Deep Saliency Prior

Kfir Aberman* **Junfeng He*** **Yossi Gandelsman**
Inbar Mosseri **David E. Jacobs** **Kai Kohlhoff**
Yael Pritch **Michael Rubinstein**
Google Research

Abstract

We show that a saliency model trained to predict human eye-gaze can drive a range of powerful editing effects for reducing distraction in images, without any additional supervision. Given an image and a region to edit, we cast the problem of reducing distraction as an optimization over a composition of a differentiable image editing operator and a state-of-the-art saliency model. We demonstrate several operators, including: a recoloring operator, which applies a color transform that camouflages and blends distractors into their surroundings; a warping operator, which warps less salient image regions to cover distractors, gradually collapsing objects into themselves and effectively removing them (an effect akin to inpainting); a GAN operator, which uses a semantic prior to fully replace image regions with plausible, less salient alternatives. The resulting effects are consistent with cognitive research on the human visual system (e.g., since color mismatch is salient, the recoloring operator learns to harmonize objects' colors with their surrounding to reduce their saliency), and, importantly, are all achieved solely through the guidance of the pretrained saliency model, with no additional training data. We present results on a variety of natural images and conduct a perceptual study to evaluate and validate the changes in viewers' eye-gaze between the original images and our edited results.

1 Introduction

Studying and modeling human attention – how and where people look at images – has been widely researched and explored. In the deep learning era, saliency models trained on eye-gaze data are now able to predict human visual attention to high accuracy. However, while the research community has so far focused on developing models for *predicting* where people look, almost no attention has been given to utilizing the knowledge embedded in such recent, deep saliency models to actually *drive and direct* editing of images and videos so as to tweak the attention drawn to different regions in them. A few recent attempts [13, 31] have focused on subtle effects designed to make minimal modifications to the image, and are therefore limited in their ability to make meaningful changes to visual attention.

In this paper, we leverage deep saliency models to drive drastic, but still realistic, edits, which can significantly change an observer's attention to different regions in the image. Such capability can have important applications, for example in photography, where pictures we take often contain objects that distract from the main subject(s) we want to portray, or in video conferencing, where clutter in the background of a room or an office may distract from the main speaker participating in the call.

We ask: using a differentiable saliency model as a guide, what types of editing effects can be achieved? How would those effects affect viewers' attention in practice when looking at the images? Our focus in this paper is on *decreasing attention* for the purpose of reducing visual distraction, but we also demonstrate some results for *increasing* attention in Section 4 (Fig. 7).

*Denotes equal contribution

Preprint. Under review.

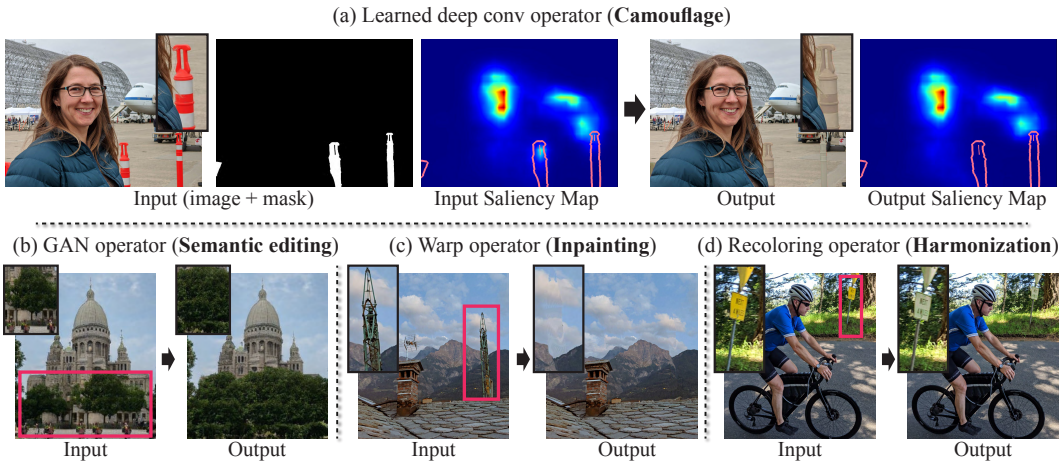


Figure 1: Given an input image and a mask of the region(s) to edit (top row, left), our method back-propagates through a visual saliency prediction model to solve for an image such that the saliency level in the region of interest is modified (top row, right). We explore a set of differentiable operators, the parameters of which are all guided by the saliency model, resulting in a variety of effects such as (a) camouflaging (b) semantic editing (c) inpainting, and (d) color harmonization.

To this end, we develop an optimization framework for guiding visual attention in images using a differentiable, predictive saliency model. Our method employs a state-of-the-art deep saliency model [20], pre-trained on large-scale saliency data [21]. Given an input image and a distractor mask, we backpropagate through the saliency model to parameterize an editing operator, such that the saliency within the masked region is reduced (Fig. 1). The space of appropriate operators in such a framework is, however, not unbounded. The problem lies in the saliency predictor—as with many deep learning models, the parametric space of saliency predictors is sparse and prone to failure if out-of-distribution samples are produced in unconstrained manner (Figure 2). Using a careful selection of operators and priors, we show that natural and realistic editing can be achieved via gradient descent on a single objective function.

