Volume 0 (1981), Number 0 pp. 1–13 COMPUTER GRAPHICS *forum*

EvIcon: Designing High-Usability Icon with Human-in-the-loop Exploration and IconCLIP

I-Chao Shen[1] Fu-Yin Cherng[2†] Takeo Igarashi[3] Wen-Chieh Lin[4] Bing-Yu Chen[5]

[1] ichaoshen@g.ecc.u-tokyo.ac.jp , The University of Tokyo, Japan
[2] fuyincherng@cs.ccu.edu.tw, National Chung Cheng University, Taiwan
[3] takeo@acm.org, The University of Tokyo, Japan
[4] wclin@cs.nycu.edu.tw, National Yang Ming Chiao Tung University, Taiwan
[5] robin@ntu.edu.tw, National Taiwan University, Taiwan

Abstract

Interface icons are prevalent in various digital applications. Due to limited time and budgets, many designers rely on informal evaluation, which often results in poor usability icons. In this paper, we propose a unique human-in-the-loop framework that allows our target users, i.e., novice and professional UI designers, to improve the usability of interface icons efficiently. We formulate several usability criteria into a perceptual usability function and enable users to iteratively revise an icon set with an interactive design tool, EvIcon. We take a large-scale pre-trained joint image-text embedding (CLIP) and fine-tune it to embed icon visuals with icon tags in the same embedding space (IconCLIP). During the revision process, our design tool provides two types of instant perceptual usability feedback. First, we provide perceptual usability feedback modeled by deep learning models trained on IconCLIP embeddings and crowdsourced perceptual ratings. Second, we use the embedding space of IconCLIP to assist users in improving icons' visual distinguishability among icons within the user-prepared icon set. To provide the perceptual prediction, we compiled IconCEPT10K, the first large-scale dataset of perceptual usability ratings over $10,000$ *interface icons, by conducting a crowdsourcing study. We demonstrated that our framework could benefit UI designers' interface icon revision process with a wide range of professional experience. Moreover, the interface icons designed using our framework achieved better semantic distance and familiarity, verified by an additional online user study.*

Keywords: vector graphics, icon, CLIP, usability, human-in-the-loop

CCS Concepts

• Human-centered computing → Interactive systems and tools; • Computing methodologies → Computer graphics;

1. Introduction

Amid the ubiquity of digital technologies, including computers, intelligent appliances, and wearable devices, interface icons play an increasingly important role in representing various functions with benefits including improving interface scannability (i.e., the ease of reading and understanding the content of the interface), saving space on small screens, and conveying information universally [SABAG*05, SM14]. The usability of icons is determined by several characteristics, including visual complexity, style, familiarity, etc. [MCdB01, IMC07]. For instance, previous research has demonstrated that users are able to recognize familiar icons more quickly compared to unfamiliar ones [IMC07]. Furthermore, the visual design of the icons significantly impacts how users perceive the usability of both the interface and the overall system [KK95, HM10, SJ16]. While existing design guidelines (like Google's Material Design) offer invaluable insights into the visual aspects of icon design, the step of gathering users' perceptual feedback on the icons is indispensable for accurately assessing icon usability [BAR92].

Yet, conducting formal usability tests (e.g., inviting real users to perform usability tasks) can be time-consuming and requires extra effort [RDF11, DHF*17b, SL19], which could significantly lengthen the iterative process of interface icon design. Moreover, when evaluating icons designed for specific users (e.g., elders or users with lower computer literacy), conducting adequate usability testing is even more laborious.

Zhao *et al.* [ZKH*20] reported that designers often consult other UI designers' feedback on icons. These informal evaluations often failed to provide comprehensive and objective information about how target users would perceive and use the icons [BAR92, RDF11], thus leading to low usability icons. As shown in Figure 1, even for icons of standard tags (e.g., "Search" and "Calendar"),

† Corresponding author

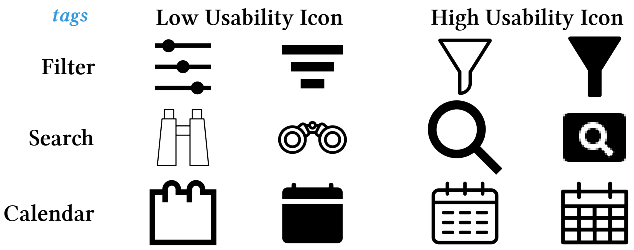


Figure 1: *Example icons in IconCEPT10K with usability rated by crowdworkers. Icons with low usability have three common shortcomings. First, these icons make users misunderstand their tag with others (e.g., Filter). Second, they use unconventional metaphors to transmit the meanings of a concept (e.g., Search). Last, these icons omit the critical features, so users fail to recognize the target concept (e.g., Calendar).*

accidentally omitting the critical visual features when adjusting icons' style could lead to poor usability [Lin94, CUG20]. These examples further illustrate the importance of getting objective and instant feedback from users. These findings underscore the need for an objective and comprehensive usability testing approach that is cost- and time-efficient. Although several automatic graphical icon synthesis methods have been proposed [LRFN04, KPL08], involving humans (i.e., "human-in-the-loop") in the design process has several advantages (e.g., solving computationally complex problems [Hol16], building users' trustworthiness to interactive systems [LGM20]).

Hence, instead of proposing an automatic icon synthesis method, we propose EvIcon, an interactive framework to reduce the workload of performing usability tests for a user-prepared interface icon set. EvIcon comprises two main parts: (i) a novel human-in-the-loop formulation of icon and icon set design and (ii) an interactive tool with instant perceptual usability feedback. Our main idea is to formulate the common icon usability criteria into perceptual usability functions. Among all icon-related features, we select *semantic distance* and *familiarity* as the usability criteria since they are the most critical indications of icons' effectiveness at conveying information [MCdB99, SABAG*05, WMLB13, SM14, CLKL16] and are commonly used by professional artists. As defined in prior icon design literature, *Semantic distance* stands for the perceived degree of closeness between an icon and the tag it represents [MCdB99, SABAG*05, SKJ17] and *familiarity* referred to users' experiences and perceived frequency of encountering specific icons [MCdB99, SABAG*05]. In Figure 2(a), we show examples of icons with different semantic distances and familiarity levels. Moreover, prior research has found that using icons with close semantic distance and high familiarity can significantly increase both user's behavioral performance on interfaces and perceived usability [MCdB99, SABAG*05, WMLB13, SM14, CLKL16, SJ16]. Hence, due to the importance of these two indications for icon's usability [SABAG*05, MI09, SM14], we focus on providing icon designers with semantic distance and familiarity predictions on icon designs in this paper. Moreover, as an icon is usually designed and displayed within an icon set [Kur00], we also use *visual distinguishability* (i.e., ability to be easily recognized and differentiated from one another at a glance) as a critical usability criterion for designing an icon set [Kur00, LRFN04, SABAG*05, MS95]. The goal is to prevent users from confusing icons of different tags.

To reach the goal of this study mentioned above, we gathered the first large-scale dataset of single-colored interface icon usability ratings coined as *IconCEPT10K*. The reason we focus on the single-colored icons is that single-colored icons have been recommended by popular online resources (e.g., Font Awesome and Noun Project) and major software providers (e.g., Google and Apple) due to their scalability in various screen sizes and applications as the prevalence of flat UI design [SRS18, LC20]. Also, icons are usually designed in single-colored in the first place and then edited their color later, tailoring to the configuration of display devices [GSF01, ZKH*20]. Moreover, prior works found that icons' coloring is more critical to icons' visual attractiveness than effectiveness in the conveyance of meaning [Hsi17, CUG20, SZL*21]. Accordingly, we consider devising icons in single-colored is common in the design process. Hence, we focus on the single-colored icons in the present study.

Our perceptual usability function comprises two components. First, we took a large-scale pre-trained joint image-text embedding (CLIP [RKH*21]) and fine-tuned it to embed icon visuals with icon tags in the same embedding space (IconCLIP). Second, we collected usability ratings for a curated icon dataset of 50 base tags. We expanded the base tags by using the tags associated with each icon; thus, our usability prediction model can recognize unseen tags and is scalable for future use. After building the perceptual usability function, we present an interactive user interface with two types of instant feedback (as shown in Figure 2) to support refining icons' usability efficiently. Users can iteratively revise the initial icon in the prepared icon set and query for predicted usability results. The first feedback is the predicted perceptual usability of the revised icon (Figure 2(a)). The second feedback is the icon's visual distinguishability to other icons in (i) the user-prepared icon set and (ii) our icon dataset. This feedback is realized by providing an interactive two-dimensional visualization of the IconCLIP embedding (see Figure 2(b)).

