資料驅動互動式回饋以輔助高易用性圖標設計

王郁婷
國立台灣大學
r08944019@ntu.edu.tw

程芙茵
國立台灣大學
fuyincherng@gmail.com

沈奕超
東京大學
jdilyshen@gmail.com

陳育淳
國立台灣大學
lesleychen@ntu.edu.tw

張元嘉
京都大學
riceball0907@gmail.com

林文杰
國立陽明交通大學
wclin@cs.nctu.edu.tw

陳炳宇
國立台灣大學
robin@ntu.edu.tw

ABSTRACT

圖標在應用軟體、網頁、和各式各樣的使用者介面上是重要的元素之一，並且經常以整組的圖標一起設計、使用。然而，設計師通常僅僅通過公司、團隊內部非正式的測試來評估新設計的圖標、圖標組的容易使用程度。即便有些較常見的功能已有具代表性的圖標樣式（例如：搜尋、下一步），但是在不同案例中仍然有許多尚未建立代表性圖標的功能的設計需求（例如：封存）。不僅如此，因為介面設計在時間和預算上的限制，設計師鮮少針對每個圖標去做正式的（具有足夠受測者的）易用性測試。因此，我們提出EvIcon，一個互動式設計評估工具來提升圖標設計反覆修改和評估階段的效率，並且針對新設計的圖標提供兩種即時資料驅動回饋。

首先，我們透過群眾外包收集大量群眾對於不同圖標的感知程度（包含語義相關性和熟悉程度）評分（共收集到$62,649$ 筆評分資料）並用以訓練深度學習模型來針對使用者上傳的圖標提供即時的容易使用程度回饋。接著，我們利用收集到的圖標資料庫（$n = 2,000$）以及孿生神經網路（Siamese Neural Network）來輔助圖標組的設計能達到足夠的視覺區分程度。我們透過新手及專業介面設計師的使用者試驗及訪談，展示EvIcon對於圖標設計反覆修改和評估階段的幫助和成效。在後續的群眾外包實驗中，可以看到在EvIcon輔助下設計的圖標對比沒有EvIcon輔助的圖標在語意相關性和熟悉程度上均達到較佳的成效。

CCS Concepts

•Human-centered computing → Interactive systems and tools;

Keywords

Icon Design, Icon Set, Crowdsourcing, Interactive Assistive Tool, Computational Design

1. INTRODUCTION

Amid the rapid development and near ubiquity of digital technologies, including computers, intelligent appliances, and wearable devices, interface icons play an increasingly important role in representing various functions with the benefits including improving scannability of interface (i.e., the ease of reading and understanding the content of interface), save space on small screens, and convey information universally [46, 47]. The usability of icons is determined by several characteristics [39, 20] (e.g., visual complexity, style, familiarity, etc.). Existing design guidelines (e.g., Google's Material Design) provide designers implications of icon design regarding the visual characteristics. However, collecting users' perceptual feedback of the icons is still an irreplaceable step to assess the icon's usability [4]. Yet, conducting formal usability testing can be time-consuming and required extra effort [44, 13, 51], which could significantly lengthen the iterative process of icons and interface design. When evaluating icons designed for specific users (e.g., elders or users with lower computer literacy), conducting adequate usability testing is even more laborious.

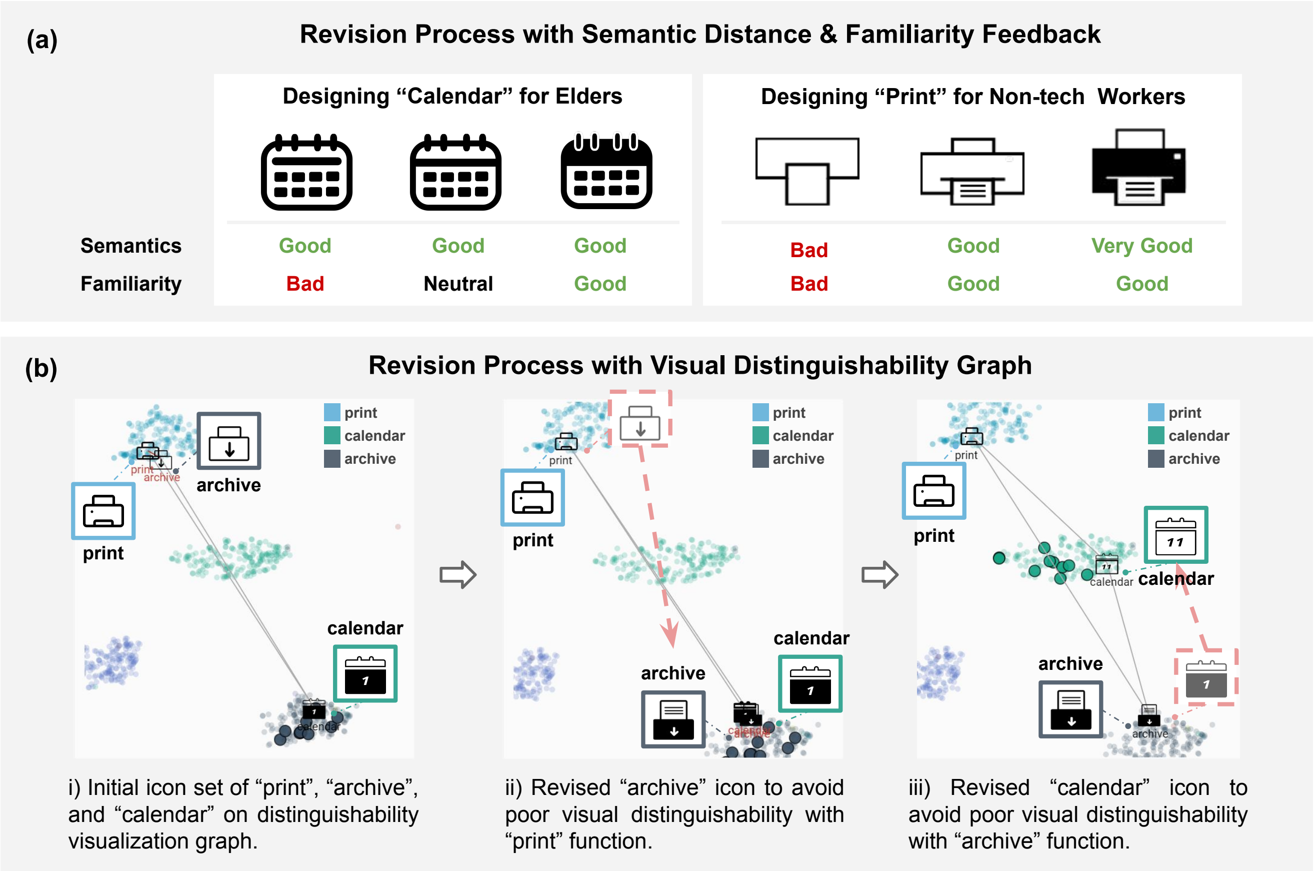
Through interviews with professional UI designers, Zhao *et al.* [56] reported that designers often conducted internal and informal evaluations by consulting co-workers' opinions and feedback on icons. These informal evaluations often failed to provide comprehensive and objective information about how target users would perceive and use the icons [4, 44]. This finding underscores the need for an objective and comprehensive usability testing approach but is cost- and time-efficient. However, most prior works focus on the applications, such as proposing novel computational model for learning icons' appearance similarity [27], exploring the layout of different mobile app's icon [35], and designing compound icons using text [56]. To the best of our knowledge, there has been little research focusing on assisting icon design validation and lower the cost of conducting usability testing of interface icons.

In this work, we present an interactive design tool, EvIcon, to ease and accelerate validating interface icon set in the iterative design process by providing two types of instant feedback. In Fig. 1, we show examples of these two types of feedback on some example icons and their evolutions along with the icons' revision process. First, EvIcon provides predicted users' perception of the icon's semantic distance and familiarity. Semantic distance stands for the degree of closeness between an icon and the function it represents [38, 46,

CGW '21, July 12–13, 2021, Keelung City, Taiwan

© 2021 ACM. ISBN .

DOI:


(a)
Revision Process with Semantic Distance & Familiarity Feedback
Designing “Calendar” for Elders
Designing “Print” for Non-tech Workers
Semantics
Familiarity
Good
Bad
Good
Neutral
Good
Good
Bad
Bad
Good
Good
Very Good
Good
(b)
Revision Process with Visual Distinguishability Graph
print
calendar
archive
print
archive
calendar

i) Initial icon set of “print”, “archive”, and “calendar” on distinguishability visualization graph.

ii) Revised “archive” icon to avoid poor visual distinguishability with “print” function.

iii) Revised “calendar” icon to avoid poor visual distinguishability with “archive” function.

Figure 1: EvIcon provides two types of feedback for supporting instant usability evaluations. (a) With the “Semantics” (abbreviation for semantic distance) and “Familiarity” feedback, designers can improve the icons’ usability targeting different demographic users (e.g., elder people or non-tech workers). (b) Designers can improve the visual distinguishability and the overall usability of an icon set with the provided interactive graph of distinguishability visualization. The bigger nodes with black stroke represent the top-10 high-usability icons in our dataset.

54], and familiarity referred to the frequency with which icons are encountered [38, 39, 21]. In the Fig. 1(a), we show examples of icons with different semantic distances and familiarity levels. They are both critical indications of good icon design, and using icons with close semantic distance and high familiarity can increase user performance on interfaces [38, 46, 54, 47, 8]. Due to the importance of these two indications for icon's usability [46, 37, 47], we focus on providing icon designers with semantic distance and familiarity predictions on icon designs in this paper. The predictions are generated by deep learning models trained on a large icon dataset with $62,649$ crowdsourced ratings. Additionally, we grouped crowdworkers by their demographics to simulate particular types of target users' perceptions of icons. Apart from offering the predicted level of users' perceptions, the second feedback EvIcon provides is individual icons' visual distinguishability to others in (i) our large icon dataset and (ii) the icon set provided by designers (See Fig. 1(b)). This feedback is realized by providing an interactive two-dimensional embedding visualization learned by a Siamese neural network [5]. This visualization provides designers with a holistic overview of icons' visual relationships and allows them to explore visually similarities between icons.