We experiment with several differentiable operators: two standard image editing operations (whose parameters are learned through the saliency model), namely recolorization and image warping (shift); and two learned operators (we do not define the editing operation explicitly), namely a multi-layer convolution filter, and a generative model (GAN). With those operators, our framework is able to produce a variety of powerful effects, including recoloring, inpainting, detail/tone attenuation, camouflage, object editing or insertion, and facial attribute editing (Figure 1). Importantly, all these effects are driven solely by the single, pretrained saliency model, without any additional supervision or training. Note that our goal is not to compete with dedicated methods for producing each effect, but rather to demonstrate how multiple such editing operations can be guided by the knowledge embedded within deep saliency models, all within a single framework.

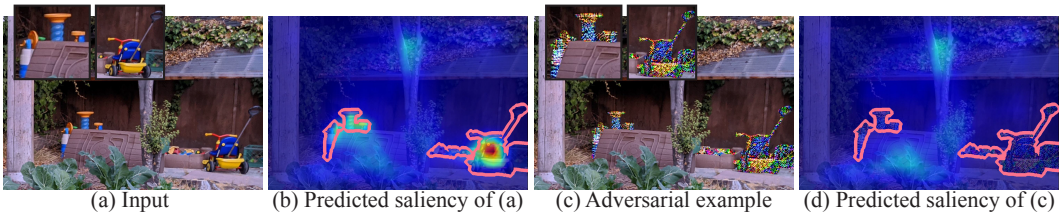
We demonstrate our approach on a variety of natural images, and conduct a perceptual study to validate the changes in real human eye-gaze between the original images and our edited results. Our experiments and user studies show that the produced image edits: a) effectively reduce the visual attention drawn to the specified regions, b) maintain well the overall realism of the images, and c) are significantly more preferred by users over more subtle saliency-driven editing effects that were proposed before.

2 Related Work

2.1 Human visual attention and saliency prediction models

Existing research on human visual attention has demonstrated that our attention is attracted to visually salient stimuli, i.e., a region sufficiently different from its surroundings, in terms of color, intensity, size, spatial frequency, orientation, shape, etc. [11, 18, 35, 36]. Moreover, studies were shown that human visual attention is drawn by particular objects like faces, texts [4], and emotion eliciting stimuli [2, 9], which are important for our survival.

Saliency prediction models aim at predicting which areas in an image will be salient to human attention and attract eye fixation. Early works in saliency prediction usually define saliency through



(a) Input (b) Predicted saliency of (a) (c) Adversarial example (d) Predicted saliency of (c)

Figure 2: An adversarial example of saliency models. Given an input image (a) with a predicted saliency (b), additive noise is applied to the image and optimized to reduce the saliency of image regions that were previously salient. However, the output (c) still exhibits salient regions which are interpreted as non-salient by the model (d).

a set of hand crafted features such as color difference. contrast, intensities, etc. [19, 25]. Recent works [17, 20, 26, 27] leverage the power of deep neural networks and are often trained/fine-tuned on large scale gaze data set [1, 21]. Deep neural networks based saliency prediction models often perform quite well, with predicted salient regions matching human gaze ground truths [20, 27]. A more thorough review on saliency prediction models can be found in [2, 11].

2.2 Saliency Driven Image Manipulation

Saliency prediction models have been applied to various applications like image/video compression [32], quality assessment [39], visualization [3], and image captioning [8]. Specifically, saliency prediction models are shown to be helpful for image editing tasks [14, 15, 37], e.g., to enhance contrast [15], improve aesthetics [37], and enhance details [14].

There are some early works [16, 29, 30] to use saliency models to guide human attention, however, they either do not use deep saliency models, or only use it as some extra input. Only recently, a few approaches [6, 13, 31] use deep saliency prediction models in the loss function with back propagation to help retarget visual attention. Gatys et al. [13] use a neural network that receives an image and a target saliency map, and generates an image satisfying that map. Chen et al. [6] use a similar architecture with an additional cyclic loss to stabilize the training procedure and reduce artifacts. Both of these approaches applies an adversarial and perceptual losses to the output, which strictly restricts its deviation from the original content of the region, resulting in a subtle and narrow effect. Recently, Mejjati et al. [31] proposed a neural network to predict a set of parameters that are applied to the image via a set of pre-defined operators, imitating the subtle changes that professional editors apply to images in order to retarget attention while maintaining fidelity to the original image. All of the previous approaches have been able to show only narrow and subtle effects, producing a single result with minimal changes to the visual saliency. In contrast, our approach utilizes the modeled perception of the saliency detector to a full extent, proposing a set of effects that are more dramatic and effective in guiding the visual attention. Moreover, unlike previous approaches that require large-scale datasets, our approach is simple and doesn't involve any training data, or training sessions, and is simple to tune. It requires a single saliency model that was pre-trained on high-quality eye-gaze tracking data.