To understand the benefits of EvIcon for designers, we conducted a user study with six UI designers and asked them to revise icon sets with and without using EvIcon. We further conducted an online user study on the revised icons to verify whether EvIcon can assist UI designers to improve icons' usability. The result shows that EvIcon can assist UI designers with a wide range of professional experiences to improve the usability of their icon designs. The major contributions and novelties of this paper include:

- We propose a novel human-in-the-loop formulation, EvIcon, for refining the usability of an icon set, while previous works focus on providing supports for designing a single icon ignoring the icon usability.
- We gathered *IconCEPT10K*, the first icon dataset with high-level perceptual usability ratings, instead of low-level visual perceptual properties such as visual saliency.

We implemented EvIconas a web application so anyone can test EvIconon their own icon set. We will also release the source code, pretrained models, and the collected dataset (*IconCEPT10K*).

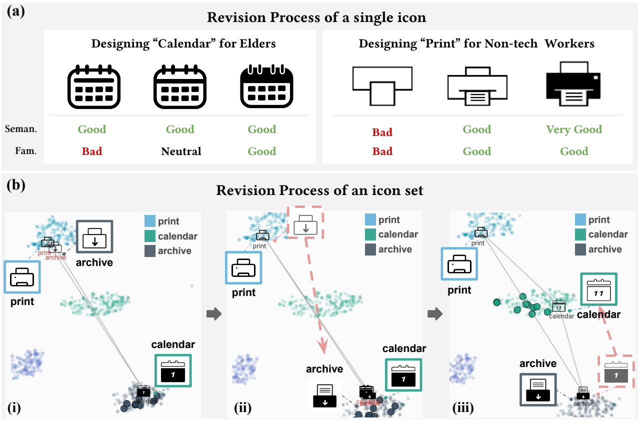


Figure 2: *EvIcon provides two types of instant perceptual usability feedback. (a) An UI designer can improve a single icon's usability and target different demographic users (e.g., elder people or non-tech workers) with the "semantic distance" and "familiarity" feedback. (b) Moreover, an UI designer can improve the usability of an icon set by (i) identifying poor visual distinguishability and (ii) revise the "archive" icon and (iii) the "calendar" icon using the visual distinguishability graph.*

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 [Git86, Hor94, Hor96] provide a thorough introduction on how to design usable icons and recommended practices. While usability is broadly defined [KK95, Tra18, Tra20], icon usability is predominantly linked to the icon's ability to convey its meaning and the users' ability to comprehend it. Previous research highlights features such as visual complexity, semantic distance, and familiarity as major influences on icon usability [MCdB99, MCdB01, MI09, IMC07, SM14, KFZI20, SJ16, SKJ17]. These features are often operationally defined through user ratings or rankings on various scales (e.g., the perceived closeness between an icon and its represented information or the perceived aesthetics of icons) [MCdB01, MI09, IMC07, SJ16]. Studies have discovered that icons with differing levels of these features can influence user behaviors [MCdB01, IMC07, SKJ17] and cognitive responses during interaction [CLKL16]. User age [LMG11] and experience [IMC07, AMW21] have also been found to impact how these features affect icon usability.

Researchers have proposed various methods to support icon design and generation due to the complex relationship between icons' features and usability. Zhao *et al.* [ZKH*20] developed a system to generate icons containing compound meanings automatically. Some works focus on generating icons based on filenames [LRFN04], data content [KPL08], and man-made object category [SC21]. Other prior works focus on learning icons' appearance similarity [LGG18], creating scale variations of icons [BL15], and selecting an icon set based on crowdsourced ratings [LKC*16]. Compared to previous works [LKC*16, LGG18], our system lets designers devise the final icon set on their own with our perceptual usability feedback instead of directly obtaining an icon set from an optimization process.

2.2. 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 utilized personal editing histories to assist 2D sketch [XCW14], 3D shape sculpturing [PXW18], and viewpoint selection [CGW*14]. On the other hand, various prior works have incorporated real-time physical simulation into their interactive tools for designing physically valid furnitures [UIM12] and model airplanes [UKSI14]. Among them, many recent works leveraged collected visual content data to assist 2D sketch [LZC11], multi-view clipart design [SLS*21], and mobile apps user interface design [LCS*18, DHF*17a, DHF*17c]. Other studies crowdsourced and modeled large-scale users' perception about tappability for the mobile interfaces [SL19] and visual importance on graphic designs [BKO*17] to assist designers in diagnosing the perceptual issues in their designs. Additionally, Rosenholtz *et al.* [RDF11] conducted a thorough qualitative study with professional design teams and showed that designers benefited from tools with low-level perceptual prediction in the agile assessment of usability. Unlike previous works that only focused on providing low-level visual perceptions feedback, we provide high-level usability feedback such as semantic distance and familiarity. Moreover, we provide visual distinguishability feedback to support revising an icon set's usability, which is rarely addressed in prior related research.

2.3. Human-in-the-loop Exploration

Prior studies have demonstrated the feasibility of conducting usability evaluation on crowdsourcing platforms via performing benchmark user testings [KRG13] and collecting human visual importance [BKO*17]. As exploring various huge design spaces with usability evaluations is a ubiquitous task in visual design, this task is realized by various interactive optimization techniques, including interactive evolutionary computation [Tak01] and human-in-the-loop Bayesian optimization [KSI14, KSSI17, KSG20, CSSI21, BBDF10]. Unlike previous methods, our human-in-the-loop framework focuses on providing instant perceptual usability feedback to support users' exploration instead of providing the final design using the optimization-based method due to the following reasons. First, the state-of-the-art human-in-the-loop optimization methods work best in relatively lower-dimensional parameter spaces (e.g., $6-15$) [KSI14, KSSI17, KSG20, CSSI21, BBDF10], whereas reducing the design dimensions of interface icons into such low dimensions would omit the nuanced features that are crucial for the high-level usability perceived by users and designers. Hence, our framework makes designers finalize the icons and the icon sets iteratively and manually. Second, previous human-in-the-loop optimization methods use "selection" as the main interaction approach, whereas the task of icon design requires more complicated design interactions than selections [ZKH*20]. Therefore, the current human-in-the-loop optimization methods are not suitable for the inputs of our framework to design high-usability icons.

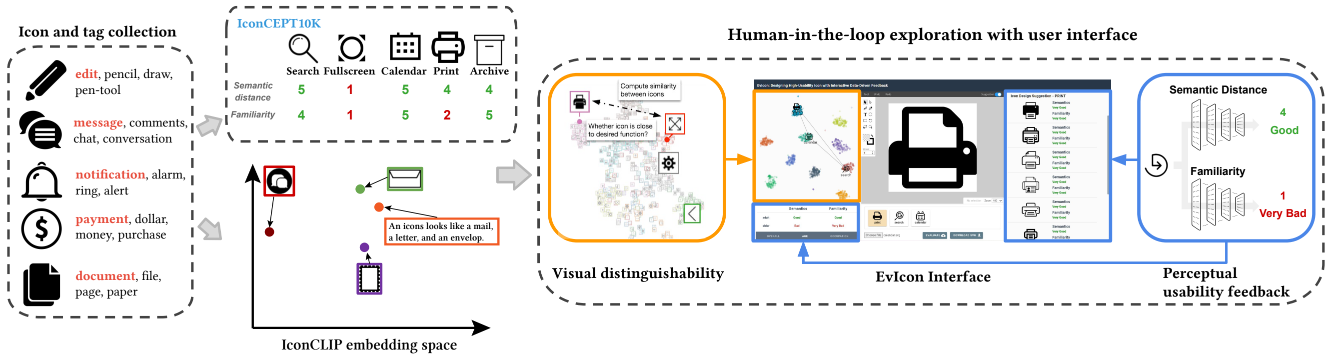


Figure 3: ***Overview of EvIcon.*** *We collect a large-scale icon and tag collection. And we compiled a dataset IconCEPT10K, comprises* $10,000$ *icons across* 50 *base tags, their associated tags, and crowdsourced semantic distance and familiarity ratings. We also fine-tune a pre-trained joint text-image embedding (CLIP) into IconCLIP using this collection. EvIcon computes and presents designers with instant perceptual usability feedback to assist revising high-usability icon sets.*

