wook Kim, Zoya Bylinskii, Michelle A. Borkin, Krzysztof Z. Gajos, Aude Oliva, Frédo Durand, and Hanspeter Pfister. 2017. BubbleView: an alternative to eye-tracking for crowdsourcing image importance. *CoRR* abs/1702.05150 (2017). http://arxiv.org/abs/1702.05150 *Submitted to TOCHI*.
10. Steven Komarov, Katharina Reinecke, and Krzysztof Z Gajos. 2013. Crowdsourcing performance evaluations of user interfaces. In *SIGCHI*. ACM, 207–216.
11. Matthias Kümmerer, Lucas Theis, and Matthias Bethge. 2014. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. *CoRR* abs/1411.1045 (2014). http://arxiv.org/abs/1411.1045
12. Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2017. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (April 2017), 640–651. DOI:http://dx.doi.org/10.1109/TPAMI.2016.2572683
13. Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2014. Learning Layouts for Single-PageGraphic Designs. *IEEE Transactions on Visualization and Computer Graphics* 20, 8 (Aug 2014), 1200–1213. DOI: http://dx.doi.org/10.1109/TVCG.2014.48



(a)



(b)



(c)

Figure 5. Three study versions of retargeting: (a) retargeting with straight seams, (b) retargeting with crops, and (c) banner retargeting with crops. In each study, we compared 6 methods of retargeting, based on different energy maps. The mean scores of each method are provided in the legend; the aggregate counts of each score are plotted.

Figure 6. Additional retargeting variants: (a) retargeting with cropping, (b) retargeting with straight seams, (c) retargeting by seam carving.