3 Method

Given an image $\mathbf{I}$ and a region of interest $\mathbf{M}$, our objective is to manipulate the content of $\mathbf{I}$, such that the attention that this region draws is modified while keeping high-fidelity to the original image in other areas. Our approach is to follow the guidance of a saliency prediction model [20] [2] that was pretrained to identify attention grabbing regions based on saliency data [21]. Formally, we want to find an image $\tilde{\mathbf{I}}$ that solves the following two-term optimization problem

$$\arg\min_{\tilde{\mathbf{I}}} \ \mathcal{L}_{\text{sal}}\left(\tilde{\mathbf{I}}\right) + \beta \mathcal{L}_{\text{sim}}\left(\tilde{\mathbf{I}}\right), \tag{1}$$

where

$$\mathcal{L}_{\text{sal}}\left(\tilde{\mathbf{I}}\right) = \left\| \mathbf{M} \circ \left(S(\tilde{\mathbf{I}}) - \mathbf{T} \right) \right\|^2 \ \text{ and } \ \mathcal{L}_{\text{sim}}\left(\tilde{\mathbf{I}}\right) = \left\| (1 - \mathbf{M}) \circ \left(\tilde{\mathbf{I}} - \mathbf{I} \right) \right\|^2, \tag{2}$$

with a saliency model $S(\cdot)$ that predicts a spatial map (per-pixel value in the range of $[0, 1]$), and a target saliency map $\mathbf{T}$. $\|\cdot\|$ and $\circ$ represent the L_2 norm and the Hadamard product, respectively.

[2] For all the experiments in this paper we use the saliency prediction model of [20], with minor modifications that are described in the supplementary material.

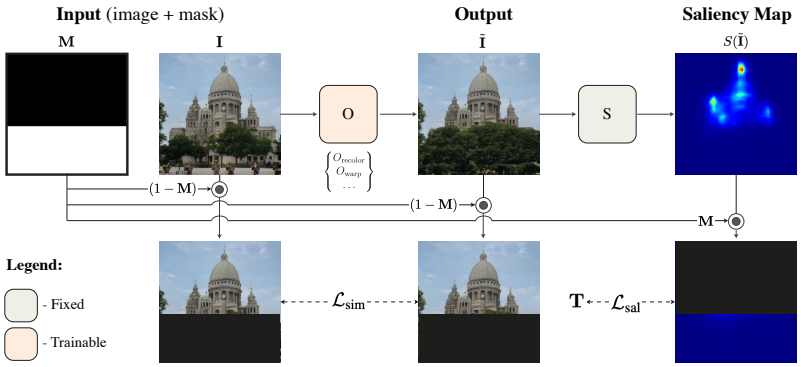


Figure 3: Our framework. Given an input image $\mathbf{I}$, a region of interest mask $\mathbf{M}$, and an operator $O \in \{O_{\text{recolor}}, O_{\text{warp}}, O_{\text{GAN}}, \dots\}$. Our approach generates an image with high-fidelity to the input image outside of the mask ($\mathcal{L}_{\text{sim}}$), and with reduced saliency inside it ($\mathcal{L}_{\text{sal}}$). The target saliency is typically selected to be $\mathbf{T} \equiv 0$.

We typically use $\mathbf{T} \equiv 0$ to reduce the saliency within the region of interest. However, $\mathbf{T}$ can be an arbitrary map, so saliency can be increased (e.g., by setting $\mathbf{T} \equiv 1$) or even controlled, as we show in a couple of examples in the paper and in the supplementary material.

Since existing saliency models are trained on natural images, a naive manipulation of the image pixels guided by (1) can easily converge into an "out-of-distribution" output. For instance, if additive noise is applied to the pixels within $\mathbf{M}$ and optimized with $\mathbf{T} \equiv 0$ the output may exhibit salient regions which are interpreted as non-salient by the model, as shown in Figure 2.

In order to avoid the vacant sub-spaces of the model, we constraint the solution space of $\tilde{\mathbf{I}}$ by substituting $\tilde{\mathbf{I}} = O_\theta(\mathbf{I})$ to (1), where O_θ is a pre-defined differentiable operator with a set of parameters θ that are used as our optimization variables. The constrained objective function can be written as

$$\arg\min_{\theta} \; \mathcal{L}_{\text{sal}}\left(O_\theta(\mathbf{I})\right) + \beta \mathcal{L}_{\text{sim}}\left(O_\theta(\mathbf{I})\right) + \gamma \Gamma(\theta), \tag{3}$$

where $\Gamma(\cdot)$ is a regularization function that is applied to θ, with weight γ.

Constraints imposed by using specific operators guarantee that manipulated images remain in the valid input domain of the saliency model where its predictive power is useful. We next show how different operators O_θ, yielding different effects, hand-crafted or learned, that comply with cognitive perception principles [11, 36].

Note that the results presented in the paper are achieved by a gradient decent optimization, however, the framework can be converted to a per-operator feed forward network, once trained on scale, as has been done in other domains, like in image style transfer [12, 22].

Recolorization - We first aim at solving a re-colorization task for our purpose, namely, maintaining the luminosity of the region of interest while modifying its chromatic values ('ab' components in the CIELab color representation) in order to reduce saliency. Here, O_θ is a re-color operator that applies a per-pixel affine transform on the 'ab' channels of the input image. The map is represented by a grid $\theta \in \mathbb{R}^{B \times B \times 6}$, that contains $B \times B$ affine transforms. We follow the idea behind Bilateral Guided Upsampling [5], and apply the map to the image in two differentiable steps: slice and apply. In the first step, we extract the affine transforms correspond to each pixel by querying the grid with the 'ab' value of the pixels. For example, a pixel with chromatic values (a, b), that lay in the (i, j)-th bin, slices θ

$$[\mathbf{A}, \mathbf{b}] = w_0(a,b)\theta(i,j) + w_1(a,b)\theta(i+1,j) + w_2(a,b)\theta(i,j+1) + w_3(a,b)\theta(i+1,j+1), \tag{4}$$

where $w_i(a,b),\ i \in \{0,1,2,3\}$ are bilinear weights that are dictated by the relative position of (a,b) within the bin, and $\mathbf{A} \in \mathbb{R}^{2\times 2}$ and $\mathbf{b} \in \mathbb{R}^2$ are the rotation and translation parts of the sliced affine transform, respectively. Then, the transformations are applied to the corresponding pixels $(a' \quad b') = (a \quad b)\,\mathbf{A} + \mathbf{b}$, where (a', b') are the modified chromatic values. In addition, to encourage color changes to be piecewise smooth, we add a smoothness term in the form of an isotropic total variation (TV) loss, $\Gamma(\theta) = \|\nabla_a \theta\|_1 + \|\nabla_b \theta\|_1$, where ∇_a and ∇_b represent the gradients of the grid with respect to the chroma axes a and b, respectively.