To understand the benefits of EvIcon for designers, we conducted an evaluation with six user interface (UI) designers with different professional levels. We asked them to improve icons with and without using EvIcon. The results of the post semi-structured interview show that all designers were optimistic about using EvIcon in the process of improving the icons. The instant feedback of users' perception helps them quickly identify the essential visual components of icons associated with close semantic distance and high familiarity. We further conducted an online crowdsourcing study to verify whether EvIcon can assist designers improve icons' usability. The result shows that the revised icons generated by the novice designers with the assistance of EvIcon have better performance of semantic distance and familiarity than those without using EvIcon. This proves that EvIcon is an effective tool to assist UI designers refining icons through the iteration and validation process.

The contributions of this work are summarized as follows.

- We propose a framework for assisting interface icon set design iteration and validation process by providing data-driven feedback of users' perception and visual distinguishability.
- We demonstrate that the critical criteria of icon design such as semantic distance, familiarity, and visual distinguishability can be modeled using data-driven approaches, such as convolutional and Siamese neural network. The reusable pre-trained computational models become a low-cost and efficient approach to obtain users' perceptual feedback without conducting formal usability testing.
- We implemented an interactive design tool based on the proposed framework and demonstrate the effectiveness of EvIcon by conducting an evaluation study. We verified that the resulting icon designs achieved better perception performance through an additional crowdsourcing study.

2. RELATED WORKS

2.1 Icon Design and Analysis

Icon plays an essential role in visual communication, including graphic design and user interface design. Prior studies [17, 19, 18] provide a thorough introduction on how to design icons and recommended practices. Icon's usability mainly associates with the ability to convey the information it represents. Previous research identified several features that heavily influence icon's usability, including visual complexity, semantic distance, and familiarity [39, 37, 47, 21]. Some studies reported that the users' age [30] and experience [20, 2] also influence the effect of these features on icon's usability.

Researchers have proposed various methods to support icon design and generation due to the complex relationship between icons' features and usability. For example, Zhao *et al.* [56] developed a system to generate icons containing compound meanings automatically. Some works focus on generating icons based on file-names [31], data content [24], and man-made object category [48]. Liu *et al.* [36] proposed a system that synthesizes novel icons by remixing portions of icons retrieved from large online repositories. Lagunas *et al.* [27] learned icons' appearance similarity to automatically recommend other icons that have coherent style and visual identity with the given query icon.

Unlike other previous works focus on generating a single icon at a time, Laursen *et al.* [28] proposed a crowdsourcing-based method for selecting an icon set among sets of candidate icons. They recruited many crowdworkers and asked them to evaluate icons for comprehensibility and identifiability. Finally, they design an optimization method to optimize an optimal icon set based on the collected crowdsourced ratings. Our system shares the same spirit that we also aim for designing an icon set based on crowdsourced perception data. Moreover, we take a step further to assist the users in refining their icon set design based on the feedback learned from crowdsourced perception data. Compared to the work of Laursen *et al.* [28], our system lets users design the final icon set on their own with our data-driven feedback and guidance instead of obtaining an icon set from an optimization process. Moreover, the provided suggestive design feedback and the human-in-the-loop workflow increase the accountability and flexibility of perceptual prediction tools for designers [44].

2.2 Crowdsourced Human Computation

Human processors are often realized through microtask-based crowdsourcing services, such as Amazon Mechanical Turk (AMT)[1], CrowdWorks[2], and Lancers[3]. The *human computation algorithm* concept is proposed in [34] where crowdsourced human processors and other algorithms are hybrid together as function calls. Previous works extend this concept to solve perceptual computer vision problems [16] and tweaking parameters in various design scenarios [26]. Prior studies in HCI also demonstrated the feasibility of conducting usability evaluation on crowdsourcing platform via performing benchmark user testings [25], collecting human visual importance [6], and building pipelines to gather crowdsourced dataset on performance and annotations of mobile apps [13, 11, 35].

[1] https://www.mturk.com/

[2] http://www.crowdworker.com/

[3] https://www.lancers.jp/

Our system follows the human computation paradigm by collecting crowdsourced perception data and modeling them using deep learning models. The major difference from the previous method is that as the users iteratively revise their icon designs, they do not need to resort to additional crowdsourcing tasks but use the model we trained on the collected crowdsourced perception data.

2.3 Assistive Authoring Tool for Visual Design

Assistive visual content authoring has gained increasing interest in the past few years since the surge of the need for novel visual content. Many works utilize the personal editing histories to assist 2D sketch [55], 3D shape sculpturing [42], and viewpoint selection [7]. On the other hand, various prior works have incorporated real-time physical simulation into their interactive tools for designing physically valid furnitures [52] and model airplanes [53]. Among them, many recent works leverage collected visual content data to assist 2D sketch [29], multi-view clipart design [49], and mobile apps user interface design [35, 12, 14]. Other studies crowdsourced and modeled large-scale users' perception about tappability for the mobile interfaces [51] and visual importance on graphic designs [6] to assist designers in diagnosing the perceptual issues in their designs. Rosenholtz *et al.* [44] conducted a thorough qualitative study with professional design teams to verify if the perceptual prediction tools aid the design process. Their results showed that design teams and designers benefited from such tools in the agile assessment of usability and communication in cross-functional teams. With the extensive examples showing the benefits of adopting assistive authoring tools in visual design, we consider the proposed framework and EvIcon have great potential in assisting interface icon design, especially in the iteration and validation process.

3. EvIcon

We propose EvIcon, an interactive and exploratory tool to present perceptual feedback of individual icons and visual distinguishability between icons in an icon set. Instead of offering an alternative tool to the existing vector graphics tools (e.g., Illustrator [1] and Sketch [50]), our main goal is to facilitate efficient usability validation for iterative refinement of icon design by providing data-driven feedback. We implemented EvIcon as a web-based system as shown in Fig. 2 for its compatibility and broad reach across different types of devices. In this section, we provide an overview of the system and the user scenario. The dataset collection and computational models that power the system are discussed in Sec. 4 and Sec. 5. The interface of EvIcon contains four main panels: (i) the main canvas panel which includes an editor for icon modification and a list to present the uploaded icon set, (ii) icon suggestion panel, (iii) perception feedback panel, and (iv) distinguishability visualization panel. We show the screenshot of EvIcon user interface in the supplemental material.

3.1 Main Canvas Panel

The main canvas located in the center comprises two modules, including (i) a list that presents the icon set uploaded by designers and (ii) a scalable vector graphics (SVG) editor to support continuous revision and submission. We customized the SVG editor from an online Javascript SVG editor papergrapher[4] to provide basic SVG editing functionalities during revision. To perform the usability testing using EvIcon, users first need to upload the initial version of the icon set to be evaluated. Next, the users can select an icon from the set and click the "evaluate" button to get the predicted perception to the selected icon and distinguishability feedback to the uploaded icon set. In the iterative revision process, in-place warnings will be presented once the predicted perception level (semantic distance and familiarity) dropped. The purpose of the warning is to draw user's attention to the perceived usability fall off and build the connection between perceptual prediction and the icon image. We implemented two types of in-place warnings, including visual warning and hint text as shown in Fig. 3(c). The goal of the visual warning is to highlight the poor adjustments compared to the last usability inspection. We highlighted the paths that we encourage the users to add in light blue and those to remove in dark blue. We specifically avoided using red and green to prevent misunderstanding of the suggestion because of their established meanings (e.g., green means go and red means stop). At the same time, hint text would pop up on the editor to inform the types of decreased perceptions as demonstrated in Fig. 3(c).

3.2 Icon Suggestion Panel

Based on the crowdsourced perception data, we present the top N ($N = 10$ in the current implementation) icons representing the identical function with the selected one as design references. The top N icons are ranked by the score of semantic distance plus the score of familiarity (Score = Mean(crowdsourced ratings of perceptions) × Standard Deviation(crowdsourced ratings of perceptions)). We also provide the overall semantic distance and familiarity level of each icon in the suggestion list for comparison. To present the levels of semantic distance and familiarity in a way designers can easily understand, instead of showing rating numbers directly, we use "Very Bad", "Bad", "Neutral", "Good", and "Very Good" to represent five different levels of user perceptions, and semantic distance is presented as "Semantics" on the interface of EvIcon. We highlighted "Very Bad“ and "Bad" in red, "Neutral" in black, and "Good" and "Very Good" in green to enhance readability.