3. Problem Overview

Given an interface icon set $\mathcal{I}$ provided by a designer. The goal of our framework is to assist this designer in revising the usability of prepared icons into a new interface icon set $\hat{\mathcal{I}}$ efficiently. We expect that each icon I in $\mathcal{I}$ is associated with n text tags ($\mathcal{T}_I = t_0, t_1, \ldots, t_{n-1}$) that represent the semantic and visual concepts of the icon such as "search", "next", "television", and "map". We characterize the usability of an icon using common perceptual usability metrics including *semantic distance*, *familiarity*, and *visual distinuguishability*, which are commonly used by professional icon designers [MCdB01, Kur00, SM14]. However, these metrics of an icon are usually hard to evaluate mathematically from the icon image since the assessments of these metrics require extensive user testing to collect users' self-reports and feedback. Hence, we collected a large-scale icon dataset and the crowdsourced perceptual ratings of these icons on Amazon Mechanical Turk (AMT). We used the collected ratings to train usability classifiers. For each tag A, we trained a separate classifier f_A^{sd} and f_A^{fam} for classifying the semantic distance and familiarity of an icon belongs to tag A. For each classifier, it predicts "Very Good", "Good", "Neutral", "Bad", and "Very Bad" as the different levels for the semantic distance and familiarity. The goal of our framework is to enable users to revise an icon $I \in \mathcal{I}$ that maximizes the following perceptual usability function:

$$i^* = \arg\max(w_{sd}\phi_{sd}(I,\mathcal{T}_I) + w_{fam}\phi_{fam}(I,\mathcal{T}_I) + w_{vd}\phi_{vd}(I)), \quad (1)$$

where ϕ is a semantic perceptual function. In our work, the semantic perceptual function comprises three parts:

- semantic distance: $\phi_{sd}(I,\mathcal{T}_I) = P(f_A^{\text{sd}} == \text{Very Close}|I,\mathcal{T}_I)$
- familiarity: $\phi_{fam}(I,\mathcal{T}_I) = P(f_A^{\text{fam}} == \text{Very Good}|I,\mathcal{T}_I)$
- visual distinguishability: $\phi_{vd}(I) = \sum_{J\in\mathcal{I}} \|\rho_I - \rho_J\|_2^2$,

where $P(f_A^{\text{sd}} == \text{Very Close})$ stands for the probability of an icon being classified as having the "Very Close" semantic distance.

To measure visual distinguishability, it is important to measure the distance with respect to the semantic concept difference instead of just pixel-level difference. To achieve this, we obtain an embedding space where icons of the same tags stay closer to each other than those of different tags. We describe how to obtain this embedding space in Section 5.2. The embedded coordinates of icon i in this space are represented by ρ_i. The goal of $\phi_{sd}(i)$ is to encourage the revised icon to be classified as "Very Close," while the aim of $\phi_{vd}(i)$ is to separate the refined icon from other icons in $\mathcal{I}$. To optimize Equation 1 and iteratively refine the icon set $\mathcal{I}$, designers need to be involved in the process to specify their design requirements. Instead of providing designers with automatic synthesis results, we have developed an interactive interface that guides them in designing highly usable icons.

4. EvIcon User Interface

We propose an interactive and exploratory design tool, EvIcon, to present perceptual usability feedback of an individual icon and visual distinguishability between icons. Our interface augments existing vector graphics design tools with additional usability feedback panels. As shown in Figure 3, our interface contains three main panels: (i) the main canvas panel which includes a vector graphics editor for icon revision and a list to present the uploaded icon set, (ii) the perceptual feedback panel (box with blue borderline), and (iii) the distinguishability visualization panel (box with orange borderline).

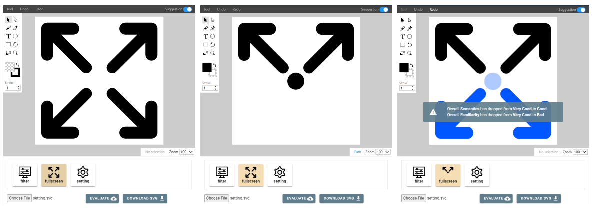
4.1. User Workflow

To use EvIcon, a designer first prepares a set of icons and corresponding tags under designing. Next, the designer can select an icon from the icon set, and EvIcon would infer its predicted usability. Designers can use the feedback panel to check predicted usability, revise icons to improve usability, and inspect visual distinguishability with an interactive graph. This process is repeated until the usability and distinguishability of icons meet satisfaction.

4.2. Interface Components

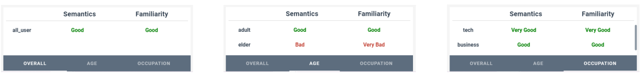
4.2.1. Main Canvas Panel

The designer can revise the icons using the vector graphics editor in this panel. During the iterative revision process, EvIcon also



(a) Initial icon (b) Modified icon (c) Two types of in-place warnings

Figure 4: *A warning will be displayed in place to draw attention to the poor adjustment compared to the last usability inspection. Blue highlights will indicate the paths suggested to be added back, while light-blue highlights will mark those suggested to be removed.*



(a) Overall (b) Age categories: i) adult and ii) elder (c) Occupation categories: i) technology, ii) business, and iii) others

Figure 5: *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.*

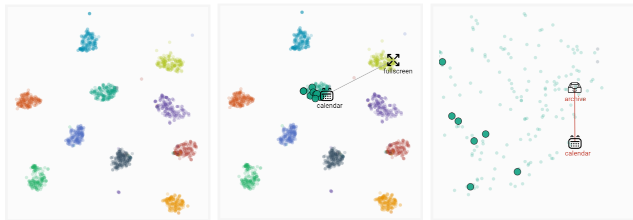
provides in-place warnings when the predicted perception usability drops. This in-place visual warning is helpful for building the connection between the revised icon and the perceptual prediction. We highlighted the paths of an icon that we encourage the designers to add and remove in two different colors as shown in Figure 4.

4.2.2. Perceptual Usability Feedback Panel

EvIcon shows the predicted level of perceptual usabilities (semantic distance and familiarity) of the icon under revision (Figure 5). Designers can switch between tabs to assess predicted usabilities for specific demographic target audiences. To present the levels of semantic distance and familiarity in a way designers can easily understand, instead of showing rating scores 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.

4.2.3. Distinguishability Visualization Panel

EvIcon presents an interactive distinguishability graph to help designers compare the relative visual distance between icons in the prepared icon set $\mathcal{I}$. After the designers revise an icon, they can check the updated embedded coordinate of the icon. We connected the icons in the prepared icon set using grey links (as shown in Figure 6(b)) and changed the color of the links into red if the connected icons were too close to each other (see Figure 6(c)). This interactive design aims to warn designers of the inadequate visual distinguishability in the prepared icon set, and prevent them from refining icons that fall into the wrong tag.



(a) (b) (c)

Figure 6: *(a) Different color-codings indicate different semantic concept clusters. The icons are linked (b) in grey but will change (c) to red in order to notify poor visual distinguishability.*

5. EvIcon Implementation

5.1. Icon and Crowdsourced Perceptual Rating Dataset

Our goal of data collection is to gather an icon dataset covering a comprehensive range of tags that UI designers are likely to design. Since unlimited tags exist for interface icons, it is impractical to enumerate them all and collect them at once. To address this issue, we expanded the tags we can cover by adopting the following data collection procedure. First, we collected icons of 212 base tags reported in prior work [LCS*18], including “Search”, “Crop”, “Message”, “Pause”, “Filter”, “Calendar”, and “Archive”. We collected these single-colored icons from multiple online resources, including Google Material Icons, Icon8, and The Noun Project. Although these icons are collected from different websites, they share similar visual styles due to the prevalence of flat UI design [Arl14, SRS18]. Overall, we collected $2,613,438$ single-colored icons and their associated tags provided by the original designers. There are $191,472$ unique tags representing a wide range of concepts, and they provided us with a rich resource to model the relationship between icons and tags. We then used this icon and tag collection to train a joint image-text embedding.

However, it is tedious and repetitive to collect users’ perceptual usability ratings for all icons; thus we selected the top 50 base tags that are semantically independent by analyzing their distribution in the Word2Vec [MCCD13] embedding space. For each selected base tag, we further selected 200 representative icons with respect to the uniqueness of icon shapes using the following process. 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. Then, 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 [AV07] on these projected icon representations and set $K = 10$ based on the results of the Elbow method (i.e., ten clusters in a subset) [KS96]. We obtained 200 icons from each base tag by randomly sampling 20 icons from each cluster. After repeating the same process to all base tags, we acquired the curated dataset with in total $10,000$ icons in which the variety of icons of each function increased compared to the raw dataset.

After obtaining the curated dataset, we used Amazon Mechanical Turk (AMT) to collect users' perceived semantic distance and familiarity with the selected 10,000 icons. We recruited 5,559 workers participating in the crowdsourcing task (3,498 males and 2,061 females; mean age = 33.1 with a standard deviation of 8.90). The workers' self-report ages and occupations were divided into three age levels (elder: age $>$ 50 yrs; adult: 50 $>$ age $>$ 20 yrs; teenager: age $<$ 20 yrs) and occupational categories (technology, business, and others) which are used as the demographic information of their ratings when building perceptual usability prediction. Each worker finished five assignments and rated icons of five tags in each assignment (i.e., 25 icons in total) with an average completion time of 8 minutes. The workers were asked to rate each icon on a 5-point Likert scale to specify their assessment of the icon's semantic distance and familiarity [MCdB99, IMC07]. The workers also rated their perceived familiarity with each tag on the same 5-point Likert scale. We described the rating distribution of the 50 base tags and the content of the questions in Section 1 of the supplementary material. The order of the icons was randomized. In general, we spent two days collecting all the rating data in parallel using MTurk API. In the final rated dataset, we collected 138,964 unique ratings. We describe the details of the distribution of the collected ratings and the AMT crowdsource task in Figure 1 of the supplemental material. We will include the selected 10,000 icons and the collected perceptual usability ratings as our *IconCEPT10K* dataset.