Warping - We next find a 2D warping field that modifies the saliency of the target region once applied. Here O_θ is a warp operator, with a sparse set of control points θ that are uniformly populated

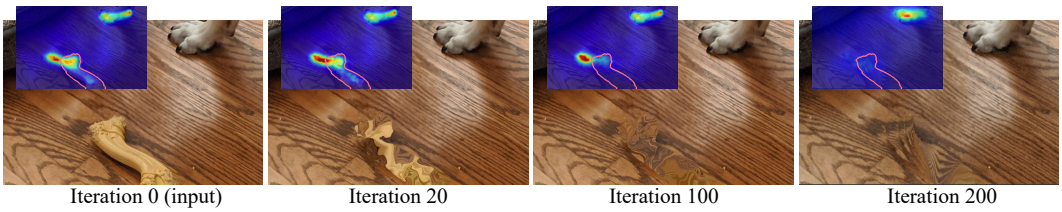


Figure 4: Saliency driven image warping. Our optimization framework gradually removes the distracting object by covering it with nearby pixels. Texture mismatch results in high saliency, thus, saliency model guides the warp operator towards a seamless completion of the region.

over the image grid. Each control point contains a 2D coordinate that indicates its displacement to the corresponding source pixel. The warp is accomplished in 2 steps. We first upsample the low-resolution grid θ to the full image size using bilinear interpolation to get the upsampled warp field $\mathbf{W}$, then we apply $\mathbf{W}$ to the source image to get $\tilde{\mathbf{I}} = \mathbf{W} \oslash \mathbf{I}$, were $\oslash$ is a warping operator that calculates the new value of each pixel by a bilinear interpolation, using the displacements specified in $\mathbf{W}$. For example, the new value of pixel (i, j) is calculated by

$$\tilde{\mathbf{I}}(\tilde{i}, \tilde{j}) = w_0(\tilde{i}, \tilde{j})\mathbf{I}(\tilde{i}, \tilde{j}) + w_1(\tilde{i}, \tilde{j})\mathbf{I}(\tilde{i}+1, \tilde{j}) + w_2(\tilde{i}, \tilde{j})\mathbf{I}(\tilde{i}, \tilde{j}+1) + w_3(\tilde{i}, \tilde{j})\mathbf{I}(\tilde{i}+1, \tilde{j}+1), \quad (5)$$

where $(\tilde{i}, \tilde{j}) = \lfloor \mathbf{W}(i, j) + (i, j) \rfloor$, and $w_i,\ i \in \{0, 1, 2, 3\}$ are bilinear weights, that are dictated by the relative position of $(\tilde{i}, \tilde{j})$ within the bin. Since this chain of operator is differentiable, we can back-propagate gradients through it to calculate the optimal warping field w.r.t (3). A similar smoothness term to the recolor operator is applied to the warping field. Our results demonstrate that the warp operation tends to remove the object, as it solves an image inpainting problem under unsupervised setting, namely, replacing the foreground object with a natural completion of the background with no explicit self-supervision. Unnatural completion of the background, or mismatch in texture, are interpreted as attention grabbing regions by the saliency model as can be seen in Figure 4.

Learning Convolutional Networks - We next use a deep convolutional neural network as our operator. The network consists of 5 convolution layers followed by non-linearity (ReLU), where θ represents the weights of the convolution kernels. Since deep networks may represent a large set of functions, the model can easily converge into an out-of-domain example. Thus, $\mathcal{L}_{\text{sim}}$ plays a crucial role in maintaining the solution in the valid region of the model. In the first 50 iterations the network weights are optimized to learn an identity mapping, then the saliency objective is applied. It can be seen that the network learns to camouflage prominent objects, and blend them with the background [7]. Another interesting insight is that the network selects to adapt colors of regions that are associated with the background, even in cases that there are multiple nearby regions which includes also foreground objects or subjects. Although the network is optimized on a single image (similarly to [33]), the saliency model that was trained on many examples refer background colors to lower saliency, and guides the network to transfer colors of background regions. To demonstrate this point, we calculate a naive baseline which adapts the colors of the surrounding pixels into the marked regions. The chromatic channels were replaced by the most dominant chromatic values of the surrounding pixels, and the brightness is translated such that its average is equal to the average brightness of the surrounding pixels. As can be seen in Figure 5, such a naive approach can not distinguish between foreground and background pixel values, while our method can by simply relying on the guidance of the saliency model.

StyleGAN as a Natural Image Prior - We can further constrain the solution space to the set of natural image patches that can fill the region of interest in a semantically-aware manner. Since this requirement is too general, we incorporate a domain specific pre-trained StyleGAN generator (e.g., human faces, towers, churches), that enables generation of high-quality images from a learned latent distribution, and define θ to be a latent vector in the $\mathcal{W}$ space [23].

