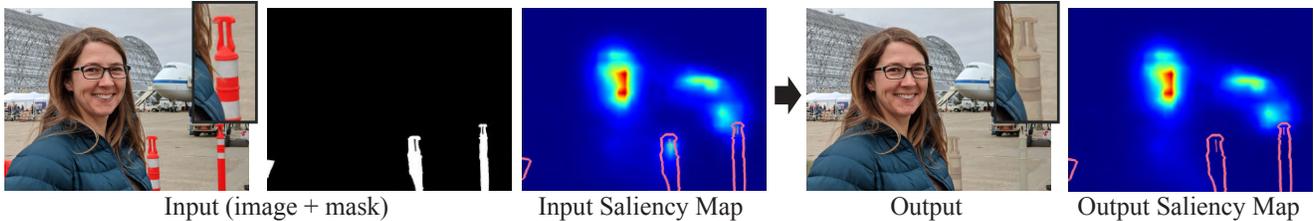


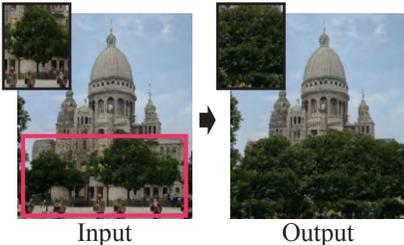
Deep Saliency Prior for Reducing Visual Distraction

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

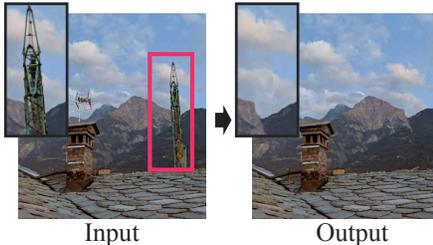
(a) Learned deep conv operator (**Camouflage**)



(b) GAN operator (**Semantic editing**)



(c) Warp operator (**Inpainting**)



(d) Recoloring operator (**Harmonization**)

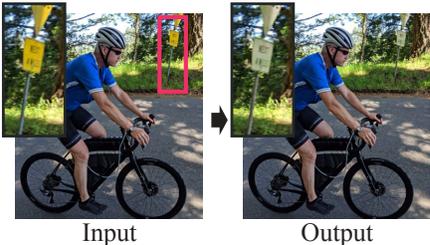


Figure 1. Given an input image and a mask of the region(s) to edit (top row, left), our method back-propagates through a visual saliency prediction model to solve for an image such that the saliency level in the region of interest is modified (top row, right). We explore a set of differentiable operators, 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.

Abstract

Using only a model that was trained to predict where people look at images, and no additional training data, we can produce a range of powerful editing effects for reducing distraction in images. Given an image and a mask specifying the region to edit, we backpropagate through a state-of-the-art saliency model to parameterize a differentiable editing operator, such that the saliency within the masked region is reduced. We demonstrate several operators, including: a recoloring operator, which learns to apply 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. 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. Project Webpage: <https://deep-saliency-prior.github.io/>

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 [15, 34] 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 dramatic, but still realistic, edits, which can significantly change an observer’s attention to different regions in an 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 drawn to image regions in Section 4 (Fig. 6).

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 [22], pre-trained on large-scale saliency data [24]. Given an input image and a distractor mask, we backpropagate through the saliency model – effectively using it as a *prior* – 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, 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 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

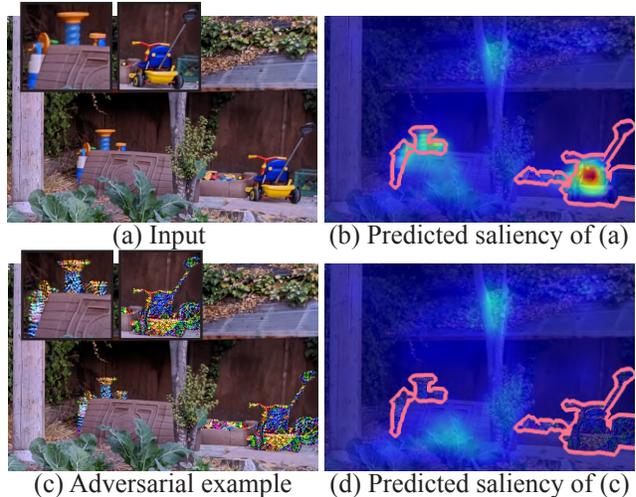


Figure 2. An adversarial example of saliency models. Given an input image (a) with a predicted saliency (b), additive noise is applied to the image and optimized to reduce the saliency of image regions that were previously salient. However, the output (c) still exhibits salient regions which are interpreted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

preferred by users over more subtle saliency-driven editing effects that were proposed before.