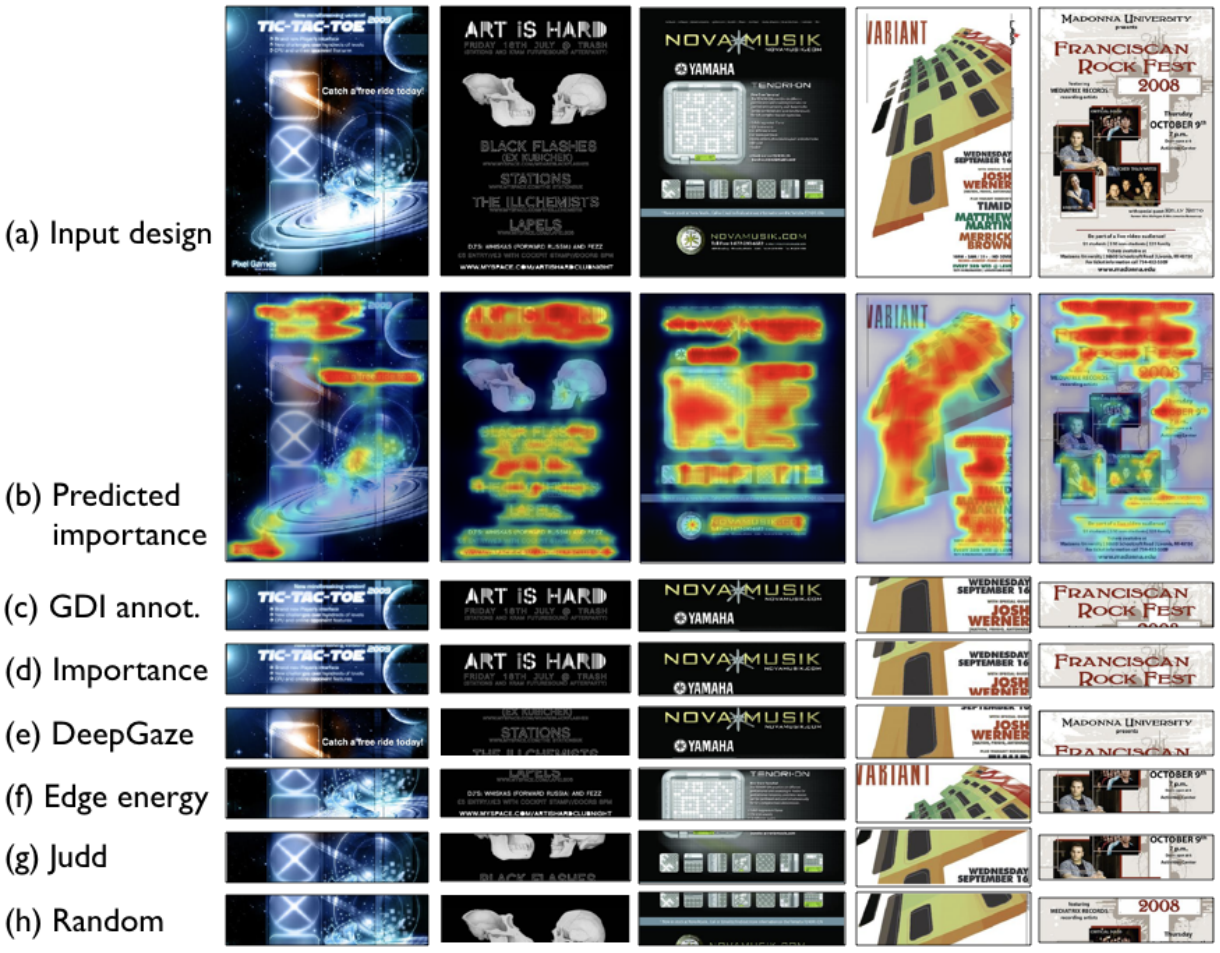


Figure 7. (a) Input design. (b) Predicted importance maps. Retargeted results using: (c) ground truth GDI importance annotations [13], (d) predicted importance map, (e) DeepGaze saliency [11], a top-performing neural network saliency model, (f) image gradient magnitudes, (g) Judd saliency [8], a commonly-used natural image saliency model, (h) a random crop baseline.

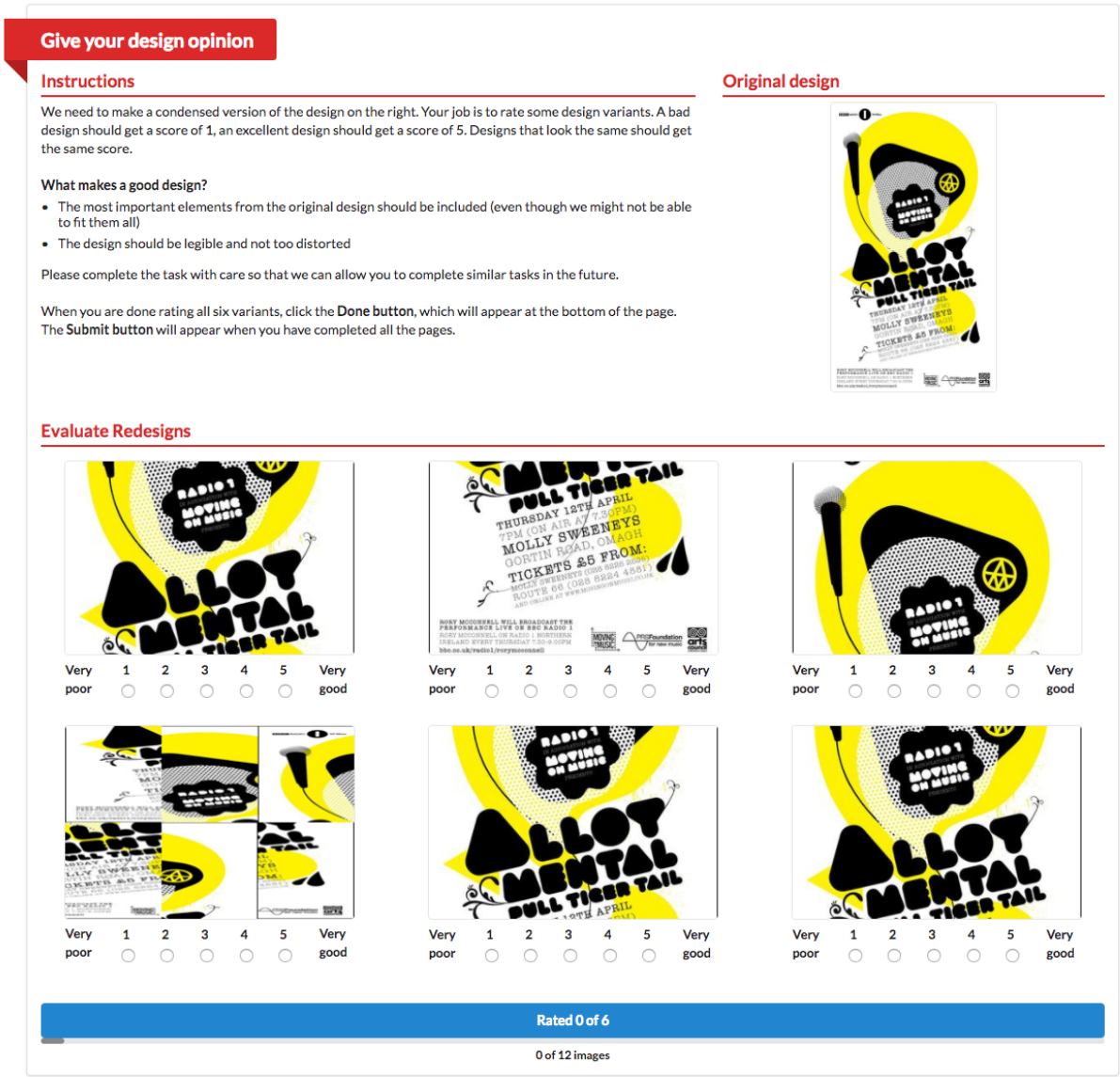


Figure 8. Screenshot of the retargeting experiment. Participants are given a design and 6 redesign variants, and their goal is to rate the quality of each redesign on a 5-point Likert scale. Instructions ask participants to more highly rate redesigns that contain the most important content from the original design, and those that are legible and not distorted. The redesigns are obtained by retargeting the original design using 5 different input energy maps. A random jumbled baseline is used to validate that participants are completing the task correctly.