5.2. Perceptual Usability and Visual Distinguishability Feedback

Given an input icon I and its associated tags $\mathbf{T} = t_0, t_1, ..., t_{n-1}$, we want to build a classifier that can predict its perceptual usability ratings (semantic distance and familiarity). However, there are unlimited possible tags designers want to design; and it is tedious to collect icons of all possible tags and their perceptual usability ratings. Thus, it is vital to design a classification method to predict the perceptual usability ratings for icons of unseen tags. To address this need, we designed our classification method based on the pre-trained joint embedding (CLIP) [RKH*21] which is learned from loose image-text pairing information.

5.2.1. Introduction to CLIP Embedding Space

CLIP [RKH*21] is a joint image-text embedding trained on 400 million text-image pairs. The representations learned by CLIP have been shown to be effective for various downstream tasks such as zero-shot image classification. CLIP jointly trains an image encoder g and a text encoder h, that map images and text into a shared embedding space. Unlike previous works on natural image editing using CLIP embedding space [PWS*21, AZF*21], the target image domain of our application (single-colored icon image) is different from the training images used in the pre-trained CLIP model. Thus, instead of using the pre-trained CLIP model to extract image and text representations directly, we finetune the original CLIP model using our icon dataset to obtain IconCLIP.

Finetuning CLIP on icon image We let $S_{\text{icon}} = \{(I_i, \mathbf{T}_i) | i = 0, ..., N\}$ denote the icon dataset used for finetuning the original CLIP model. For each icon I_i, we converted the associated tags $\mathbf{T}_i$ into a sentence s_i using the prompt template "A icon looks like a $\{tag_0$, tag_1, ..., $tag_{n-1}\}$". We use a pre-trained CLIP ViT-B/32 model as the base model, which uses ViT-B/32 [DBK*21] as the image encoder and Transformer [VSP*17] as the text encoder. We follow the training procedure described in the original CLIP [RKH*21]. Given a training pair (an icon image I_i and a sentence s_i), CLIP produces a scalar score: $g(I)^T h(s_i)$ that is high when the image and text are mismatched. We finetune the pre-trained model by minimizing a symmetric InfoNCE loss [VdOLV18].

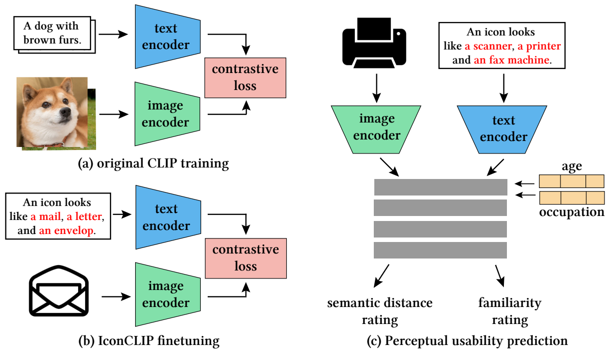


Figure 7: *(a) The general-purpose CLIP [RKH*21] is a joint image-text embedding trained on 400 million text-image pairs. (b) We fine-tune the general-purpose CLIP into IconCLIP using "icon tags"-"icon image" pairs. (c) We predicted the perceptual usability ratings of an input icon and its associated tags using the IconCLIP embedding space and the target demographic information.*

5.2.2. Perceptual Usability Prediction

We designed our classifier F_Θ using a deep fully-connected neural network without convolutional layers (i.e., a MLP). As illustrated in Figure 7(c), F_Θ takes three different inputs: the input image embedding, the input sentence embedding, and the discrete demographics vector (*age: three levels* and *occupation: three categories*) and the output are five usability ratings of semantic distance and familiarity. We obtained the image and sentence embedding using the image encoder and the text decoder of IconCLIP. And F_Θ process these inputs with four fully-connected layers (using ReLU activations and 256 channels per layer.

5.2.3. Visual Distinguishability

As discussed in Section 4.2.3, EvIcon provides a distinguishability graph to help users compare the relative visual distance between icons in the prepared icon set. We directly use the embedding space of the finetuned IconCLIP as our similarity measurement space. We used Uniform Manifold Approximation and Projection (UMAP) [MHSG18] to project the 512d feature vector to 2d.

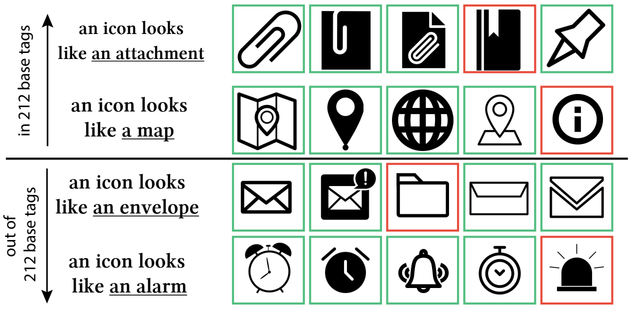


Figure 8: ***text-to-image retrieval by IconCLIP.*** *Among the four tags we used as queries, only "attachment" and "map" are in the 212 base tags. The IconCLIP embedding space recognizes the meaning of "envelop" and "alarm" because we used the tags associated with icons (we use the green box for positive results and the red box for negative results).*

6. Evaluation

6.1. Evaluation of IconCLIP

To evaluate the IconCLIP embedding space, we performed a top-k image retrieval evaluation. We split the overall collected icons into a training set and a testing set. We used the training set to fine-tune IconCLIP. Using the fine-tuned IconCLIP, we conducted the retrieval test in the following manner: for each icon image within the testing set, we use it as a query and retrieved the most similar icon image from the entire testing set. And we consider a retrieved icon image as a positive result if it shares a common tag with the query image. The MAP@5 (mean average precision at rank 5) of the retrieval test is 74.3. On the other hand, we also performed a text-to-image retrieval test and showed the qualitative image retrieval results in Figure 8. We can observe that the top-5 nearest neighbors match the tags in the sentence even when the concepts are not in the 212 base tags we used for collecting the icons. This suggests that the tags associated with the icons expand the embedding space of IconCLIP.

6.2. Evaluation of perceptual usability feedback models

As mentioned in Section 5.2.2, we trained a unified network to predict usability ratings based on the icon's image embeddings, tag embeddings, and demographic information. To demonstrate the ability to predict usability ratings of unseen tags, we split the icons of 50 base tags into 45 tags as seen and 5 tags as unseen tags. It should be noted that we only use the base tags as the selection criteria, but we used all the associated tags within the base tags as training signals, so it is not restricted to these base tags. We performed two types of evaluation of our prediction model.

6.2.1. In-domain Evaluation

First, we evaluated the prediction precision and recall on the 45 base functions we used for training. Among the icons of these 45 base functions, we randomly split the data into 90/10 as training/testing data. For *semantic distance*, our models achieved 83.6% for precision and 84.1% for recall. For *familiarity*, our models achieved 76.3% for precision and 77.6% for recall.

6.2.2. Out-of-domain Evaluation

Second, we also evaluated the prediction precision and recall on all icons belonging to the 5 base tags we held out during training. For *semantic distance*, our models achieved 66.4% for precision and 69.5% for recall. For *familiarity*, our models achieved 67.1% for precision and 68.4% for recall.

6.3. Evaluation with UI Designers

To evaluate how EvIcon's interaction design can support designers' revision process, we conducted a user study with six professional UI designers (five females and one male; ages ranging from 22 to 34 years old). We recruited a similar number of domain experts with prior similar works [PXW18, XKG*16, SSII18]. The self-reported professional experience of the designers ranges from one to ten years. All of them used Adobe Illustrator[†] for initial icon design.

6.3.1. Procedure and Tasks

After introducing EvIcon and the meaning of two types of feedback, the designers practiced using EvIcon for ten minutes. We then asked them to complete the practice tasks (e.g., reporting the perceptual usability of an icon in different age groups of users) to ensure they understand how to use EvIcon. In the formal sessions, the designers were asked to improve the usability of two icon sets. For the design brief to guide the designers when revising icons, we informed the designers the scenario of the evaluation is that a client asked them to improve the icon sets so that these icons can be used in a wide range of applications and users (e.g., elders). Each icon set contains three icons of tags "Archive", "Print", and "Filter". We selected these tags based on their average familiarity level collected via the crowdsourced study in Section 5.1 ("Archive": 3.8; "Filter": 3.9; "Print": 4.2) to ensure we included established and uncommon tags in the evaluation. Moreover, "Archive" and "Print" are in the 45 seen tag set, and "Filter" is in the 5 unseen tag set used in Section 6.1. We denoted these icons as the original icons in the following sections.