2. Related Work

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. [12, 20, 42, 43]. Moreover, studies were shown that human visual attention is drawn by particular objects like faces, texts [5], and emotion eliciting stimuli [1, 10], which are important for our survival.

Saliency prediction models [19, 21, 22, 28–30] aim at predicting which areas in an image are salient to human attention. Recent works [19, 22, 29, 30, 36] leverage the power of deep neural networks and are often trained/fine-tuned on large scale gaze data sets [2, 24]. A more thorough review on saliency prediction models can be found in [1, 12].

Saliency Driven Image Manipulation Saliency prediction models have been applied to various applications such as image/video compression [35], quality assessment [47], visualization [4], and image captioning [9]. Specifically, saliency models are shown to be helpful for image editing tasks [16, 17, 44], e.g., to enhance contrast [17], improve aesthetics [44], and enhance details [16].

There are some early works on using saliency models to guide human attention [18, 32, 33], however, they either do not use deep saliency models, or only use it as an extra input. Only recently, a few approaches [7, 15, 34] suggested using deep saliency prediction models in the loss function with

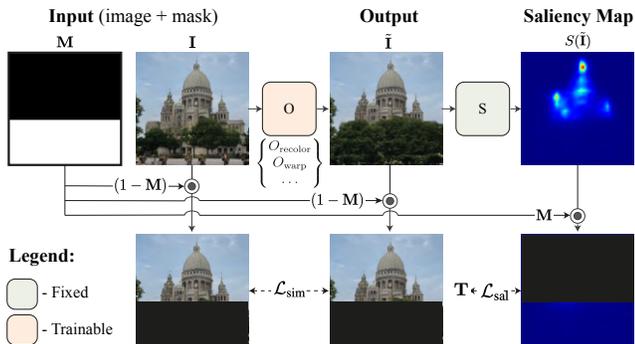


Figure 3. Our framework. Given an input image I , a region of interest mask M , and an operator $O \in \{O_{\text{recolor}}, O_{\text{warp}}, O_{\text{GAN}}, \dots\}$. Our approach generates an image with high-fidelity to the input image outside of the mask (\mathcal{L}_{sim}), and with reduced saliency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

back propagation to help retarget visual attention. Gatys et al. [15] and Chen et al. [7] use a neural network that receives an image and a target saliency map, and generates an image satisfying that map. However, both methods strictly restricts its deviation from the original content of the region, resulting in a subtle and narrow effect.

Recently, Jiang et al. [23] proposed a GAN-based image translation method to manipulate saliency via object removal and insertion. Mejjati et al. [34] proposed a neural network to predict a set of parameters that are applied to the image via 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. While their approach intentionally aims to only apply subtle changes to the image, our output effects are more diverse and dramatic in the form of inpainting, camouflage and semantic manipulation effects, making a more significant effect on the viewers’ attention. In addition, while both [23] and [34] requires a large dataset to train their network, our approach works in a zero-shot setting, namely, does not require additional data.

3. Method

Given an input image I and a region of interest M , our objective is to manipulate the content of I such that the attention drawn to region M 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 [22]¹ that was pretrained to identify attention grabbing regions based on saliency data [24]. Formally, we seek to find an image \tilde{I} that solves the following optimization problem:

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

where

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

¹We use the saliency prediction model of [22], with minor modifications described in the SM pdf.

and

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

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.

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 set to specific values in the range $[0, 1]$, as we show in examples in the paper and supplementary material (SM).

Since existing saliency models are trained on natural images, a naive manipulation of the image pixels guided by Eq. (1) can easily converge into ”out-of-distribution” outputs. For instance, if additive noise is applied to the pixels within 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 prevent convergence into the vacant regions of the saliency model, we constrain the solution space of \tilde{I} by substituting $\tilde{I} = O_{\theta}(I)$ in Eq. (1), where O_{θ} is a pre-defined differentiable operator with a set of parameters θ that are used as the optimization variables. The constrained objective function can be written as

$$\arg \min_{\theta} \mathcal{L}_{\text{sal}}(O_{\theta}(I)) + \beta \mathcal{L}_{\text{sim}}(O_{\theta}(I), I) + \gamma \Gamma(\theta), \quad (2)$$

where $\Gamma(\cdot)$ is a regularization, with weight γ .

Constraints imposed by using specific operators guarantee that the manipulated images remain within the valid input domain of the saliency model where its predictive power is useful. We next show how different operators O_{θ} can yield different effects, hand-crafted or learned, that comply with cognitive perception principles [12, 43].