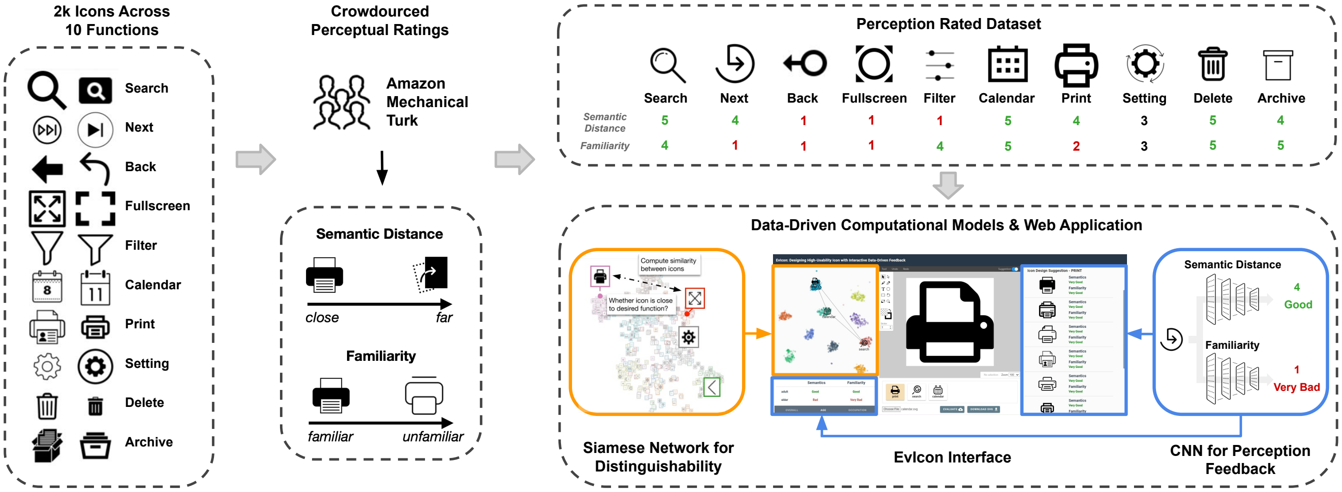
3.3 Perception Feedback Panel

In this panel, EvIcon allows the users to view the predicted level of perceptual usability of the selected icon (see Fig. 4). As shown in Fig. 4(a), we used pre-trained deep learning models to predict the expected semantic distance and familiarity of the icon. Moreover, the users can check the predicted semantic distance and familiarity for target audiences with particular demographics. The users can switch to "age" (Fig. 4(b)) and "occupation" (Fig. 4(c)) tabs to inspect the additional perception predictions. About how to collect the dataset of perceptual usability and construct computational models will be explained in Sec. 4 and Sec. 5.

3.4 Distinguishability Visualization Panel

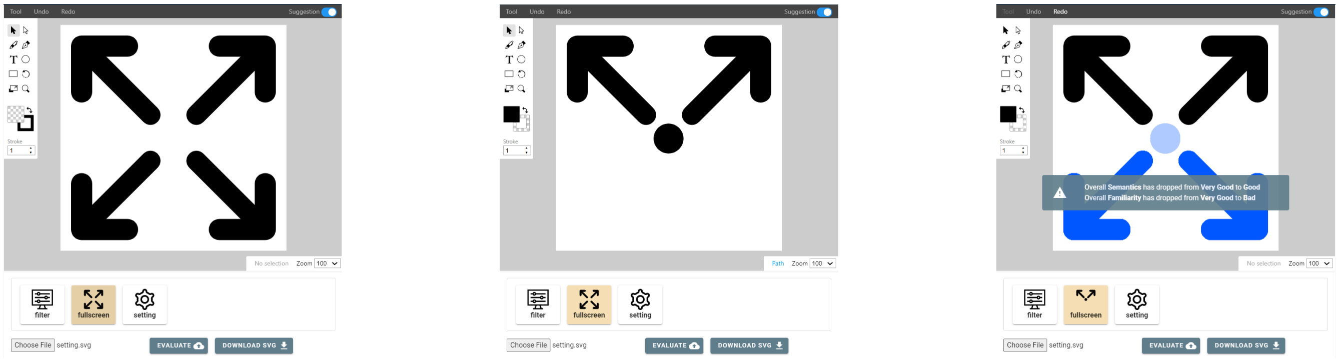
EvIcon presents interactive distinguishability visualization to help designers compare the relative visual distance between icons in (i) the uploaded icon set and (ii) other icons in the collected icon dataset. Color-coded nodes representing different functions are plotted on the graph according to

[4] https://github.com/w00dn/papergrapher



	Search	Next	Back	Fullscreen	Filter	Calendar	Print	Setting	Delete	Archive
Semantic Distance	5	4	1	1	1	5	4	3	5	4
Familiarity	4	1	1	1	4	5	2	3	5	5

Figure 2: An overview of EvIcon. A dataset comprises 2,000 icons across 10 functions (left) are labeled with semantic distance and familiarity level by crowdworkers (center). EvIcon computes and presents designers with data-driven feedback to assist designing of high-usability icon sets based on the perception labeled dataset (right).



Overall Semantics has dropped from Very Good to Good
Overall Familiarity has dropped from Very Good to Bad

(a) Initial icon.

(b) Modified icon.

(c) Two types of in-place warnings, including visual warning and hint text.

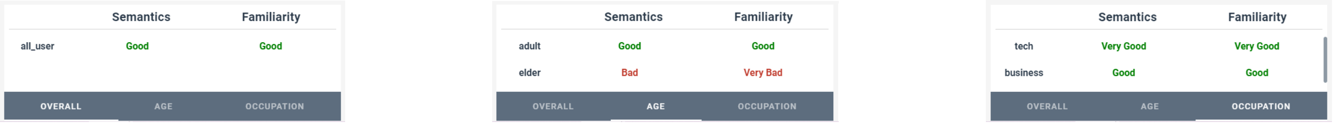
Figure 3: In-place warning to highlight the poor adjustment compared to the last usability inspection.

their calculated coordinates. Description of computational models that support the calculation of the coordinates is provided in Sec. 5.2. The icon image and its function name would be shown on a pop-up tooltip once users hover on the node. The elements mentioned above form the basic interactive graph as demonstrated in Fig. 5(a) in which the icons were roughly grouped into 10 clusters. We can observe from the Fig. 5(a) that the majority of icons in a cluster represent the same function with the same color. This finding indicates that most icons representing the same functions are visually distinguishable from icons of other functions in the icon dataset we collected. Therefore, users can know if an icon would be easily misrecognized as incorrect functions by observing its location on the graph of distinguishability visualization.

Whenever users click "evaluate" after selected an icon in the uploaded icon set, the images and function names of all icons in the set would be shown on the scatter plot at their calculated coordinates. At the same time, the graph would be zoomed in automatically to the local view in which the selected icon is centered so that users can browse the nodes nearby easier (see Fig. 5(b)). We highlighted the top-K suggested icons described in Sec. 3.2 in bigger nodes with black stroke for ease of searching and comparison. By double-clicking on the blank canvas, users can switch back to the global view as shown in Fig. 5(c). We fully connected the icons in the uploaded icon set using grey links (as shown in Fig. 5(c)) and change the color of the links into red if the connected icons are too close to each others (see Fig. 5(d)). This design aims to provide users information about the potential problem of inadequate visual distinguishability in the icon set.

4. DATASET COLLECTION AND CROWD-SOURCED PERCEPTUAL RATING

To provide feedback for icon's perceptual usability, collecting an dataset with rating of user perception to icons



	Semantics	Familiarity
all_user	Good	Good

(a) Overall

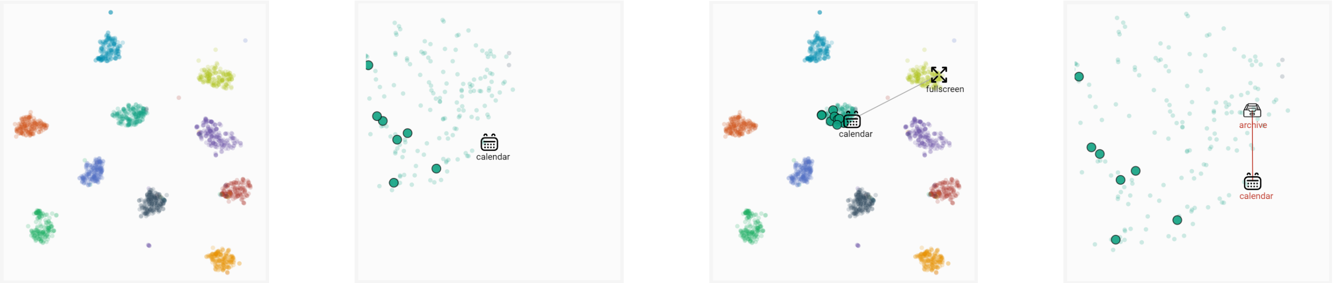
	Semantics	Familiarity
adult	Good	Good
elder	Bad	Very Bad

(b) Age categories: i) adult, ii) elder

	Semantics	Familiarity
tech	Very Good	Very Good
business	Good	Good

(c) Occupation categories: i) technology, ii) business, and iii) others

Figure 4: EvIcon provides predicted perceptual usability feedback. Apart from viewing perception feedback for general people (a), users can inspect the feedback from different demographics categories including (b) age and (c) occupation.



(a) Basic graph with color-coded nodes.

(b) Local view after clicking the "evaluate" button and top-K suggested icons are highlighted.

(c) Global view after clicking the "evaluate" button, icons in the uploaded icon set are connected by links.

(d) Links between icons will be marked in red to notify poor visual distinguishability.

Figure 5: Interactive graph of distinguishability visualization

is an essential steps. However, there is no existing dataset fullfilling the requirements for enabling the perception feedback that we want to provide on EvIcon. In this section, we describe the process of (i) collecting dataset contains unique icons, and (ii) crowdsourced perceptual ratings with semantic distance and familiarity levels.