Given an image $\mathbf{I}_{w_0} = G(w_o)$ that was generated by a generator G with a latent code $w_0 \in \mathcal{W}$, we initialize θ to be $\theta_0 = w_0$, and optimize it w.r.t (3). To avoid our-of-distribution solutions the output image is restricted to lay in the $\mathcal{W}$ space, by $\tilde{\mathbf{I}} = G(\theta)$. The optimization guides the latent code into meaningful directions that maintain the details of the image anywhere outside the region of interest, but modify the region's content in a semantically meaningful manner that affects the saliency. For example, in order to reduce the saliency of a structure that contains fine grained details (arcs, poles and windows), the saliency model guides the network to cover the structure by trees. In addition, the model can remove facial accessories such as glasses and to close the eyes of a person (Figure 6), which comply with cognitive perception principles [9].

(a) Input image (b) Background color adaptation (c) Our method

Figure 5: A comparison against a naive method for adaptation of background colors. (a) The input image, where we wish to reduce the saliency of the sign/post in the back. (b) The result when replacing the chromatic channels with the dominant chromatic values of the surrounding pixels + equalizing the average brightness level with the surrounding pixels by a translation. (c) Our result using the deep conv operator.

While increasing the saliency of a region is a less-constrained problem that can be solved in various ways with the aforementioned hand-crafted operators (e.g, 'recolor' can modify the colors of the region to be shiny and unnatural, and warp can lead to unnatural attention grabbing distortions), here, the dense latent space of StyleGAN contains a variety of meaningful directions that enable to also increase the saliency level of the region. For instance, the saliency model can guide the network to add facial details such as a moustache to increase the saliency in the mouth region, and also add prominent structure details to churches, as shown in figure 7.

We show semantic editing examples, that are applied to both purely generated images, and examples that were reconstructed from real images using GAN inversion techniques in the supplementary.

4 Results and Experiments

A gallery demonstrating our results with the different operators presented in Section 3 is depicted in Figure 6. More results can be found in the supplementary material. Note that the saliency model guide the operators to mitigate mismatch on color, intensity, texture (spatial frequency), shape, etc., between regions of interest and their surroundings, consistent with existing research on cognitive perception and human visual attention [11, 18, 35, 36].

In order to evaluate our method, we collected 120 images and asked professional photography editors to mark regions that draw attention away from the main subjects and reduce the overall user experience [10]. For the domain-specific GAN approach, we use images from the FFHQ dataset [23] and the LSUN dataset [38] for churches and towers. Our framework is implemented in TensorFlow and the parameters of the operators are optimized with the loss term in (3) using the Adam optimizer [24]. More detail can be found in the supplementary material.

We also demonstrate how our approach can be applied to video conference calls, aiming at reducing background clutter while maintaining the overall appearance of the room or the office. To apply our approach to videos, we segment the regions where the predicted saliency is above a threshold ($t = 0.15$). For each distracting region, we apply our different operators and select the one that yields the lowest saliency value within the region and apply the per-distractor parameters to the corresponding regions in all the frames. Figure 8 shows representative frames from the original video, a standard background blur effect, and our effect combined with background blur. It can be seen that our approach selects to inpaint some of the regions using a warp operator while other regions are camouflaged or recolorized. While background blur still includes dominant colorful blobs in the background that may distract from the main speaker, our approach further reduces the attention to the distracting regions while maintaining the overall "atmosphere" of the subject's environment.

Evaluating Changes in Eye-Gaze In order to evaluate the change in eye-gaze that our approach applies to images, we conducted a user study that tracks with high accuracy the eye fixation of 20 subjects, using the front camera of a smart phone and a dedicated app, as described in [34][3]. The subjects were asked to look at 31 images, one at a time, where each was presented for 5 seconds followed by a 1 second break. In order to ensure that their perception is unbiased, each subject was exposed either to the original image or its modified version, but not to both. We calculated the gaze saliency map of each image following the common procedure in gaze/saliency study [28].

[3]The gaze data was collected for research purposes only with participants' explicit consent. In addition, participants were allowed to opt out of the study at any point and request their data to be deleted.