Note that the results presented in the paper are achieved by a gradient descent optimization, however, the framework can be converted to a per-operator feed forward network, once trained on scale, as done in other domains such as image style transfer [14, 25].

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 CIE Lab color representation) in order to reduce saliency. Here, O_{θ} is a recolor 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. Following the idea of Bilateral Guided Upsampling [6], we apply the map to the image in two differentiable steps. 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 lies in the (i, j) -th bin, yields the following affine transform

$$\mathbf{T}_{(a,b)} = w_0(a, b)\theta(i, j) + w_1(a, b)\theta(i + 1, j) + w_2(a, b)\theta(i, j + 1) + w_3(a, b)\theta(i + 1, j + 1), \quad (3)$$

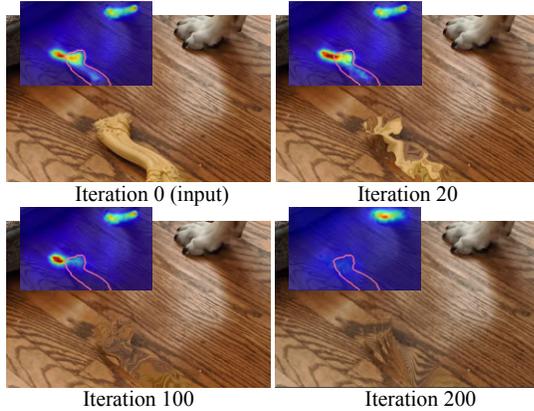


Figure 4. Saliency driven image warping. Our optimization framework gradually removes the distracting object by covering it with nearby pixels. Texture mismatch results in high saliency, thus, the saliency model guides the warp operator towards a seamless completion of the region.

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 $\mathbf{T}_{(a,b)} \in \mathbb{R}^6$ is a vector that can be reshaped into the rotation $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ and translation $\mathbf{b} \in \mathbb{R}^2$ parts of the affine transform. The extracted transformation is applied to the pixel via $(a' \ b') = (a \ b) \mathbf{A} + \mathbf{b}$, where (a', b') are the output chromatic values. 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 to the image. Here O_θ is a warp operator, represented by a sparse set of control points θ that are uniformly populated over the image grid. Each control point contains a 2D coordinate indicating its displacement to the corresponding source pixel. The warp is accomplished by upsampling the low-resolution grid θ to full image size (bilinear interpolation) to get the upsampled warp field \mathbf{W} , then we apply \mathbf{W} to the source image. The output value of each pixel is computed by

$$\tilde{\mathbf{I}}(\tilde{i}, \tilde{j}) = w_0(\tilde{i}, \tilde{j})\mathbf{I}(\tilde{i}, \tilde{j}) + w_1(\tilde{i}, \tilde{j})\mathbf{I}(\tilde{i} + 1, \tilde{j}) + w_2(\tilde{i}, \tilde{j})\mathbf{I}(\tilde{i}, \tilde{j} + 1) + w_3(\tilde{i}, \tilde{j})\mathbf{I}(\tilde{i} + 1, \tilde{j} + 1), \quad (4)$$

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. Due to the differentiability of the operators, the gradients can be backpropagated through this chain, enabling calculation of the optimal warping field w.r.t (2). In addition, in order to enable better propagation of pixels warped from the exterior region into the interior region of the mask, in each iteration the input image is updated by the warped image $\tilde{\mathbf{I}} \rightarrow \mathbf{I}$. A similar smoothness term to the one added to the recolor operator is applied to the warping field. Our results demonstrate that the warp operator tends to remove objects, as it solves an image inpainting problem under unsupervised setting, namely, replacing the foreground object with a natural completion of the background with no

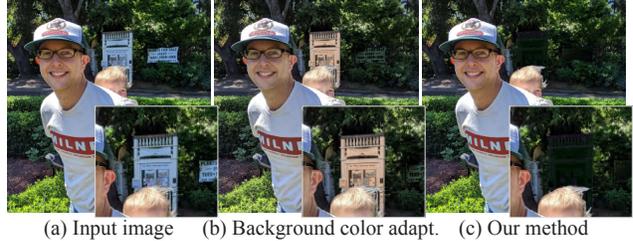


Figure 5. Comparison with a naive method for adapting background colors. (a) The input image, where we wish to reduce the saliency of the sign/post in the back. (b) The result when replacing the chromatic channels with the dominant chromatic values of surrounding pixels (equalizing the average brightness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

explicit self-supervision. Unnatural completion of the background, or mismatch in texture, are interpreted as attention grabbing regions by the saliency model (Figure 4).