4.1 Icon Dataset Collection

The first block at the left of Fig. 2 lists the ten functions we selected to include in our dataset. Eight among the selected functions are reported to be frequently used in various applications by the prior work of Liu *et al.* [35]. We included the functions "Filter" and "Archive" since they are relatively new compared to other selected functions. We considered that the validation of such functions can benefit from EvIcon the most because designers have not learned the design pattern for the icons of these functions. The connection between their semantic meaning and visual appearance has not been established. We collected the icons from online resources, including Google Material Icons[5], icon library of IBM design[6], Icon8[7], and The Noun Project[8]. We collected 10,000 icons in our raw icon dataset (1,000 icons for each function). To remove the influence of visual style and color, all icons in the raw icon dataset are in black-and-white with minimalist and flat design. We found that many icons in the raw dataset are very similar to each other (i.e., slight visual difference such as different line thickness). In order to collect users' perceptions for sufficiently diverse icons of each function, we performed the following process to curate the raw icon dataset for each function separately. We divided the raw dataset into ten subsets in which contained the icons of the same function. For each subset, after normalizing the size of icons from different resources into 28×28 pixels, we applied the principal component analysis (PCA) on icons' pixel values after removing the duplicated icons. We set the projection to preserve 90% of the variances to generate the final principal components and utilize them to represent each icon. Next, we performed K-Means clustering [3] on these projected icon representations and set $K = 10$ based on the results of Elbow method (i.e., ten clusters in a subset) [22] . We obtained 200 icons from each subset by randomly sampling 20 icons from each cluster. After repeating the same process to all subset, we acquired the curated dataset (denoted as the dataset in the following sections) with 2,000 icons in which the variety of icons of each function increased compared to the raw dataset.

4.2 Crowdsourced Perceptual Ratings

After finalizing the icon dataset, we used Amazon Mechanical Turk to collect users' perceived semantic distance and familiarity to each icon in the dataset. Before rating the icons, crowdworkers were asked to report the demographic questions, including age, sex, and occupation. Then, they were asked to read the instructions about (i) the definition of semantic distance and familiarity of icons and (ii) the conditions of rejections. In the rating task, crowdworkers were first asked to rate the familiarity of the presented function on a scale from 1 (very unfamiliar) to 5 (very familiar). Next, five icons of this function were displayed and crowd-

[5] https://material.io/resources/icons/
[6] https://www.ibm.com/design/language/elements/icon-library/
[7] https://icons8.com/
[8] https://thenounproject.com/

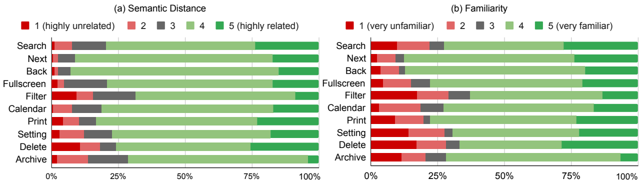


Figure 6: The percentage of the (a) semantic distance and (b) familiarity levels for each function in the crowdsourced labeling dataset.

workers were asked to rate the semantic distance of each icon by *"How semantically related is this icon to the function from 1 (highly unrelated) to 5 (highly related)?"* [38, 20], and the familiarity of each icon by *"How familiar is this icon to you from 1 (very unfamiliar) to 5 (very familiar)?"* [38]. We inserted one repeated icon as a sanity check question among five icons to detect whether the crowdworkers provided contradictory ratings to the same icon. The workers would rate five functions in each assignment (i.e., 25 icons in total), and the order of the functions and icons were randomized. The average completion time was 8 minutes, and we compensated each crowdworker 0.5$ for completing the task.

We recruited $2,698$ workers participating in the crowdsourcing task. Those workers who meet any of the following criteria would be treated as outliers and removed from the rating dataset: (i) contradictory answers for the sanity check questions, (ii) give the same rating to all icons, and (iii) fail to rate all icons. In the final rating dataset, we collected $62,649$ unique ratings for semantic distance and familiarity for 2,000 icons spanning ten functions. We found that rating score of semantic distance and familiarity are strong positive correlated with each other ($r = 0.9$, $p < .001$). This finding echo the results in the prior study [39] in which reported that the semantic closeness between icon and function could be learned by being more familiar with the icon and its connection with the function. We discuss how this issue influence the performance of EvIcon in the Sec. 7. Each icon received 31.3 ratings on average. The mean age of workers is 30.6 (SD $= 7.75$), and 1,565 crowdworkers are reported to be male. The mode in all ratings an icon received was used as the final score of semantic distance and familiarity of the icon.

Fig. 6 illustrates the percentage of rating level (1-5) in each function for semantic distance and familiarity. We can see that the distribution of rating levels varies for each function, and the level "4" occupied the largest proportion among other levels. We explain the solution to balance the number of different levels for training the computational model of perception feedback in Sec. 5.1. Next, we investigated whether the crowdworkers with particular demographics provide different perceptual ratings to icons because prior works observed users of different ages and experiences reacted same icons uniquely [30, 20, 2]. To examine the effects of the demographics we collected (i.e., age, sex, and occupation), we applied ANOVA with each demographics factor as the independent variable and perceptual ratings as dependent variables. Tukey's test was applied in post-hoc tests. We found that female and male crowdworkers provided similar perceptual ratings, yet the adults (20 to 51 years; $n = 2,394$) have higher rating scores of familiarity than the elder (51 and older; $n = 93$) crowdworkers ($F(1, 62647) = 189.6, p < .001$). The crowdworkers from technology-oriented occupation ($n = 978$) provided higher semantic distance ($F(2, 62646) = 350.2, p < .001$) and familiarity ($F(1, 62647) = 319.4, p < .001$) than business-oriented ($n = 890$) and other types of occupations ($n = 634$). Based on the results, we selected "age" and "occupation" as the two types of additional perception predictions mentioned in Sec. 3.3. We used the entire dataset and the subsets divided by crowdworkers' age and occupations to train the separate computational models for data-driven feedback described in the next section.

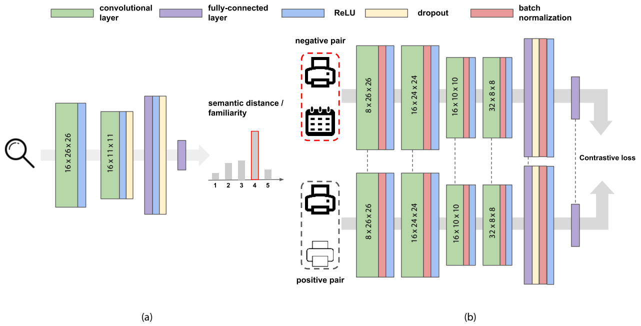


Figure 7: (a) The architecture of our convolutional neural network model for predicting perception level. (b) The architecture of our Siamese neural network for extracting visual features using positive/negative icon pairs. The dotted lines indicate those two layers share weights between each other.

5. COMPUTATIONAL MODELS FOR DATA-DRIVEN FEEDBACK

In EvIcon, we provide two types of feedback for icon set design. First, to support perception prediction of individual icon, we collected icons for ten different functions that are frequently used in various applications. We collected users' perception data using Amazon Mechanical Turk and trained separate classifiers to predict both semantic distance and familiarity of novel icons. Second, we learned an embedding space to illustrate the visual distinguishability using Siamese network. We discuss the implementation details of these two computational models.

5.1 Computational Model for Perception Feedback

Given a novel icon design, we want to build a classifier to predict semantic distance and familiarity. We implemented the classifier using a convolutional neural network (CNN) and we show the architecture of our classifier in Fig. 7(a). Our classifier comprises two convolutional layers with a kernel of 3×3 each, and they contain 16 filters each and are followed by a max-pooling layer. We utilized two fully-connected layers and dropout layers with 0.5 dropout rate after extracting the visual features using convolutional layers. We used ReLU [41] as the activation function. The input of the network is an icon image in 28×28 pixels, and the output of the network is the probabilities of different semantic distance/familiarity levels. The categorical corssentropy loss function was minimized using the ADAM optimizer [23]. We trained a separate network for each function because the visual feature of icons with high semantic distance or familiarity ratings is different across different functions. For example, the visual feature to make a "Search" icon with close semantic distance might not make a "Print" icon with close semantic distance. As Fig. 6 shown, we addressed the issue of imbalance number of each rating level by oversampling the icons of the less frequent levels. The icons of level "1" and "2" (i.e., far semantic distance and less familiar) were oversampled by adding icons from different functions. On the other hand, we oversampled the icons of level "3" (neutral) and "5" by duplicating the same icons.

Results and Findings of Classification Models.

We evaluated each model using the 10-fold cross-validation. In each fold, we randomly selected 90% of the data for training and 10% of the data for validation. We report the mean average precision and recall across 10-fold cross-validation followed by their standard deviation (SD). For *semantic distance*, our models achieved 86.7% (SD=6.4%) for mean precision, and 85.5% (SD=7.3%) for mean recall. For *familiarity*, the mean precision is 81.0% (SD=7.4%), and the mean recall is 79.9% (SD=7.7%), which are slightly lower than the performance of semantic distance prediction. The results of the classification models using the separate dataset containing ratings from crowdworkers of different categories of demographics (i.e., age and occupation) revealed similar performance with the models trained with the entire dataset. Detailed precision and recall numbers for separate functions are shown in Table 1 in the supplemental material. The performance of familiarity prediction is lower than the semantic distance prediction because users' own experience determines their perception of the icon's familiarity; thus, a higher level of individual difference might be involved.

5.2 Computational Model for Visual Distinguishability

As discussed in Sec. 3.4, EvIcon provides distinguishability visualization to help users compare the relative visual distance between icons in the uploaded icon set. To obtain such visualization, we would like to learn the visual similarity from the icons' appearance to understand how an icon visually associated with the icons from the same function and other functions in our dataset. We implemented a Siamese Network inspired by [5, 45] with the contrastive loss function [9] to enforce inter-function separability while preserving intra-function compactness. The input to the contrastive loss function is a pair of icons. If the two icons in the pair represent the same function, the pair is denoted as a positive pair, and if they represent different functions, the pair is treated as a negative pair. The concept of the contrastive loss is to encourage the distance between positive pair icons to be minimized while the negative pair icons are push apart from each other [9].