As shown in Figure 9, the icons in the two sets are different, and we instructed each designer to improve the usability of one icon set with EvIcon and another set without EvIcon, both in fifteen minutes. The combination of the icon set and two conditions were randomly assigned, and the order of conditions was counterbalanced to avoid the learning effect. Under both conditions, the designers can freely edit icons using the design tool of their choice and search online for the information. However, under the without EvIcon condition, the designers can not access the icons and perceptual ratings we collected. We recorded the revision process and the revised icons. In the end, we obtained 36 revised icons from six designers in total, and we found that all designers spent the entire time budget (15 minutes) for each condition.

† https://www.adobe.com/products/illustrator.html

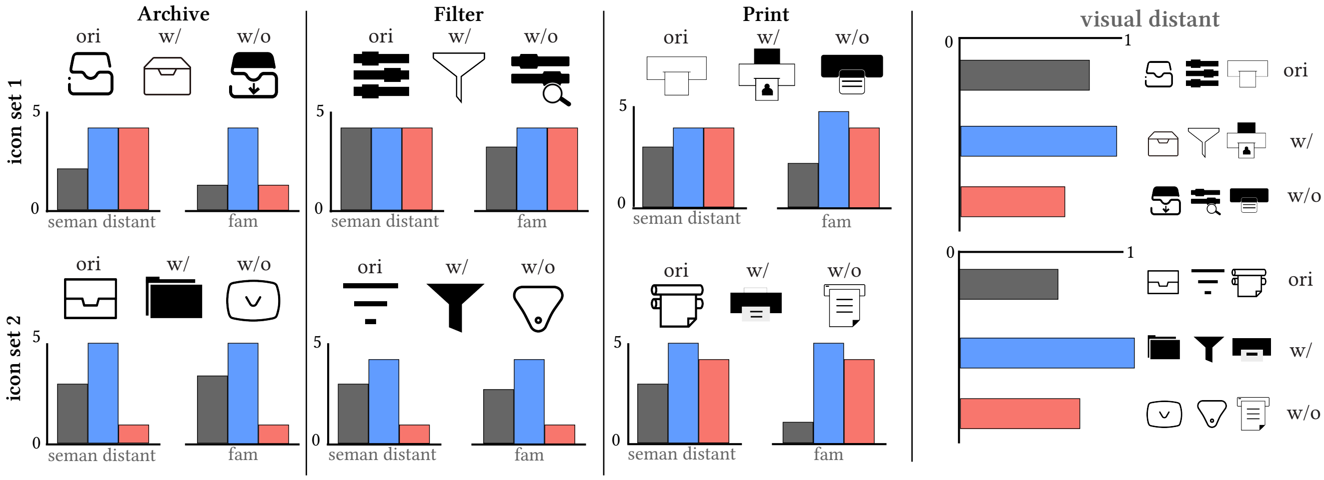


Figure 9: *The six original icons and their examples of revised icons with and without EvIcon. We plot the crowdsourced evaluation results ("AMT rating") of each icon. The gray/blue/red bar denotes the AMT rating of original icon/icon revised with EvIcon/icon revised without EvIcon. The ratings ranged from 1 ("Very Bad") to 5 ("Very Good") for the bar chart of semantic distance (seman distant) and familiarity (fam). We can see that most of the icons revised with EvIcon received higher AMT ratings than icons revised without EvIcon. We also show the visual distinguishability score between each icon in the embedding space. The visual distinguishability between icons revised with EvIcon is the furthest.*

6.3.2. Result

Revised icons In Figure 9, we show the revised icons of all three tags with and without using EvIcon. To further verify that EvIcon can help designers improve icons' usability, we launched an additional crowdsourced evaluation on AMT to collect 213 (140 males and 73 females; 19 to 64 years old) crowdworkers' usability ratings of all original and revised icons pair in an assignment. We collected averaged 57.8 unique ratings for each original/revised icon pair. Each crowdworker would only rate an icon pair revised by the same designers to eliminate the influence of individual designers' abilities. We used the majority vote of all received ratings as the final rating of each revised icon to reduce the effects of spammers as shown in Figure 9. To reduce the mutual influence of icons in different pairs of revised and original icons, we calculated the final ratings of the original icons by averaging the majority vote ratings across the different pairs of revised icons provided by the designers.

We can see from Figure 9 that most of the revised icons with EvIcon (blue bars) obtained higher AMT ratings of semantic distance and familiarity than those without EvIcon. Moreover, to demonstrate the usefulness of EvIcon in improving the visual distinguishability within the icon set, we computed the mutual distances between the 512-dim embedded vector of each revised icon. In the rightmost panel of Figure 9, the mutual distance between icons revised with EvIcon is farther than the original icons and icons revised without EvIcon, which suggests better visual distinguishability.

Figure 10 illustrates example revision processes for the icons "Archive" and "Print." Throughout each design step, EvIcon provided feedback on "Semantics" (semantic distance) and "Familiarity." The crowdsourced evaluation results, collected via Amazon Mechanical Turk (AMT), are displayed next to the finalized icon (the right-most icon of each block in Figure 10) for each revision process. The evaluation outcomes demonstrate that icons revised with EvIcon generally outperformed those revised without it, as shown in Figure 10, with higher ratings given for both "Semantics" and "Familiarity."

Revised icons for diverse demographics We investigated whether icons revised with EvIcon result in higher usability ratings from older users (> 50 years old) to demonstrate the tool's ability to create more inclusive designs for users with diverse demographics. We found that the revised icons obtained higher AMT mean semantic distance (with: 3.49; without: 3.35) and familiarity (with: 2.83; without: 2.81) ratings across tags. We show the two examples of the revised icons using EvIcon in Figure 11. Additionally, Figure 2 in the supplementary material showcases visible differences among top-rated usability icons across various demographics, supporting the need for designing diverse icons tailored to specific demographic groups.

Crowdsourced evaluation on revised icons As shown in Figure 12, we compared the averaged mode ratings of the revised icons by their tag and whether they were revised with EvIcon using Cohen's d. We can see that for the icons revised by all designers, the "Archive" and "Filter" icons revised with EvIcon received a higher level of semantic distance (Archive: d = 1.48; Filter: d = 0.63; Figure 12(a)) and familiarity (Archive: d = 1.02; Filter: d = 0.4; Figure 12(b)) than those without using EvIcon with the moderate to the large magnitude of the mean difference. Yet, the "Print" icons revised with EvIcon obtained the same level of semantic distance and familiarity as those without EvIcon. These results suggest that the designers benefit most from using EvIcon in improving the usability of icons of unestablished tags (i.e., "Archive" and "Filter"). Since most of the designers and users have not formed the common visual metaphors for the unestablished tags, EvIcon's