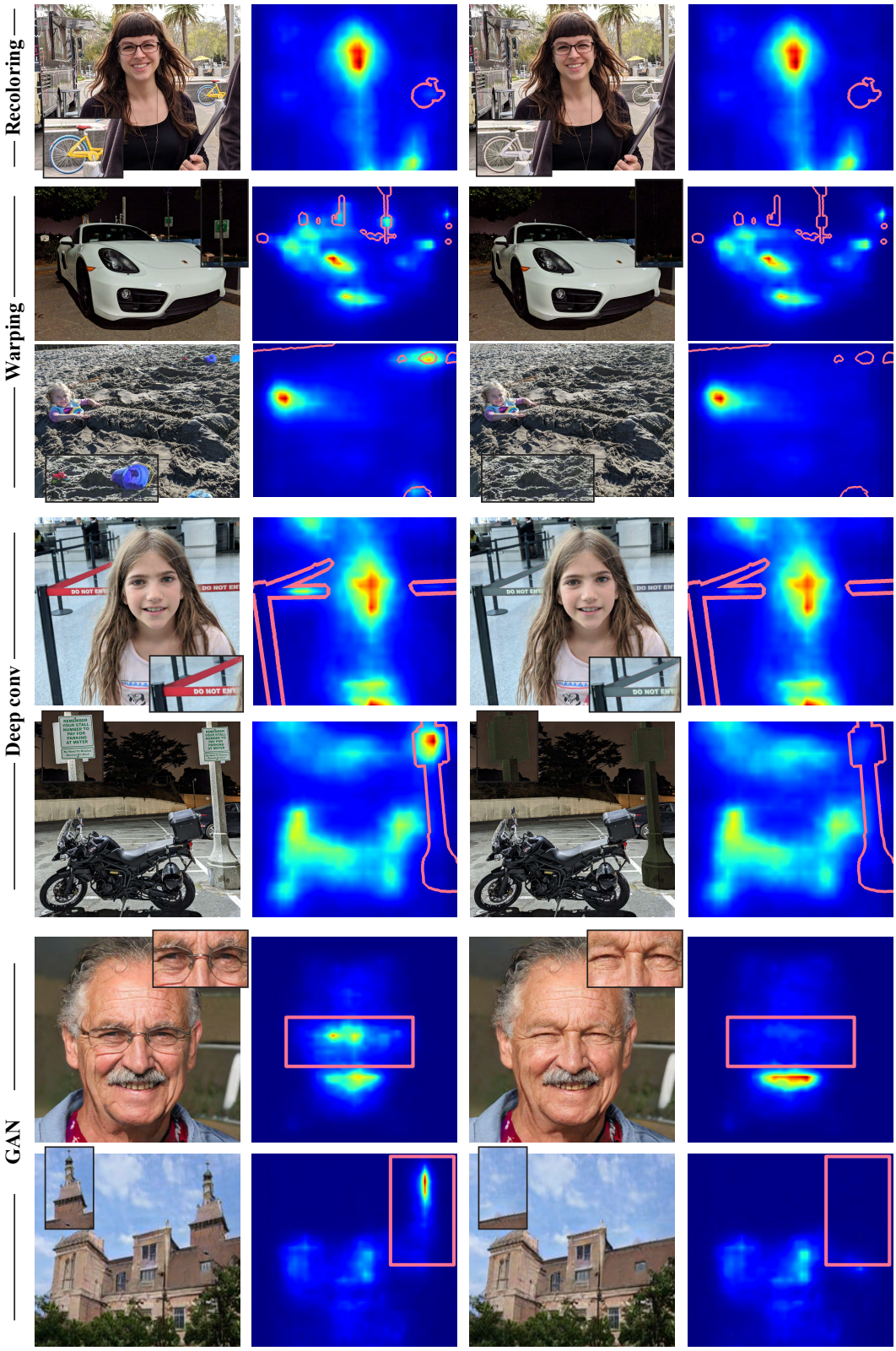
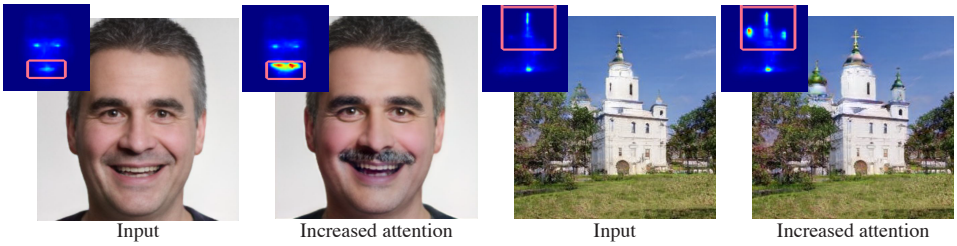
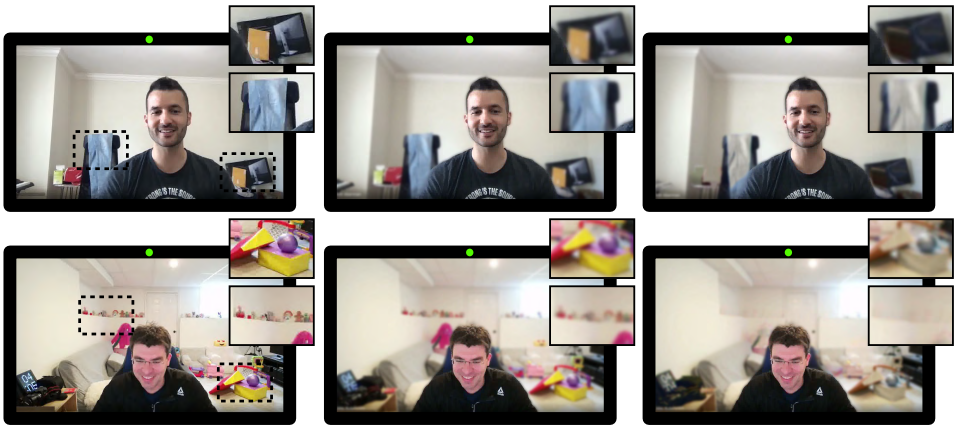


Figure 6: Additional results of reducing visual distractions, guided by the saliency model with several operators. The region of interest is marked on top of the saliency map (red border) in each example. More results are available in the supplementary material.



Input Increased attention Input Increased attention

Figure 7: Saliency *increase* by StyleGAN. For each image pair, the output image (right) was achieved by learning directions in the latent space, such that the saliency of the original image (left) is increased in the region of interest (marked in red on the corresponding saliency map). The found directions are semantically meaningful and natural (adding a moustache and adding prominent domes).