Figure 7(b) shows the architecture of our Siamese Network. The input X and Y are the normalized pixel values (from 0-255 to 0-1) of two icons with the size of 28×28 pixels from a negative or positive pair. Both inputs were passed into two identical Convolutional Neural Networks (CNNs) with the sequences of layers showed in Fig. 7. The last max-pooling layer is connected to a fully-connected (FC) layer followed by a dropout layer with 0.5 dropout rate and a batch normalization layer. The final output is two 32-dimensional representations of X and Y. Then, we optimize the contrastive loss function, by minimizing the distance between the icons in the positive pair and maximizing the distance between icons in the negative pairs. We chose ReLU [41] as the activation function and use ADAM optimizer [23] to minimize the loss.

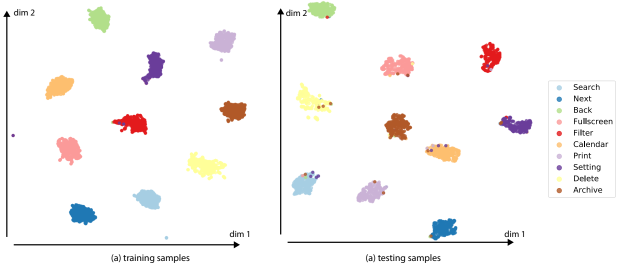


Figure 8: The learned embedding space of the icons for our (a) training and (b) testing samples, where each dot represented an icon, and the color of a dot indicates the function of the icon. Both training and testing samples are grouped nicely according to their functions in the learned embedding space.

Results and Findings of Siamese Network.

We used 70% of our icon dataset as the training set to train the Siamese network with 200 epochs on a 2.7 GHz Intel Core i5 CPU and 8 GB DDR3 RAM. Each epoch contained 50 positive and 50 negative pairs of icons. After the training process, we used the model to compute 32-dimensional features for each icon from the test set (30% of our icon dataset). In Fig. 8, we show the learned embedding space of the icons in our training and testing set, where each dot represented an icon, and the color of a dot indicates the function of the icon. We used Uniform Manifold Approximation and Projection (UMAP) [40] to project the learned 32-dim feature vector to 2-dim. We can observe the dots with the same color are grouped nicely for both training samples (Fig. 8(a)) and testing samples (Fig. 8(a)), where the icons of the same function are closer to each other than the icons with different functions. However, we can also see from the Fig. 8(b) that there are icons of some functions scattered between icons, e.g., the icons of "Archive" (brown dots), "Setting" (light blue dots), and "Print" (light purple dots) are scattered in between the icon groups of other functions. We considered these differences between functions are related to whether the function has an established design pattern of icons or not, and how the icons with different function related visually. By observing the UMAP visualizations of the training and test set, we concluded that we successfully learned an embedding space that matches the functions of different icons; thus is useful for providing visual distinguishability information while evaluating a icon set.

6. EVALUATION WITH UI DESIGNERS

To evaluate how EvIcon can support designers' iteration and validation process, we conducted a study with six UI designers (five females and one male; age ranging from 22 to 34 years old). We grouped three designers with less than five years of professional experience in UI designs as novice designers (P2, P4, P5) and the other three with five or more than five years of experience as professional designers (P1, P3, P6). All of them used Adobe Illustrator and Mac computers in their design practice except P1 and P5 using Windows computers.

In the online recruitment questionnaire, we asked the UI designers to briefly describe their difficulties when designing interface icons. P3 and P6 mentioned that they often found little reference and established metrics to evaluate icon's usability, especially when designing icons for uncommon functions. P4 and P5 noted a gap between the designer's perception and target users' perception of an icon. Sometimes, an icon considered to be suitable by designers turns out to have poor usability for average users. P4 and P5 also mentioned the challenge of designing icons for users of different ages and cultural backgrounds.

6.1 Procedure and Tasks

After introducing the EvIcon and the meaning of two types of feedback, the designers practiced how to use EvIcon for at least ten minutes. We instructed them to complete the given practice tasks (e.g., reporting the perceptual usability of an icon in different groups of users) to ensure they fully understand how to use EvIcon. In the formal sessions, the designers were asked to improve the usability of two icon sets containing one "Archive" icon, one "Filter" icon, and one "Print" icon, which we denoted as the original icons in the following sections. We selected these functions based on their average familiarity level collected via the crowdsourcing study in Sec. 4 (1: very unfamiliar, 5: very familiar; "Archive": 3.8; "Filter": 3.9; "Print": 4.2) to ensure we included established and uncommon functions in the evaluation. Each designer was instructed to improve the usability of one icon set with EvIcon and another icon set without EvIcon. The combination of icon set and the condition of using EvIcon or not were randomly assigned. The order of conditions that designers went through was counterbalanced to avoid the ordering and learning effect. The designers can freely edit icons using Adobe Illustrator [1] and search on the Internet for the information. We present all original and revised icon pair for all designers participated in our evaluation in the supplementary material. In Sec. 6.2.1, we discuss the process and results of the revision sessions with and without using EvIcon.

The designers were given fifteen minutes for each session. We recorded the entire revision process and the final revised icons. After the designers finished both sessions, we conducted a semi-structured interview to understand how they integrate EvIcon when revising icons, their general feedback on EvIcon's usability, and what benefits EvIcon can bring to their current design practice.

To further verify that EvIcon can help designers generate icons with better perceptual usability, we launched the crowdsourcing task on Amazon Mechanical Turk to collect crowdworkers' semantic distance and familiarity ratings of the original icons and the revised icons. Crowdworkers would only rate all the revised icons by the same designers in one assignment to eliminate the influence of individual designers' abilities. We collected fifty unique ratings for each revised and original icon. There were 166 crowdworkers (108 males and 58 females) participated in the crowdsourced evaluation with ages ranging from 19 to 64. Each crowdworker could complete up to two assignments. Each of them completed 1.8 assignment on average.

6.2 Result

6.2.1 *Revised Icons*

In Fig. 9, we show steps of the icon revision process of designing "archive" and "print" with and without EvIcon. Due to the limited functionalities of the implemented editor in EvIcon, some designers modified the icons with Adobe Illustrator [1] and uploaded the edited icon to EvIcon. Therefore, we can only observe the final step of revision process in the session with EvIconfor the examples shown in Fig. 9(b).

In Fig. 9, the "Semantics" (abbreviation for semantic distance) and "Familiarity" feedback predicted by EvIcon were provided along with the revision process under the icons of each design steps. The perception feedback of icons designed without EvIcon were generated afterward for comparison. Crowdsourced evaluation results collected on Amazon Mechenical Turk (AMT) of the revised icons are shown right next to the finalized icon in each revision process.

Because the icons of the same function in both sets have similar perceptual usability according to the results of the crowdsourced evaluation, we investigate the influence of using EvIcon by comparing the icons representing the same function from each set revised by the same designers to eliminate the bias of individual designers' ability. We highlighted the evaluation outcomes of perceptual usability that outperformed the other icon designed by the same designer in orange for better readability. In Fig. 9, we can see that the revision steps with EvIcon achieved better predicted semantic distance and familiarity in the example of "Archive" icons by P6 and "Print" icons by P5. Moreover, the crowdsourced evaluation outcomes shows that most of the icons revised with EvIcon outperformed the ones revised without EvIcon on both "Semantics" and "Familiarity" as demonstrated in Fig. 9.

6.2.2 *Revised Icons For Specific Demographics*

We also want to investigate whether providing perception feedback can successfully help designers revise icons tailor to the target audience with particular demographics. We selected "the elders" group as the demonstration. Fig. 10 shows the examples of revised icons with the predicted semantic distance and familiarity levels by the classifier trained using elders' perceptual ratings. The AMT ratings shown in Fig. 10 were given by the crowdworkers over fifty years old in our crowdsourced evaluation on revised icons. We can see that the icons with EvIcon received higher levels of semantic distance and familiarity predicted by models of elders than those without EvIcon. Moreover, the revised icons optimized for elders also received better perceptual usability according to the AMT ratings (age $>$ 50 yrs). These examples show that EvIcon helps designers to generate elder-friendly icons, and these icons indeed received higher ratings from the older crowdworkers.

6.2.3 *Post-study Interview*

In the post-study interviews, all six designers gave positive attitudes towards EvIcon. The designers mentioned that when revising icons with EvIcon, they got the idea of how to revise an icon to meet public understandings more easily

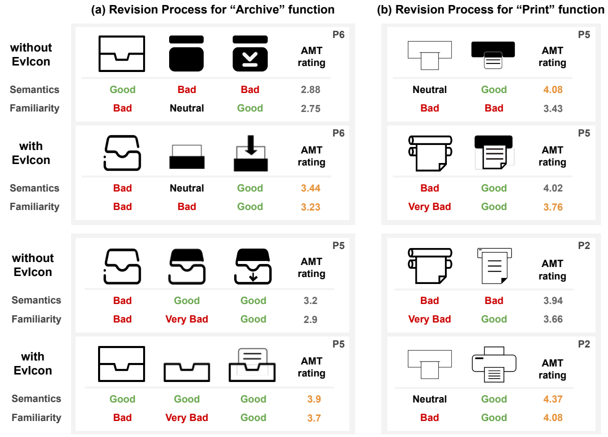


Figure 9: Icon revision process by designers with and without EvIcon. (a) presents the revision processes for “archive” function of two designers, and (b) presents the revision processes for “print” function of two designers. Eight groups of revision process with prediction of perception feedback and the crowdsourced evaluation results (“AMT rating”) of the finalized icons are presented. Both process of the identical function revised by the same designer with and without EvIcon are shown for comparison.

by checking the perception feedback constantly. They found the perception feedback convincing as it was generated based on data labeled by over two thousand crowdworkers:

- *“EvIcon keeps me on the right track and ensures that my design can be understood by others while I modify the icon design based on my creativity. ”(P3)*
- *“The good or bad rating provided by the system is promising and helpful in designing high-usability interface icons, compared to designing the icons on my own.”(P5)*

Some designers were amazed by the perception feedback for specific demographics since they have experience in struggling to design interface targeting a specific category of users while having limited knowledge or access to the users:

- *“The feedback from specific demographic is very useful. I can adjust the icons according to the feedback from my target user’s category provided by the system. This tool definitely helps this.”(P4)*
- *“I am touched to see how this tool supports elders’ feedback! Since icons play an important role in interface design, while there are not much information about which icons are friendly or recognizable to elders. ”(P6)*

Designers also found the distinguishability visualization panel helpful, both P2 and P6 said they would check the related distance between the uploaded icon and the icons in suggestion panel to see how they could improve their design. P2, P3, P5, and P6 mentioned they could derive some graphical design feature from the icon suggestion panel that can be added to their own designs:

- *“It is interesting that the system provides designs from other designers based on current target function.”(P3)*
- *“I can see those good icons in the suggestion panel, and think about how to start my design based on the recommendations. It will help save my time to grasp users’ thought at the beginning of the design flow.”(P5)*

Designers also talked about possible benefits EvIcon could bring if it is applied in their current workflow. P5 said it would save lots of time to notice the perception gap between designers, engineers, and average users earlier with EvIcon, instead of finding out in usability testing after several design iterations and discussion. As designers, participants usually care a lot about aesthetic while designing icons, EvIcon could also provide assistances to balance between aesthetic and usability.