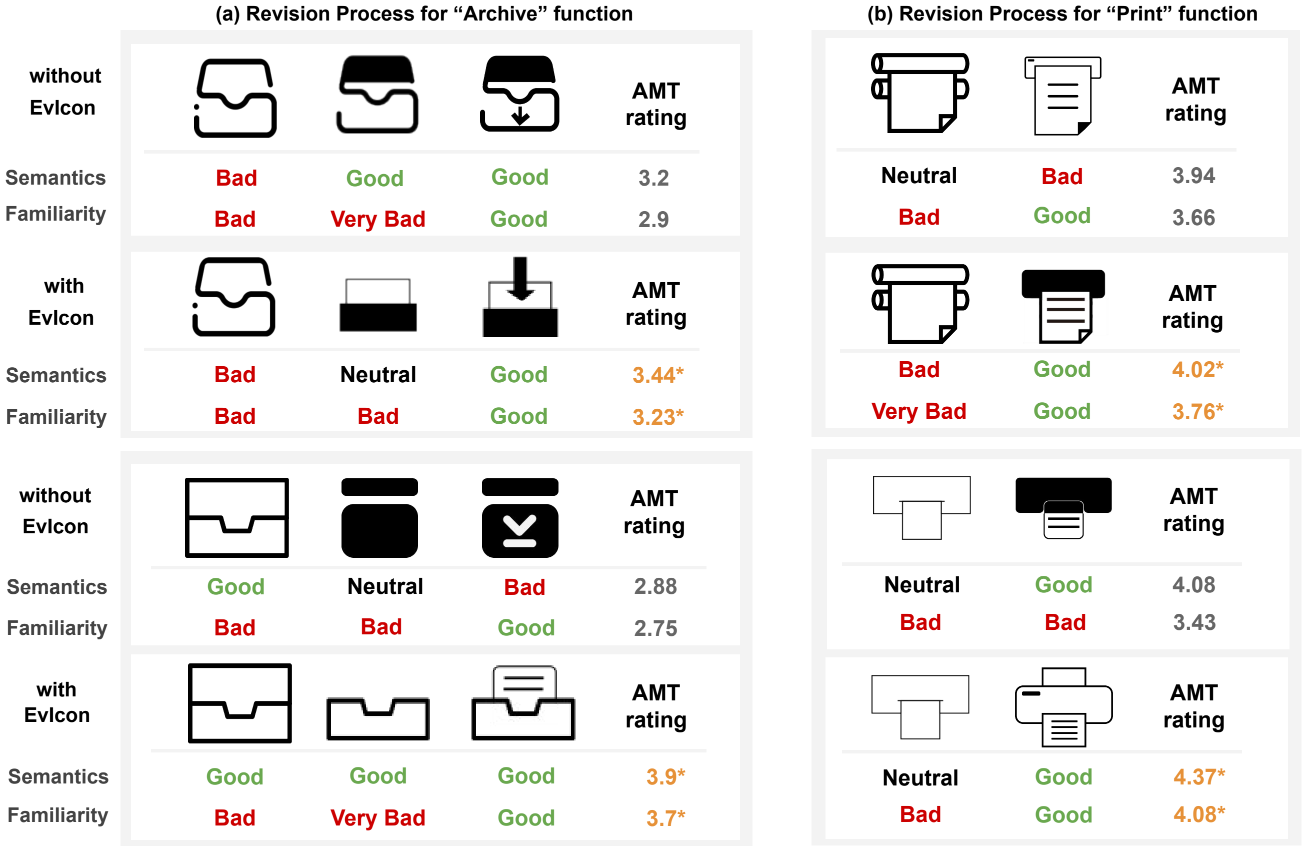


Figure 10: *Icon revision process starting from left to right by designers with and without EvIcon. (a) presents the revision processes for "Archive" icons of two designers, and (b) presents the revision processes for "Print" icons of two designers. Eight groups of revision process with the prediction of perception feedback and the crowdsourced evaluation results ("AMT rating") of the finalized icons are presented (the better AMT ratings are highlighted with the star symbol).*

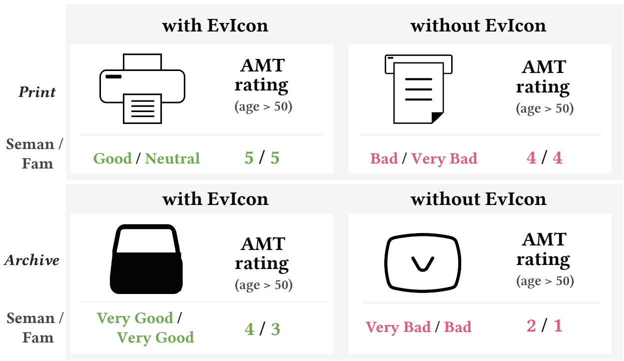


Figure 11: *We show two examples of revised icons with and without EvIcon. We can see that the icons revised by EvIconwith better predicted semantic distance and familiarity levels also achieved higher AMT ratings from the elder crowdworkers (age > 50).*

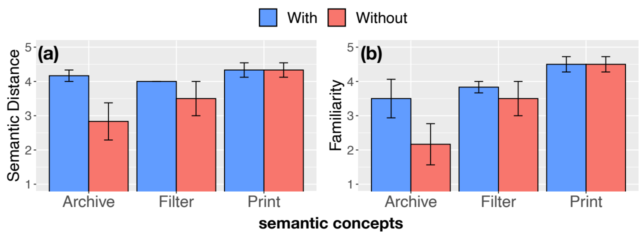


Figure 12: *The AMT ratings of the icons revised by all designers. (a) The ratings of semantic distance rating. (b) The ratings of familiarity rating. The error bars represent the standard deviation.*

feedback helps designers navigate the vast variations of "Archive" and "Filter" icons and find the best way to revise the icons.

Revised icon retrieval test The retrieval test aims to demonstrate that the existing icons in our dataset are used primarily as inspiration rather than copied directly. In Figure 13, we present the closest example from our dataset for each revised icon and provide the PSNR/SSIM scores. We observe that for the semantic concept with simpler shapes, such as "Filter," the revised icons are generally closer to the existing icons in our dataset. However, for the semantic concept with more complex shapes, designers tend to make more significant revisions (e.g., "Print", "Archive"), resulting in greater distances between the revised icons and their closest examples.

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 understanding more easily 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 experienced struggling to design interface icons targeting a specific category of users while having limited knowledge or access to the users:

- *"The feedback from a 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! Although icons play an important role in interface design, there is 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 the suggestion panel to see how they could improve their design. P2, P3, P5, and P6 mentioned they could derive some graphical design features 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' thoughts at the beginning of the design flow."*(*P5*)

Designers also discussed possible benefits EvIcon could bring if 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 discussions. As designers, participants usually care a lot about aesthetics while designing icons, EvIcon could also provide assistances to balance between aesthetics 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 go 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 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 generally be useful and mentioned possibilities of how EvIcon can assist in different design phases. Moreover, they are willing to use EvIcon in their design process if it becomes a mature product in the future.

7. Limitations and Future Work

Single icon style As flat design continues to be a popular trend in digital design, our framework currently focuses on improving the usability of single-colored icons. However, we recognize the importance of expanding EvIcon's capabilities to include a wider range of icon styles. To achieve this, we plan to build a diverse dataset of icons and use metrics proposed in [GAHG17] to explore and compare icons in different styles. By doing so, we aim to improve the generalizability of EvIcon and make it more adaptable to the changing trends and preferences in digital design. Expanding the dataset and incorporating new metrics will enable EvIcon to produce icons in a wider variety of styles, ensuring that it remains a valuable tool for UI designers across different industries and contexts.

User interface with limited icon editing functions While EvIcon's vector graphics editor offers a basic set of design tools, it may not be sufficient for the needs of some UI designers. The absence of advanced features could limit creativity and lead to a less efficient workflow. Our plan to integrate EvIcon as a plugin for professional design tools such as Adobe Illustrator and Sketch[‡] is aimed at addressing these limitations. By providing access to a more comprehensive suite of design tools, UI designers can expand their creative options and improve their efficiency. The integration will enable designers to access EvIcon's icon creation and editing features within their preferred design software, eliminating the need to switch between multiple tools. Ultimately, this will enable designers to produce higher quality icons more efficiently, resulting in better user experiences for their products.

Supporting validations for general use of icons While the proposed framework is primarily focused on supporting designers in validating and revising icons for user interface design, it is important to note that icons have many other applications beyond UI design. Icons are widely used in presentation slides, infographics, and other forms of visual communication, where their design requirements may differ from those in UI design. For example, the icons used in infographics may require better abilities to convey information rather than better familiarity with viewers. Given the versatile

‡ https://www.sketch.com/

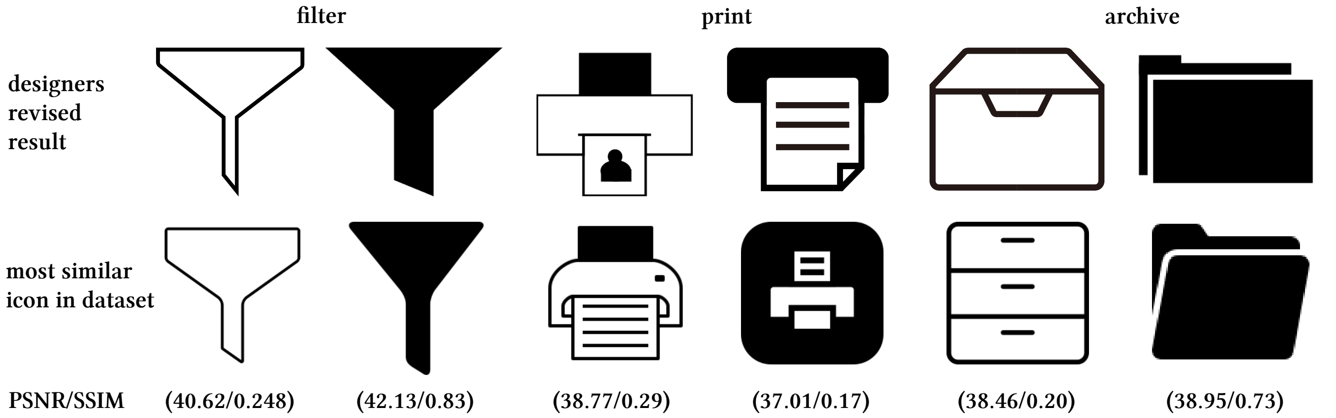


Figure 13: *For each icon revised by the designers, we show its closest example in our dataset and its corresponding PSNR/SSIM at the bottom.*

nature of icons, we plan to extend the usage scenario and target audience of EvIcon to support icon improvement and selection for more general purposes. By doing so, we aim to make EvIcon a more versatile tool that can assist designers across various fields and contexts, not just limited to UI design. Expanding EvIcon's capabilities to accommodate different design requirements and user needs will enhance its value and relevance, making it an even more valuable tool for designers working on a wide range of projects. Ultimately, this will enable designers to create more effective and engaging visual content across various domains, resulting in better user experiences for their audiences.