(a) Original Frame (b) Background Blur (c) Ours + background blur

Figure 8: Reducing distraction in a video conference call (a). Our approach + background blur (c) can reduce visual attention drawn to distracting regions, while maintaining the structural integrity of the subject's environment. Compare with the common background blur effect (b), which leaves colorful, attention-grabbing blobs in the background. See supplementary material for the full video.

Figure 9 depicts 2 examples (original and edited) and their average gaze map. It can be seen that the subjects' gaze saliency is reduced in the selected regions (red box), as expected by our approach. In addition, we compute the mean saliency value within the region, and calculate its average across all the images under each operator. The average reduction, $|\mathbf{M}(S_g(\tilde{\mathbf{I}}) - S_g(\mathbf{I}))|/|\mathbf{M}S_g(\mathbf{I})|$ where S_g is gaze saliency, (per-effect) is reported in Table 1 (a). Evidently, our effects successfully reduce the average saliency after the manipulation, demonstrating that our approach guides human attention as expected. To ensure that our data is statistically significant we ran a statistical test for three features: (i) gaze saliency within the mask, (ii) consecutive gaze duration within the mask, and (iii) first time gaze stays within the mask for more than 50 ms. We computed the average value (across participants) of each feature and ran a paired samples T-Test (original and edited images as control group and observation group, respectively). The results are shown in Table 1 (b). In all cases, p-value is < 0.003, implying statistical significance of gaze saliency/duration reduction and first gaze time increase.

Realism Modifying image saliency does not guarantee that the output image is visually plausible, or seem realistic. Hence, we asked 32 users to tell whether a given image looks natural to them. Each user saw 16 images from our dataset, where 4 of them are original and 12 are edited. 85% of the users marked the original images as realistic, while 78% of them marked the same answer to our outputs. The correlation between the numbers implies that our method preserves realism as seem by the users.

Comparison to state-of-the art In order to understand what kind of effects users prefer for the task of saliency reduction, we conducted a study with 32 participants. The users were asked to look at 16 images with a marked region of interest, together with two outputs, ours (various effects) and

Recolor	Warp	ConvNet	GAN
43.1%	92.9%	53.3%	34.8%

(a) Gaze saliency reduction

Gaze saliency	Duration	First gaze
t=-4.5 p=6×10^{-5}	t=-3.3 p=2.2×10^{-5}	t=3.5 p=1.2×10^{-3}

(b) Paired Samples T-Test

Table 1: Perceptual study using real gaze - analysis. (a) Reduction of average gaze saliency within the region of interest, per-effect. (b) Paired Samples T-Test (edited images vs. original images) , p-value < 0.003, demonstrating statistical significance.

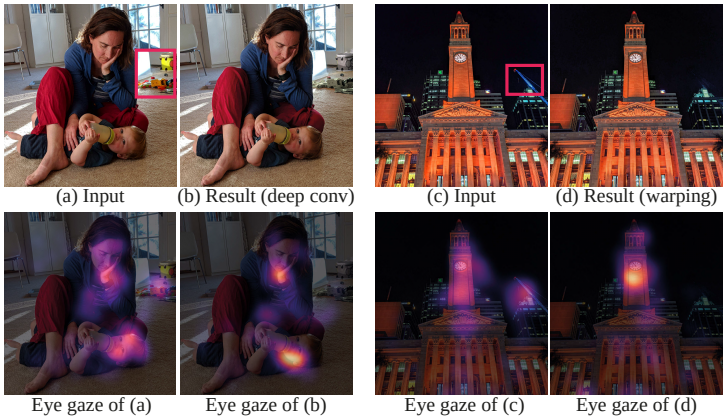


Figure 9: Samples of real eye-gaze saliency maps measured in our perceptual study, involving 20 subjects and 31 images. Each pair in the first row show an original image (left) with a region of interest on top (red border) and our result (right). The second row depicts the corresponding average eye-gaze maps across participants in the study.

Figure 10: Qualitative comparison with "look-here!" [31]. More in the supplementary material. In Table 2 we compare numerically with [31] the effective change to the saliency maps and the users' preferences as found in our user study.

"look-here!", and were asked: "The following two results attempt to draw LESS attention to the region marked in red on the original image. Which one do you like better?". Table 2(b) reports the breakdown of user selections between our method and "look-here!" [31]. Our results received clear preference for each of the effects, indicating that users in general preferred more aggressive effects to more subtle ones for the purpose of removing distractions.
Figure 10 compares our effects visually to "look-here!" (more in the supplementary material), and Table 2(a) reports the percentage of saliency reduction (comparing to the original image) for each of our effects and "look-here!". Our method enables a larger reduction in saliency compared to "look-here!", as expected from the more dramatic effects we design it for.

5 Conclusion

We introduced a novel framework that utilizes the power of a saliency model trained to predict human eye-gaze, to guide a range of editing effects (e.g., recoloring, inpainting, camouflage, semantic object and attribute editing) that result in meaningful changes to visual attention in images. This is done without any additional training data or direct supervision for the specific editing tasks. In contrast to other methods that use saliency to drive editing which produce subtle changes to image appearance, our method results in drastic, yet realistic edits. We have validated that our results indeed achieve the desired effects on observer's attention by analysing the changes in real human eye-gaze between the original images and our edited results.