- *“It was nice that I could see the perception differences between public users and my personal thoughts and styles.”(P2)*
- *“Designers often want to design an aesthetic and unique icon, but sometimes they went too far that the icon becomes unrecognizable to users. With EvIcon, it would be easier to take both aesthetic and usability into consideration at the same time.”(P3)*
- *“Designers often add more styling details in the later phase of the iteration and worsen the icons’ distinguishability. With EvIcon , we can check the perception feedback in each iteration to ensure the quality of our designed icons.”(P4)*

The designers also mentioned that the perception feedback could improve the communication with their colleagues or clients if EvIcon is included in their design process.

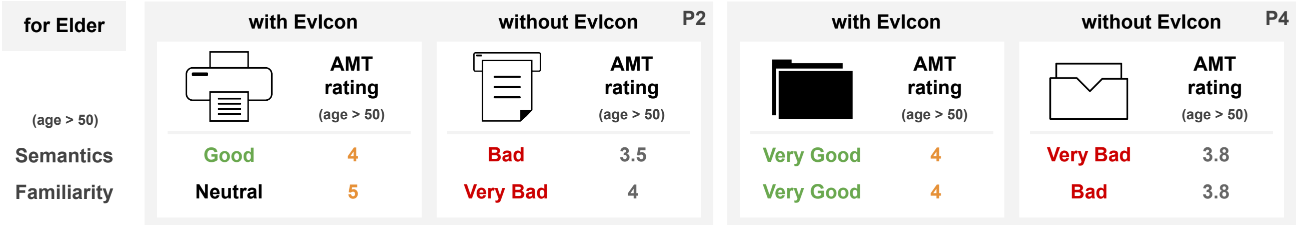
- *“I could convince the clients that my design is good with EvIcon.”(P3)*
- *“The results from EvIcon would be a promising report to defend our design against clients.”(P4)*

The designers confirmed that EvIcon could be generally useful and mentioned possibilities of how EvIcon can provide assistances in different design phases. Moreover, they are willing to use EvIcon in their design process if it becomes a mature product in the future.

6.3 Statistical Results of Crowdsourced Evaluation

First, we want to verify if revised icons have better usability than the original ones. We conducted ANCOVA in which the performances of semantic distance and familiarity were treated as the dependent variables, and the UI designers’ ID and the icon set number were used as the control variables. We utilized Tukey’s method for post-hoc tests. The result shows that the revised icons gained significantly higher level of semantic distance ($F(1, 3696) = 79.5, p < .001$) and familiarity ($F(1, 3696) = 81.2, p < .001$) than the original icons.

After confirming that revised icons’ usability improved compared to the original icons, we conducted another ANCOVA to verify whether the revised icons using EvIcon achieve better performances in semantic distance and familiarity than the revised icons without using EvIcon (i.e., only using Adobe Illustrator to revise icons). Moreover, we want to investigate if EvIcon provides different levels of support for different functions and designers with different professional



for Elder	with EvIcon	AMT rating	without EvIcon	AMT rating	with EvIcon	AMT rating	without EvIcon	AMT rating
(age > 50)		(age > 50)		(age > 50)		(age > 50)		(age > 50)
Semantics	Good	4	Bad	3.5	Very Good	4	Very Bad	3.8
Familiarity	Neutral	5	Very Bad	4	Very Good	4	Bad	3.8

Figure 10: Revised icons by designers with EvIcon show better usability ("AMT rating") for elder crowdworkers than those revised without EvIcon. The results aligned with the perception feedback for elders provided by EvIcon (shown below the icon).

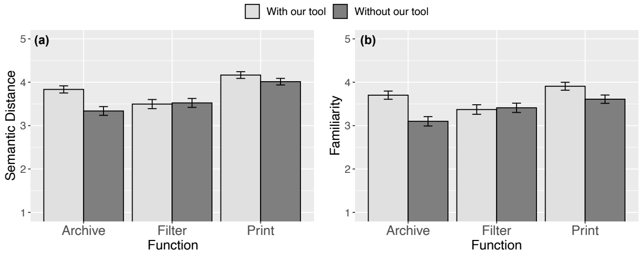


Figure 11: The results of evaluating the novice designers' revised icons on Amazon Mechanical Turk. (a) The rating level of semantic distance (1: highly unrelated; 5:highly related). (b) The rating level of familiarity (1: very unfamiliar; 5:very familiar). The error bars represent the standard error.

levels (i.e., novice and professional). We used semantic distance and familiarity as the dependent variables and the same control variables in the analysis. Tukey's method was adopted for post-hoc tests.

We found that using EvIcon has the main effect on semantic distance ($F(1, 1841) = 7.5, p < .001$) and familiarity ($F(1, 1841) = 6.6, p < .01$). Moreover, the effect of EvIcon has interactions with functions and professional levels of designers. We found that the positive effect of using EvIcon on icon's usability only exists for the icons revised by novice designers. The icons revised by professional designers with or without EvIcon have similar semantic distance and familiarity performances. From Fig. 11, we can see that for the icons revised by the novice designers, the "Archive" icons revised by EvIcon received higher level of semantic distance (Fig. 11(a), $p < .01$) and familiarity (Fig. 11(b), $p < .01$) than the "Archive" icons without using EvIcon. The "Prin" icons revised by EvIcon also gained higher level of semantic distance and familiarity than "Print" icons revised without EvIcon, but we did not find the significant difference in this comparison EvIcon ($p > .05$). For the "Filter" function, the icons revised with and without the support of EvIcon obtained similar level of semantic distance and familiarity.

The above results indicate that novice designers significantly benefit from using EvIcon in improving the usability of icons. Because novice designers are less experienced in icon design, the feedback provided in EvIcon helps significantly improve the original icons. On the other hand, the professional designers have relatively sufficient abilities to improve icons' usability without EvIcon's supports, so there is no significant difference in usability between the icons revised by EvIcon and icons without using EvIcon. Although we did not observe the significant benefits of EvIcon on professional designers in this crowdsourced quantitative evaluation, we considered that the aids of EvIcon for professional designers mainly reflected in their mental models and design strategies, as we mentioned in the qualitative results of the post-study interview above.

Among the three functions selected in the evaluation, "Archive" was rated as the most unfamiliar function. This result also matches our observation that we found diverse "Archive" icon designs while collecting the icon dataset, since most of the designers and users have not formed the common and specific visual metaphor for the unfamiliar function "Archive." Hence, EvIcon's feedback helps novice designers navigate the vast variations of "Archive" icons and find the best way to increase icons' perceptual usability.

7. DISCUSSION AND LIMITATIONS

7.1 Interactions between Designers and Authoring Assistive Tool

Given the performance of our classification models, EvIcon provide adequate and meaningful predictions for designers to derive informative messages of how to improve icons, which echoes the findings of prior research [44]. It is notable that we only observed the statistically significant supports of EvIcon for the "Archive" icons refined by novice designers in the crowdsourced evaluation. We consider that designers with different professional levels might have different strategies of adopting such assistive authoring tools in their design practice. This issue can further extend to the Human-AI interaction [32] research direction, which focus on understanding how human perceive and react to information and suggestions provided by intelligent systems. Building trustworthy and accountable collaboration between users and AI is one of primary goals for Human-AI interaction. We want to investigate this issue and provide different design tools tailor to designers of various professional levels in the future.

7.2 Extending Dataset and Using Advanced Computation Models

As a proof-of-concept, the current version of EvIcon only supports the instant evaluation of semantic distance and familiarity to the icons representing the ten functions we selected in the icon dataset. In the future, we will extend the categories of functions, the number of icons, and the perceptual usability derived from human ratings. On the other hand, the current visual distinguishability feedback is derived from the visual appearance of icon images and the designer's perceptions (the function label of an icon was given by the designers). As the gaps between designers' and users' perceptual usability were reported in our post-study interview, we want to integrate end users' perception into the embedding space of visual distinguishability between icons in the future. For example, we can utilize the positive and negative icon pairs grouped by users or crowdworkers to train Siamese neural network.