8. Conclusion

In this paper, we propose a human-in-the-loop framework called EvIconthat aims to enhance the usability of interface icon sets. Our framework includes a novel perceptual usability formulation and an interactive design tool that enable users to modify icons' effectiveness in conveying information. We also introduce the first icon dataset, *IconCEPT10K*, which features high-level perceptual usability ratings, such as semantic distance and familiarity, from over 5,000 crowdworkers. To demonstrate the effectiveness of EvIcon, we conducted a user study with six UI designers. Our quantitative and qualitative results show that using EvIconresulted in an icon set with improved usability, as rated by over 200 crowdworkers. These findings suggest that EvIconis an effective tool for facilitating the design process.

9. Acknowledgement

This work was supported in part by JSPS Grant-in-Aid JP23K16921, JP21F20075, and National Science and Technology Council, under Grant 111-2222-E-194-008-MY2, NSTC111-2634-F-002-022, 111-2221-E-002-145-MY3, 111-3111-E-002-002, 112-2218-E-002-029, and 108-2221-E-009-089, and National Taiwan University (NTU112L900902).

References

[AMW21] ALI A. X., MCAWEENEY E., WOBBROCK J. O.: Anachronism by design: Understanding young adults' perceptions of computer iconography. *International Journal of Human-Computer Studies* (2021), 102599. 3

[Arl14] ARLEDGE C.: Filled-in vs. outline icons: the impact of icon style on usability. In *Master Paper, Master of Science in Information Science* (2014), University of North Carolina at Chapel Hill. 5

[AV07] ARTHUR D., VASSILVITSKII S.: K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (USA, 2007), SODA '07, Society for Industrial and Applied Mathematics, p. 1027–1035. 5

[AZF*21] ABDAL R., ZHU P., FEMIANI J., MITRA N. J., WONKA P.: Clip2stylegan: Unsupervised extraction of stylegan edit directions. *arXiv preprint arXiv:2112.05219* (2021). 6

[BAR92] BAILEY R. W., ALLAN R. W., RAIELLO P.: Usability testing vs. heuristic evaluation: A head-to-head comparison. In *Proceedings of the human factors society annual meeting* (1992), vol. 36, SAGE Publications Sage CA: Los Angeles, CA, pp. 409–413. 1

[BBDF10] BROCHU E., BROCHU T., DE FREITAS N.: A bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Goslar, Germany, 2010), pp. 103–112. 3

[BKO*17] BYLINSKII Z., KIM N. W., O'DONOVAN P., ALSHEIKH S., MADAN S., PFISTER H., DURAND F., RUSSELL B., HERTZMANN A.: Learning visual importance for graphic designs and data visualizations. In *Proceedings of the 30th Annual ACM symposium on user interface software and technology* (New York, NY, USA, 2017), pp. 57–69. 3

[BL15] BERNSTEIN G. L., LI W.: Lillicon: Using transient widgets to create scale variations of icons. *ACM Transactions on Graphics (TOG) 34*, 4 (2015), 1–11. 3

[CGW*14] CHEN H.-T., GROSSMAN T., WEI L.-Y., SCHMIDT R. M., HARTMANN B., FITZMAURICE G., AGRAWALA M.: History assisted view authoring for 3d models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI '14, ACM, pp. 2027–2036. URL: http://doi.acm.org/10.1145/2556288.2557009, doi:10.1145/2556288.2557009. 3

[CLKL16] CHERNG F.-Y., LIN W.-C., KING J.-T., LEE Y.-C.: 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* (2016), ACM, pp. 4378–4389. 2, 3

[CSSI21] CHONG T., SHEN I.-C., SATO I., IGARASHI T.: Interactive optimization of generative image modelling using sequential subspace search and content-based guidance. In *Computer Graphics Forum* (2021), vol. 40, Wiley Online Library, pp. 279–292. 3

[CUG20] CHAJADI F., UDDIN M. S., GUTWIN C.: Effects of visual distinctiveness on learning and retrieval in icon toolbars. In *Proceedings of the Graphics Interface Conference* (Ontario, Canada, 2020), ACM. 2

[DBK*21] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J., HOULSBY N.: An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021). 6

[DHF*17a] DEKA B., HUANG Z., FRANZEN C., HIBSCHMAN J., AFERGAN D., LI Y., NICHOLS J., KUMAR R.: Rico: A mobile app dataset for building data-driven design applications. UIST '17, Association for Computing Machinery, p. 845–854. URL: https://doi.org/10.1145/3126594.3126651, doi:10.1145/3126594.3126651. 3

[DHF*17b] DEKA B., HUANG Z., FRANZEN C., NICHOLS J., LI Y., KUMAR R.: Zipt: Zero-integration performance testing of mobile app designs. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada, 2017), pp. 727–736. 1

[DHF*17c] DEKA B., HUANG Z., FRANZEN C., NICHOLS J., LI Y., KUMAR R.: Zipt: Zero-integration performance testing of mobile app designs. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2017), UIST '17, Association for Computing Machinery, p. 727–736. URL: https://doi.org/10.1145/3126594.3126647, doi:10.1145/3126594.3126647. 3

[GAHG17] GARCES E., AGARWALA A., HERTZMANN A., GUTIERREZ D.: Style-based exploration of illustration datasets. *Multimedia Tools Appl. 76*, 11 (jun 2017), 13067–13086. 10

[Git86] GITTINS D.: Icon-based human-computer interaction. *International Journal of Man-Machine Studies 24*, 6 (1986), 519–543. 3

[GSF01] GOONETILLEKE R. S., SHIH H. M., FRITSCH J.: Effects of training and representational characteristics in icon design. *International Journal of Human-Computer Studies 55*, 5 (2001), 741–760. 2

[HM10] HASSENZAHL M., MONK A.: The inference of perceived usability from beauty. *Human–Computer Interaction 25*, 3 (2010), 235–260. 1

[Hol16] HOLZINGER A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics 3*, 2 (2016), 119–131. 2

[Hor94] HORTON W. K.: *The ICON Book: Visual Symbols for Computer Systems and Documentation*. John Wiley & Sons, Inc., USA, 1994. 3

[Hor96] HORTON W.: Designing icons and visual symbols. In *Conference Companion on Human Factors in Computing Systems* (New York, NY, USA, 1996), CHI '96, Association for Computing Machinery, p. 371–372. URL: https://doi.org/10.1145/257089.257378, doi:10.1145/257089.257378. 3

[Hsi17] HSIEH T.-J.: Multiple roles of color information in the perception of icon-type images. *Color Research & Application 42*, 6 (2017), 740–752. 2

[IMC07] ISHERWOOD S. J., MCDOUGALL S. J., CURRY M. B.: 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 (2007), 465–476. 1, 3, 6

[KFZI20] KAMARULZAMAN N. A., FABIL N., ZAKI Z. M., ISMAIL R.: Comparative study of icon design for mobile application. In *Journal of Physics: Conference Series* (2020), vol. 1551, IOP Publishing, p. 012007. 3

[KK95] KUROSU M., KASHIMURA K.: Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability. In *Conference companion on Human factors in computing systems* (Denver, USA, 1995), pp. 292–293. 1, 3

[KPL08] KOLHOFF P., PREUSS J., LOVISCACH J.: Content-based icons for music files. *Computers & Graphics 32*, 5 (2008), 550–560. URL: https://www.sciencedirect.com/science/article/pii/S009784930800006X, doi:https://doi.org/10.1016/j.cag.2008.01.006. 2, 3

[KRG13] KOMAROV S., REINECKE K., GAJOS K. Z.: Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems* (New York, NY, USA, 2013), pp. 207–216. 3

[KS96] KETCHEN D. J., SHOOK C. L.: The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal 17*, 6 (1996), 441–458. 5

[KSG20] KOYAMA Y., SATO I., GOTO M.: Sequential gallery for interactive visual design optimization. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 88–1. 3

[KSI14] KOYAMA Y., SAKAMOTO D., IGARASHI T.: Crowd-powered parameter analysis for visual design exploration. In *Proceedings of the 27th annual ACM symposium on User interface software and technology* (New York, United States, 2014), pp. 65–74. 3

[KSSI17] KOYAMA Y., SATO I., SAKAMOTO D., IGARASHI T.: Sequential line search for efficient visual design optimization by crowds. URL: https://doi.org/10.1145/3072959.3073598, doi:10.1145/3072959.3073598. 3