Recolor	Warp	ConvNet	Look-here
43.1%	92.9%	53.3%	25.8%

(a)

"preferred method"	Recolor	Warp	ConvNet
Look-here	31.3%	9.4%	18.8%
Ours	62.5%	84.4	75%
"Roughly similar"	6.3%	6.3%	6.3%

(b)

Table 2: Comparison of our effects to "look-here!" [31]. (a) Reduction of average predicted saliency. (b) User study results. We show representative qualitative comparisons to [31] in Fig. 10, and more are available in the supplementary material.

References

[1] Mit saliency benchmark. *http://saliency.mit.edu*.

[2] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[3] Zoya Bylinskii, Nam Wook Kim, Peter O'Donovan, Sami Alsheikh, Spandan Madan, Hanspeter Pfister, Fredo Durand, Bryan Russell, and Aaron 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.

[4] Moran Cerf, E Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12):10–10, 2009.

[5] Jiawen Chen, Andrew Adams, Neal Wadhwa, and Samuel W Hasinoff. Bilateral guided upsampling. *ACM Transactions on Graphics (TOG)*, 35(6):1–8, 2016.

[6] Yen-Chung Chen, Keng-Jui Chang, Yu Chiang Frank Wang, Yi-Hsuan Tsai, and Wei-Chen Chiu. Guide your eyes: Learning image manipulation under saliency guidance. In *30th British Machine Vision Conference, BMVC 2019*, 2019.

[7] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM Trans. Graph.*, 29(4):51–1, 2010.

[8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2):1–21, 2018.

[9] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 7521–7531, 2018.

[10] Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. Finding distractors in images. In *Proceedings of the IEEE Conference on Computer Vision and pattern Recognition*, pages 1703–1712, 2015.

[11] Simone Frintrop, Erich Rome, and Henrik I Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):1–39, 2010.

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[13] Leon A Gatys, Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Guiding human gaze with convolutional neural networks. *arXiv preprint arXiv:1712.06492*, 2017.

[14] Sanjay Ghosh, Ruturaj G Gavaskar, and Kunal N Chaudhury. Saliency guided image detail enhancement. In *2019 National Conference on Communications (NCC)*, pages 1–6. IEEE, 2019.

[15] Ke Gu, Guangtao Zhai, Xiaokang Yang, Wenjun Zhang, and Chang Wen Chen. Automatic contrast enhancement technology with saliency preservation. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(9):1480–1494, 2014.

[16] Aiko Hagiwara, Akihiro Sugimoto, and Kazuhiko Kawamoto. Saliency-based image editing for guiding visual attention. In *Proceedings of the 1st international workshop on pervasive eye tracking & mobile eye-based interaction*, pages 43–48, 2011.

[17] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015.

[18] Laurent Itti. Visual salience. *doi:10.4249/scholarpedia.3327*, 2007.

[19] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[20] Sen Jia and Neil DB Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020.

[21] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[26] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

[27] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798, 2017.

[28] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013.

[29] Victor A Mateescu and Ivan V Bajić. Attention retargeting by color manipulation in images. In *Proceedings of the 1st International Workshop on Perception Inspired Video Processing*, pages 15–20, 2014.

[30] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Saliency driven image manipulation. *Machine Vision and Applications*, 30(2):189–202, 2019.

[31] Youssef A Mejjati, Celso F Gomez, Kwang In Kim, Eli Shechtman, and Zoya Bylinskii. Look here! a parametric learning based approach to redirect visual attention. In *European Conference on Computer Vision*, pages 343–361. Springer, 2020.

[32] Yash Patel, Srikar Appalaraju, and R Manmatha. Saliency driven perceptual image compression. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 227–236, 2021.

[33] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[34] Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature communications*, 11(1):1–12, 2020.

[35] Jeremy M Wolfe and Todd S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, 5(6):495–501, 2004.

[36] Jeremy M Wolfe and Todd S Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):1–8, 2017.

[37] Lai-Kuan Wong and Kok-Lim Low. Saliency retargeting: An approach to enhance image aesthetics. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 73–80. IEEE, 2011.

[38] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[39] Wei Zhang, Ali Borji, Zhou Wang, Patrick Le Callet, and Hantao Liu. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE transactions on neural networks and learning systems*, 27(6):1266–1278, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]

(c) Did you discuss any potential negative societal impacts of your work? [Yes]
Ethical Considerations. Our technology focuses on world-positive use cases and applications. Guiding visual attention in images through saliency models has a variety of beneficial and impactful uses, such as removing distractors from photos and video calls, or calling attention to specific areas of a poster or sign to improve the readability and understanding of its content, to name a few. However, we acknowledge the potential for misuse, given the use of generative models to edit images. We emphasize the importance of acting responsibly and taking ownership of synthesized content. To that end, we strive to take special care when sharing images or other material that has been synthesized or modified using these techniques, by clearly indicating the nature and intent of the edits. Finally, we also believe it is imperative to be thoughtful and ethical about the content being generated. We follow these guiding principles in our work.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Data and detailed instructions to reproduce our experiment results are in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] Our method does not need training.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Yes, we have run statistics test.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] There is no training needed, and the total compute amount is minimal.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The Salicon data, FFHQ, and LSUN data set used in our paper are all cited.
   (b) Did you mention the license of the assets? [Yes] The license is mentioned in the supplemental material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] It is discussed in the supplemental material.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Face images in FFHQ contains personally identifiable information.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Included in the supplemental material.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] The study is conducted with volunteers from our colleagues.