Currently, we did not fully explore the benefits of EvIcon on improving the quality of icon set in the evaluation with UI designers due to our design of task. However, we learn from the UI designers that the validation of the icon set is more style-oriented and includes more higher-level factors (e.g., aesthetic, layout, design of UI, and the brand identity). Although our visual distinguishability feedback can provide a primary evaluation to avoid confusing icons in an icon set, more advanced computational models and relevant datasets are needed to provide more meaningful feedback to evaluate icon sets. For example, we can combine the tappability models [51] or the automatic mobile UI labeling [35] in our framework to gain the information of users' perception of different UI layouts and elements. We also want to integrate the deep learning models that support style similarity [10] and style transfer [15] in our future works. Lastly, due to the importance of color for visual design and perception [33, 43], we want to include colored icons in our dataset and investigate the influence of different colors on icons' perceptual usability in the future.

7.3 Supporting validations for General Use of Icons

Although the goal of the proposed framework is to support designers validate and revise icons for user interface design, icons can be used in various scenarios such as presentation slides and infographics. The icons used in these scenarios may need to optimize different users' perceptions other than semantic distance and familiarity. For example, infographic icons may require better abilities to convey information rather than better familiarity with viewers. Therefore, we want to extend the usage scenario and target audience of EvIcon to support icon improvement and selection for more general purposes.

8. CONCLUSION

We propose EvIcon, an interactive design tool with two types of feedback to facilitate effective icon set design iteration and validation. The core of EvIcon is a framework that provides comprehensive and objective feedback based on crowdsourced user perceptual ratings and deep learning models. We demonstrated the effectiveness of EvIcon by conducting a user study with six designers. From the post-study interview and the additional crowdsourced evaluation results, we conclude that EvIcon can assist designers to improve icons' semantic distance and familiarity performances. Our framework is not limited to model the perception criteria used in this paper and can be extended to more criteria. We believe that the proposed framework and EvIcon will significantly ease the usability testing process for icon set design and open more opportunities for novel data-driven feedback designs.

9. ADDITIONAL AUTHORS

10. REFERENCES

[1] Adobe. Adobe illustrator 2021, 2021.

[2] A. X. Ali, E. Mcaweeney, and J. O. Wobbrock. Anachronism by design: Understanding young adults' perceptions of computer iconography. *International Journal of Human-Computer Studies*, page 102599, 2021.

[3] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

[4] R. W. Bailey, R. W. Allan, and P. Raiello. Usability testing vs. heuristic evaluation: A head-to-head comparison. In *Proceedings of the human factors society annual meeting*, volume 36, pages 409–413. SAGE Publications Sage CA: Los Angeles, CA, 1992.

[5] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[6] 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 *Proceedings of the 30th Annual ACM symposium on user interface software and technology*, pages 57–69, 2017.

[7] H.-T. Chen, T. Grossman, L.-Y. Wei, R. M. Schmidt, B. Hartmann, G. Fitzmaurice, and M. Agrawala. History assisted view authoring for 3d models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2027–2036, New York, NY, USA, 2014. ACM.

[8] F.-Y. Cherng, W.-C. Lin, J.-T. King, and Y.-C. Lee. An eeg-based approach for evaluating graphic icons from the perspective of semantic distance. In *Proceedings of the 2016 chi conference on human factors in computing systems*, pages 4378–4389. ACM, 2016.

[9] S. Chopra, R. Hadsell, Y. LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546, 2005.

[10] J. Collomosse, T. Bui, M. J. Wilber, C. Fang, and H. Jin. Sketching with style: Visual search with sketches and aesthetic context. In *ICCV*, pages 2679–2687, 2017.

[11] 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 ACM Symposium on User Interface Software and Technology*, pages 845–854, 2017.

[12] 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. UIST '17, page 845–854, New York, NY, USA, 2017. Association for Computing Machinery.

[13] B. Deka, Z. Huang, C. Franzen, J. Nichols, Y. Li, and R. Kumar. Zipt: Zero-integration performance testing of mobile app designs. In *Proceedings of the 30th*

Annual ACM Symposium on User Interface Software and Technology, pages 727–736, 2017.

[14] B. Deka, Z. Huang, C. Franzen, J. Nichols, Y. Li, and R. Kumar. Zipt: Zero-integration performance testing of mobile app designs. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, page 727–736, New York, NY, USA, 2017. Association for Computing Machinery.

[15] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[16] Y. Gingold, A. Shamir, and D. Cohen-Or. Micro perceptual human computation. *ACM Transactions on Graphics (TOG)*, 31(5):119:1–119:12, Aug. 2012.

[17] D. Gittins. Icon-based human-computer interaction. *International Journal of Man-Machine Studies*, 24(6):519–543, 1986.

[18] W. Horton. Designing icons and visual symbols. In *Conference Companion on Human Factors in Computing Systems*, CHI '96, page 371–372, New York, NY, USA, 1996. Association for Computing Machinery.

[19] W. K. Horton. *The ICON Book: Visual Symbols for Computer Systems and Documentation*. John Wiley & Sons, Inc., USA, 1994.

[20] S. J. Isherwood, S. J. McDougall, and M. B. Curry. Icon identification in context: The changing role of icon characteristics with user experience. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(3):465–476, 2007.

[21] N. A. Kamarulzaman, N. Fabil, Z. M. Zaki, and R. Ismail. Comparative study of icon design for mobile application. In *Journal of Physics: Conference Series*, volume 1551, page 012007. IOP Publishing, 2020.

[22] D. J. Ketchen and C. L. Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458, 1996.

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] P. Kolhoff, J. Preuß, and J. Loviscach. Content-based icons for music files. *Computers & Graphics*, 32(5):550–560, 2008.

[25] S. Komarov, K. Reinecke, and K. Z. Gajos. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 207–216, 2013.

[26] Y. Koyama, I. Sato, D. Sakamoto, and T. Igarashi. Sequential line search for efficient visual design optimization by crowds. 36(4), July 2017.

[27] M. Lagunas, E. Garces, and D. Gutierrez. Learning icons appearance similarity. *Multimedia Tools and Applications*, pages 1–19, 2018.

[28] L. F. Laursen, Y. Koyama, H.-T. Chen, E. Garces, D. Gutierrez, R. Harper, and T. Igarashi. Icon set selection via human computation. 2016.

[29] Y. J. Lee, C. L. Zitnick, and M. F. Cohen. Shadowdraw: Real-time user guidance for freehand drawing. *ACM Trans. Graph.*, 30(4):27:1–27:10, July 2011.

[30] R. Leung, J. McGrenere, and P. Graf. Age-related differences in the initial usability of mobile device icons. *Behaviour & Information Technology*, 30(5):629–642, 2011.

[31] J. P. Lewis, R. Rosenholtz, N. Fong, and U. Neumann. Visualids: Automatic distinctive icons for desktop interfaces. *ACM Trans. Graph.*, 23(3):416–423, Aug. 2004.

[32] Q. V. Liao, D. Gruen, and S. Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[33] Y.-P. Lim and P. C. Woods. Experimental color in computer icons. In *Visual Information Communication*, pages 149–158. Springer, 2010.

[34] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. Turkit: Human computation algorithms on mechanical turk. UIST '10, page 57–66, New York, NY, USA, 2010. Association for Computing Machinery.

[35] 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*, pages 569–579. ACM, 2018.

[36] Y. Liu, A. Agarwala, J. Lu, and S. Rusinkiewicz. Data-driven iconification. In *International Symposium on Non-Photorealistic Animation and Rendering (NPAR)*, May 2016.

[37] S. McDougall and S. Isherwood. What's in a name? the role of graphics, functions, and their interrelationships in icon identification. *Behavior research methods*, 41(2):325–336, 2009.

[38] S. J. Mcdougall, M. B. Curry, and O. de Bruijn. Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols. *Behavior Research Methods, Instruments, & Computers*, 31(3):487–519, 1999.

[39] S. J. McDougall, M. B. Curry, and O. de Bruijn. The effects of visual information on users' mental models: An evaluation of pathfinder analysis as a measure of icon usability. *International Journal of Cognitive Ergonomics*, 5(1):59–84, 2001.

[40] L. McInnes, J. Healy, N. Saul, and L. Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[41] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress.

[42] M. Peng, J. Xing, and L.-Y. Wei. Autocomplete 3d sculpting. *ACM Trans. Graph.*, 37(4):132:1–132:15, July 2018.

[43] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, and K. Z. Gajos. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and

colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2049–2058. ACM, 2013.

[44] R. Rosenholtz, A. Dorai, and R. Freeman. Do predictions of visual perception aid design? *ACM Transactions on Applied Perception (TAP)*, 8(2):1–20, 2011.

[45] P. Sangkloy, N. Burnell, C. Ham, and J. Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119, 2016.

[46] V. Setlur, C. Albrecht-Buehler, A. A. Gooch, S. Rossoff, and B. Gooch. Semanticons: Visual metaphors as file icons. In *Computer Graphics Forum*, volume 24, pages 647–656, 2005.

[47] V. Setlur and J. D. Mackinlay. Automatic generation of semantic icon encodings for visualizations. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 541–550, 2014.

[48] I.-C. Shen and B.-Y. Chen. Clipgen: A deep generative model for clipartvectorization and synthesis. *arXiv preprint*, 2021.

[49] I.-C. Shen, K.-H. Liu, L.-W. Su, Y.-T. Wu, and B.-Y. Chen. Clipflip : Multi-view clipart design. *Computer Graphics Forum*, 2021.

[50] Sketch. Sketch, 2021.

[51] A. Swearngin and Y. Li. Modeling mobile interface tappability using crowdsourcing and deep learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[52] N. Umetani, T. Igarashi, and N. J. Mitra. Guided exploration of physically valid shapes for furniture design. *ACM Trans. Graph.*, 31(4):86–1, 2012.

[53] N. Umetani, Y. Koyama, R. Schmidt, and T. Igarashi. Pteromys: Interactive design and optimization of free-formed free-flight model airplanes. *ACM Trans. Graph.*, 33(4), July 2014.