[Kur00] KURNIAWAN S. H.: A rule of thumb of icons' visual distinctiveness. In *Proceedings on the 2000 conference on Universal Usability* (New York, United States, 2000), pp. 159–160. 2, 4

[LC20] LEGLEITER A. M., CAPORUSSO N.: Flat-design icon sets: A case for universal meanings? In *International Conference on Applied Human Factors and Ergonomics* (Virtual Only, 2020), Springer, pp. 211–217. 2

[LCS*18] LIU T. F., CRAFT M., SITU J., YUMER E., MECH R., KUMAR R.: Learning design semantics for mobile apps. In *The 31st Annual ACM Symposium on User Interface Software and Technology* (2018), ACM, pp. 569–579. 3, 5

[LGG18] LAGUNAS M., GARCES E., GUTIERREZ D.: Learning icons appearance similarity. *Multimedia Tools and Applications* (2018), 1–19. 3

[LGM20] LIAO Q. V., GRUEN D., MILLER S.: Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA, 2020), pp. 1–15. 2

[Lin94] LIN R.: A study of visual features for icon design. *Design Studies 15*, 2 (1994), 185–197. URL: https://www.sciencedirect.com/science/article/pii/0142694X94900248, doi:https://doi.org/10.1016/0142-694X(94)90024-8. 2

[LKC*16] LAURSEN L. F., KOYAMA Y., CHEN H.-T., GARCES E., GUTIERREZ D., HARPER R., IGARASHI T.: Icon set selection via human computation. In *Proceedings of the 24th Pacific Conference on Computer Graphics and Applications* (Okinawa Japan, 2016), Eurographics Association. 3

[LMG11] LEUNG R., MCGRENERE J., GRAF P.: Age-related differences in the initial usability of mobile device icons. *Behaviour & Information Technology 30*, 5 (2011), 629–642. 3

[LRFN04] LEWIS J. P., ROSENHOLTZ R., FONG N., NEUMANN U.: Visualids: Automatic distinctive icons for desktop interfaces. *ACM Trans. Graph. 23*, 3 (Aug. 2004), 416–423. URL: https://doi.org/10.1145/1015706.1015739, doi:10.1145/1015706.1015739. 2, 3

[LZC11] LEE Y. J., ZITNICK C. L., COHEN M. F.: Shadowdraw: Real-time user guidance for freehand drawing. *ACM Trans. Graph. 30*, 4 (July 2011), 27:1–27:10. URL: http://doi.acm.org/10.1145/2010324.1964922, doi:10.1145/2010324.1964922. 3

[MCCD13] Mikolov T., Chen K., Corrado G., Dean J.: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013). 5

[MCdB99] McDougall S. J., Curry M. B., de Bruijn O.: Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols. *Behavior Research Methods, Instruments, & Computers 31*, 3 (1999), 487–519. 2, 3, 6

[MCdB01] McDougall S. J., Curry M. B., de Bruijn O.: 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 (2001), 59–84. 1, 3, 4

[MHSG18] McInnes L., Healy J., Saul N., Grossberger L.: Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software 3*, 29 (2018), 861. 6

[MI09] McDougall S., Isherwood S.: What's in a name? the role of graphics, functions, and their interrelationships in icon identification. *Behavior research methods 41*, 2 (2009), 325–336. 2, 3

[MS95] Mullet K., Sano D.: *Designing visual interfaces: communication oriented techniques*. Prentice-Hall, Inc., USA, 1995. 2

[PWS*21] Patashnik O., Wu Z., Shechtman E., Cohen-Or D., Lischinski D.: Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Virtual Only, October 2021), pp. 2085–2094. 6

[PXW18] Peng M., Xing J., Wei L.-Y.: Autocomplete 3d sculpting. *ACM Trans. Graph. 37*, 4 (July 2018), 132:1–132:15. URL: http://doi.acm.org/10.1145/3197517.3201297, doi:10.1145/3197517.3201297. 3, 7

[RDF11] Rosenholtz R., Dorai A., Freeman R.: Do predictions of visual perception aid design? *ACM Transactions on Applied Perception (TAP) 8*, 2 (2011), 1–20. 1, 3

[RKH*21] Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., et al.: Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (Virtual Only, 2021), PMLR, pp. 8748–8763. 2, 6

[SABAG*05] Setlur V., Albrecht-Buehler C., A. Gooch A., Rossoff S., Gooch B.: Semanticons: Visual metaphors as file icons. In *Computer Graphics Forum* (2005), vol. 24, pp. 647–656. 1, 2

[SC21] Shen I.-C., Chen B.-Y.: Clipgen: A deep generative model for clipart vectorization and synthesis. *IEEE Transactions on Visualization and Computer Graphics 28*, 12 (2021), 4211–4224. 3

[SJ16] Silvennoinen J. M., Jokinen J. P.: Aesthetic appeal and visual usability in four icon design eras. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (San Jose, California, USA, 2016), pp. 4390–4400. 1, 2, 3

[SKJ17] Silvennoinen J. M., Kujala T., Jokinen J. P.: Semantic distance as a critical factor in icon design for in-car infotainment systems. *Applied ergonomics 65* (2017), 369–381. 2, 3

[SL19] Swearngin A., Li Y.: Modeling mobile interface tappability using crowdsourcing and deep learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK, 2019), pp. 1–11. 1, 3

[SLS*21] Shen I.-C., Liu K.-H., Su L.-W., Wu Y.-T., Chen B.-Y.: Clipflip : Multi-view clipart design. *Computer Graphics Forum* (2021). doi:10.1111/cgf.14190. 3

[SM14] Setlur V., Mackinlay J. D.: Automatic generation of semantic icon encodings for visualizations. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (Toronto, Ontario, Canada, 2014), pp. 541–550. 1, 2, 3, 4

[SRS18] Spiliotopoulos K., Rigou M., Sirmakessis S.: A comparative study of skeuomorphic and flat design from a ux perspective. *Multimodal Technologies and Interaction 2*, 2 (2018), 31. 2, 5

[SSII18] Simo-Serra E., Iizuka S., Ishikawa H.: Real-Time Data-Driven Interactive Rough Sketch Inking. vol. 37. 7

[SZL*21] Shen Z., Zhang L., Li R., Hou J., Liu C., Hu W.: The effects of color combinations, luminance contrast, and area ratio on icon visual search performance. *Displays 67* (2021), 101999. 2

[Tak01] Takagi H.: Interactive evolutionary computation: Fusion of the capabilities of ec optimization and human evaluation. *Proceedings of the IEEE 89*, 9 (2001), 1275–1296. 3

[Tra18] Tractinsky N.: The usability construct: a dead end? *Human–Computer Interaction 33*, 2 (2018), 131–177. 3

[Tra20] Tractinsky N.: The usability construct: a concern for both theory and practice. *Human–Computer Interaction 35*, 4 (2020), 338–353. 3

[UIM12] Umetani N., Igarashi T., Mitra N. J.: Guided exploration of physically valid shapes for furniture design. *ACM Trans. Graph. 31*, 4 (2012), 86–1. 3

[UKSI14] Umetani N., Koyama Y., Schmidt R., Igarashi T.: Pteromys: Interactive design and optimization of free-formed free-flight model airplanes. *ACM Trans. Graph. 33*, 4 (July 2014). URL: https://doi.org/10.1145/2601097.2601129, doi:10.1145/2601097.2601129. 3

[VdOLV18] van den Oord A., Li Y., Vinyals O.: Representation learning with contrastive predictive coding. *arXiv e-prints* (2018), arXiv–1807. 6

[VSP*17] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I.: Attention is all you need. In *Conference on Neural Information Processing Systems* (Long Beach, CA, USA, 2017). 6

[WMLB13] Warnock D., McGee-Lennon M., Brewster S.: Multiple notification modalities and older users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France, 2013), pp. 1091–1094. 2

[XCW14] Xing J., Chen H.-T., Wei L.-Y.: Autocomplete painting repetitions. *ACM Trans. Graph. 33*, 6 (Nov. 2014), 172:1–172:11. URL: http://doi.acm.org/10.1145/2661229.2661247, doi:10.1145/2661229.2661247. 3

[XKG*16] Xing J., Kazi R. H., Grossman T., Wei L.-Y., Stam J., Fitzmaurice G.: Energy-brushes: Interactive tools for illustrating stylized elemental dynamics. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan, 2016), UIST '16, Association for Computing Machinery, p. 755–766. URL: https://doi.org/10.1145/2984511.2984585, doi:10.1145/2984511.2984585. 7

[ZKH*20] Zhao N., Kim N. W., Herman L. M., Pfister H., Lau R. W., Echevarria J., Bylinskii Z.: Iconate: Automatic compound icon generation and ideation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA, 2020), pp. 1–13. 1, 2, 3