Figure 9. More examples of (a) input data visualizations and (b) corresponding automatically-generated thumbnails. The extremes of the data are predicted important, and the middle regions and data points are removed during thumbnailing. The boundaries of the remaining regions are blurred using the importance map as an alpha-mask with a fade to white.

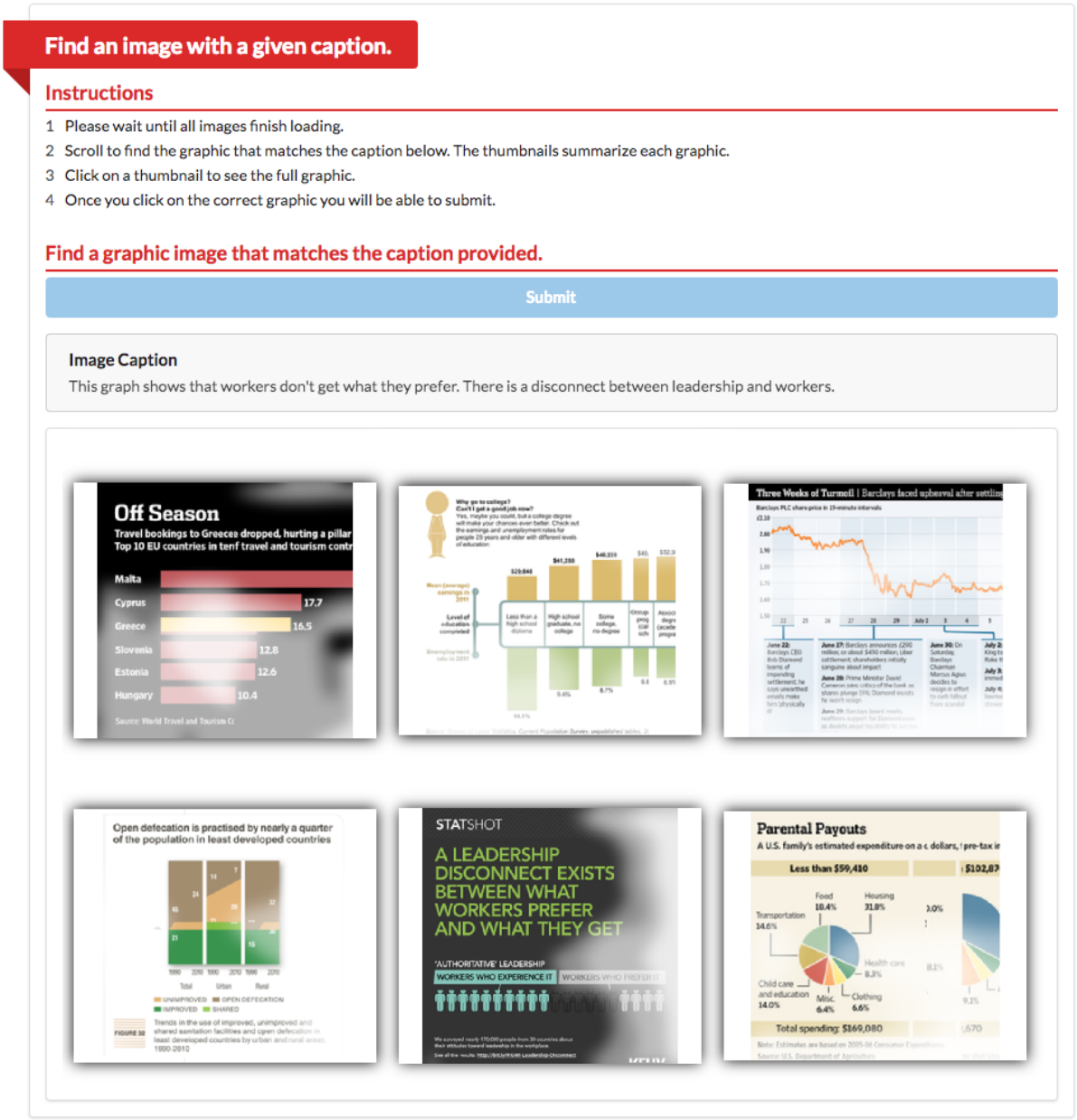


Figure 10. Screenshot of the the thumbnail search task. Participants are given an image caption and instructed to scroll through a list of 60 thumbnails to find the data visualization matching the caption. Clicking the correct thumbnail ends the task. Clicking the wrong visualization brings up a modal window with the full sized visualization, which slows down the task, and discourages participants from clicking around randomly. This task is intended to simulate a search through a database, and we measure the effectiveness of importance-based thumbnails at facilitating the search, measured as the number of clicks until the correct visualization is found.