[54] D. Warnock, M. McGee-Lennon, and S. Brewster. Multiple notification modalities and older users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1091–1094, 2013.

[55] J. Xing, H.-T. Chen, and L.-Y. Wei. Autocomplete painting repetitions. *ACM Trans. Graph.*, 33(6):172:1–172:11, Nov. 2014.

[56] N. Zhao, N. W. Kim, L. M. Herman, H. Pfister, R. W. Lau, J. Echevarria, and Z. Bylinskii. Iconate: Automatic compound icon generation and ideation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

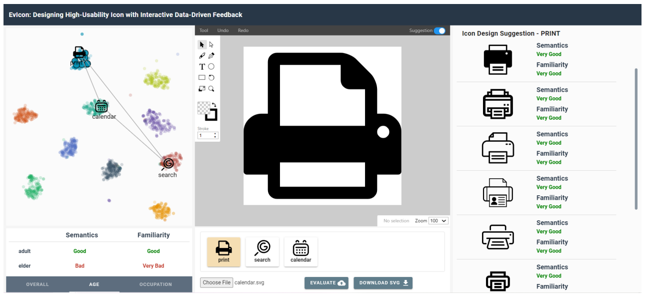
APPENDIX

.1 EvIcon interface

In Fig. 12, we show the full user interface of EvIcon. Please refer to the paper for the introduction of the interface.

.2 Computational Model

In Table 1, we reported the precision, recall, and F1 performance for separate functions using all crowdworkers' usability rating.



EvIcon: Designing High-Usability Icon with Interactive Data-Driven Feedback

	Semantics	Familiarity
adult	Good	Good
elder	Bad	Very Bad

Icon Design Suggestion - PRINT

Semantics Very Good Familiarity Very Good

Semantics Very Good Familiarity Very Good

Semantics Very Good Familiarity Very Good

Semantics Very Good Familiarity Very Good

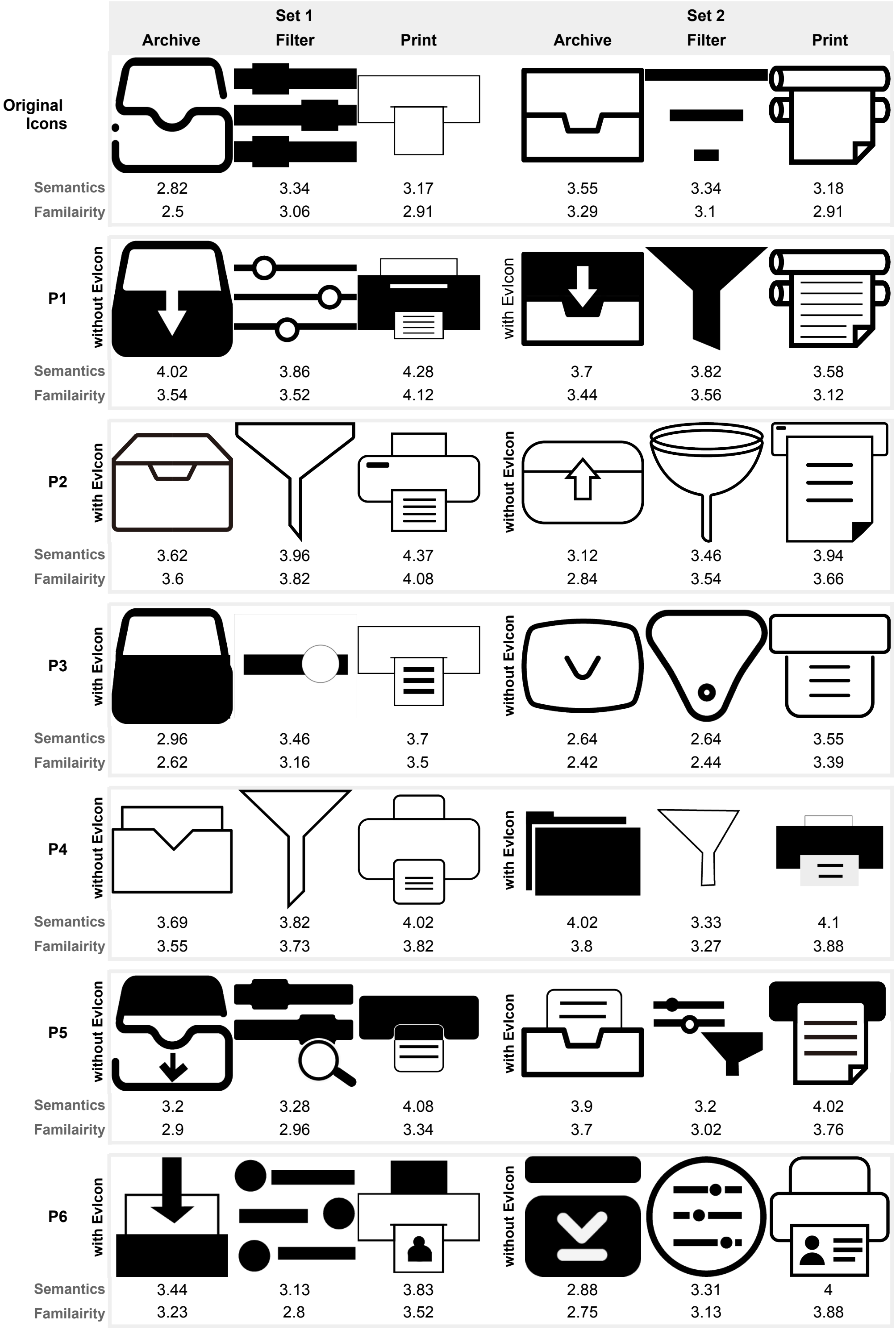
Semantics Very Good Familiarity Very Good

Semantics Very Good Familiarity

Choose File calendar.svg

Figure 12: EvIcon user interface.

.3 The original Icon and Revised Icons in Evaluation



			Set 1			Set 2	
		Archive	Filter	Print		Archive	Filter	Print
Original Icons	Semantics	2.82	3.34	3.17		3.55	3.34	3.18
	Familairity	2.5	3.06	2.91		3.29	3.1	2.91
P1	without EvIcon				with EvIcon			
	Semantics	4.02	3.86	4.28		3.7	3.82	3.58
	Familairity	3.54	3.52	4.12		3.44	3.56	3.12
P2	with EvIcon				without EvIcon			
	Semantics	3.62	3.96	4.37		3.12	3.46	3.94
	Familairity	3.6	3.82	4.08		2.84	3.54	3.66
P3	with EvIcon				without EvIcon			
	Semantics	2.96	3.46	3.7		2.64	2.64	3.55
	Familairity	2.62	3.16	3.5		2.42	2.44	3.39
P4	without EvIcon				with EvIcon			
	Semantics	3.69	3.82	4.02		4.02	3.33	4.1
	Familairity	3.55	3.73	3.82		3.8	3.27	3.88
P5	without EvIcon				with EvIcon			
	Semantics	3.2	3.28	4.08		3.9	3.2	4.02
	Familairity	2.9	2.96	3.34		3.7	3.02	3.76
P6	with EvIcon				without EvIcon			
	Semantics	3.44	3.13	3.83		2.88	3.31	4
	Familairity	3.23	2.8	3.52		2.75	3.13	3.88

Figure 13: The original icons used in our evaluation and the revised icons by six UI designers. The semantics (semantic distance) and familiarity are the mean score provided by the crowdworkers in the crowdsourced evaluation.

Table 1: CNN classification results for each function. The last row computed the overall value of the metrics by computing the mean and standard deviation of each function's recall, precision, and F1 score, respectively.

	Semantic Distance		
Function	**Precision**	**Recall**	**F1 Score**
Search	86.5% (7.1%)	84.7% (7.7%)	85.6% (7.3%)
Next	95.4% (2.4%)	95.2% (2.2%)	95.3% (2.3%)
Back	94.7% (4.7%)	94.7% (4.8%)	94.7% (4.8%)
Fullscreen	86.5% (7.9%)	85.3% (8.5%)	85.4% (8.5%)
Filter	86.2% (9.4%)	85.2% (9.7%)	85.3% (9.8%)
Calendar	87.8% (5.4%)	87.4% (5.6%)	87.1% (6.1%)
Print	94.1% (6.8%)	93.8% (6.9%)	93.7% (7.1%)
Setting	79.0% (8.4%)	76.6% (8.6%)	76.4% (9.0%)
Delete	74.9% (10.7%)	72.5% (11.5%)	71.8% (12.6%)
Archive	82.2% (5.6%)	79.8% (6.1%)	79.5% (5.7%)
Overall	86.7% (6.4%)	85.5% (7.3%)	85.5% (7.5%)
	Familiarity		
Function	**Precision**	**Recall**	**F1 Score**
Search	78.1% (9.4%)	77.4% (9.9%)	76.6% (9.9%)
Next	90.8% (5.2%)	90.5% (5.3%)	90.5% (5.2%)
Back	89.7% (5.3%)	88.9% (5.1%)	88.8% (5.1%)
Fullscreen	82.8% (8.1%)	81.1% (10.1%)	80.4% (10.0%)
Filter	83.2% (4.4%)	82.5% (4.3%)	82.8% (4.0%)
Calendar	81.9% (8.9%)	80.2% (9.5%)	80.1% (9.7%)
Print	87.6% (6.3%)	86.8% (6.6%)	86.7% (6.5%)
Setting	71.1% (9.3%)	70.2% (8.5%)	69.8% (9.3%)
Delete	74.5% (12.3%)	72.5% (12.3%)	71.2% (12.8%)
Archive	70.0% (14.6%)	68.4% (14.0%)	67.1% (14.8%)
Overall	81.0% (7.4%)	79.9% (7.7%)	79.4% (8.1%)