Learning Convolutional Networks We use an untrained deep convolutional neural network as an image-to-image 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}_{sim} plays a key role in maintaining the solution in the valid region of the model. In the first tens of iterations the network weights are optimized to only reconstruct the original image (identity mapping), then the saliency objective is added. It can be seen that the network learns to camouflage prominent objects, and blend them with the background [8]. Another interesting insight is that the network selects to adapt colors of regions that are associated with the background, even when multiple regions are presented nearby the region of interest (including foreground objects or subjects). Although the network is optimized on a single image (similarly to [13, 40]), the saliency model that was trained on many examples prefers 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 (e.g., human faces, towers, churches) pre-trained StyleGAN generator space [26], 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 [26]. Similarly to previous approaches we edit the image in the

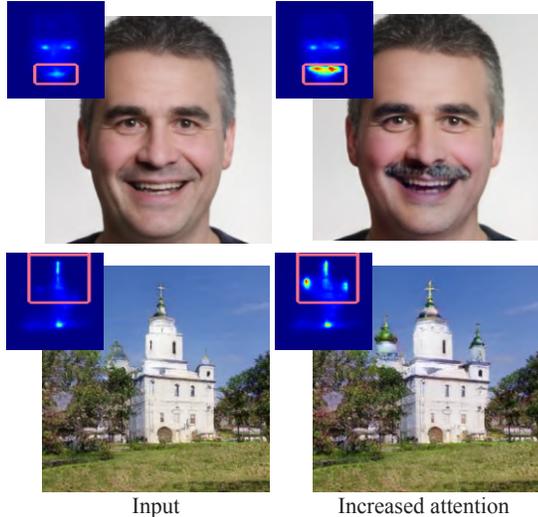


Figure 6. Saliency *increase* by StyleGAN. For each image pair, the output image (right) was achieved by learning directions in the latent space such that the saliency of the original image (left) is increased in the region of interest (marked in red on the saliency maps). The found directions are semantically meaningful and natural (adding a moustache and prominent domes).

latent space [38], but in our case, the editing is driven by the saliency model.

Given an image $I_{w_0} = G(w_0)$ 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 (2). To avoid out-of-distribution results the output image is restricted to lay in the \mathcal{W} space, such that $\tilde{I} = G(\theta)$. The optimization guides the latent code into 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 (Fig. 7), which comply with cognitive perception principles [10].

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 result in saliency increase. 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 geometric structures such as domes to churches (Fig. 6).

We show semantic editing examples, that are applied to both purely generated images, and examples reconstructed from real images using GAN inversion techniques [45] in the SM pdf (Sec. 2.4) and html (Sec. 2, 3).

4. Results and Experiments

A gallery demonstrating our results with different operators in Sec. 3 is shown in Fig. 7. More results can be found in the SM html (Sec. 2). Note that the saliency model guides the operators to mitigate mismatch in 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 [12, 20, 42, 43].

To evaluate our method, we collected 800 images and asked professional photographers to mark regions that draw attention away from the main subjects and reduce the visual experience [11]. The regions were marked by a bounding box and then an instance segmentation module [39] was used to extract a mask. To further clean the data, 15% of the masks were fine-tuned manually. For the domain-specific GAN approach, we use images from the FFHQ dataset [26] for faces and the LSUN dataset [46] for churches and towers. Our framework is implemented in TensorFlow and the parameters of the operators are optimized with the loss term in (2) using the Adam optimizer [27]. More detail about the hyper-parameters can be found in the SM pdf (Sec. 2.3).

For all the results in the paper we use a variant of EML-Net [22] as the guiding saliency model, which is extensively evaluated and considered state-of-the-art [2, 22, 36]. However, our framework is not limited to a specific model and any differential saliency model can fit into our pipeline. In the SM html (Sec. 7), we show results that were driven by a different saliency model [36]. Another small nuance is that EML-Net (as are most saliency models) is trained and evaluated on *natural* images, whereas we use it for providing saliency predictions also on *edited* content. For reassurance, we also ran extra experiments to evaluate the saliency model’s accuracy in predicting attention on our edited images, showing little to no change (compared to the accuracy on natural, unedited images) in standard saliency evaluation metrics: AUC-Judd, NSS, SIM and KLD [3]. Those experiments and results too are given in detail in the SM pdf (Sec. 2.5), for the interested reader.

We also demonstrate how our approach can be applied to video conference calls, aiming at reducing background clutter that may distract from the main speaker. To apply our approach to videos, we manually segment the regions where the predicted saliency is above a threshold ($t = 0.15$) in a single frame (assuming static background throughout the video). For each distracting region, we apply our different operators and automatically 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. The video in the SM shows representative the original video, a standard background blur effect, and our effect combined with background blur. 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, our approach further reduces distracting regions

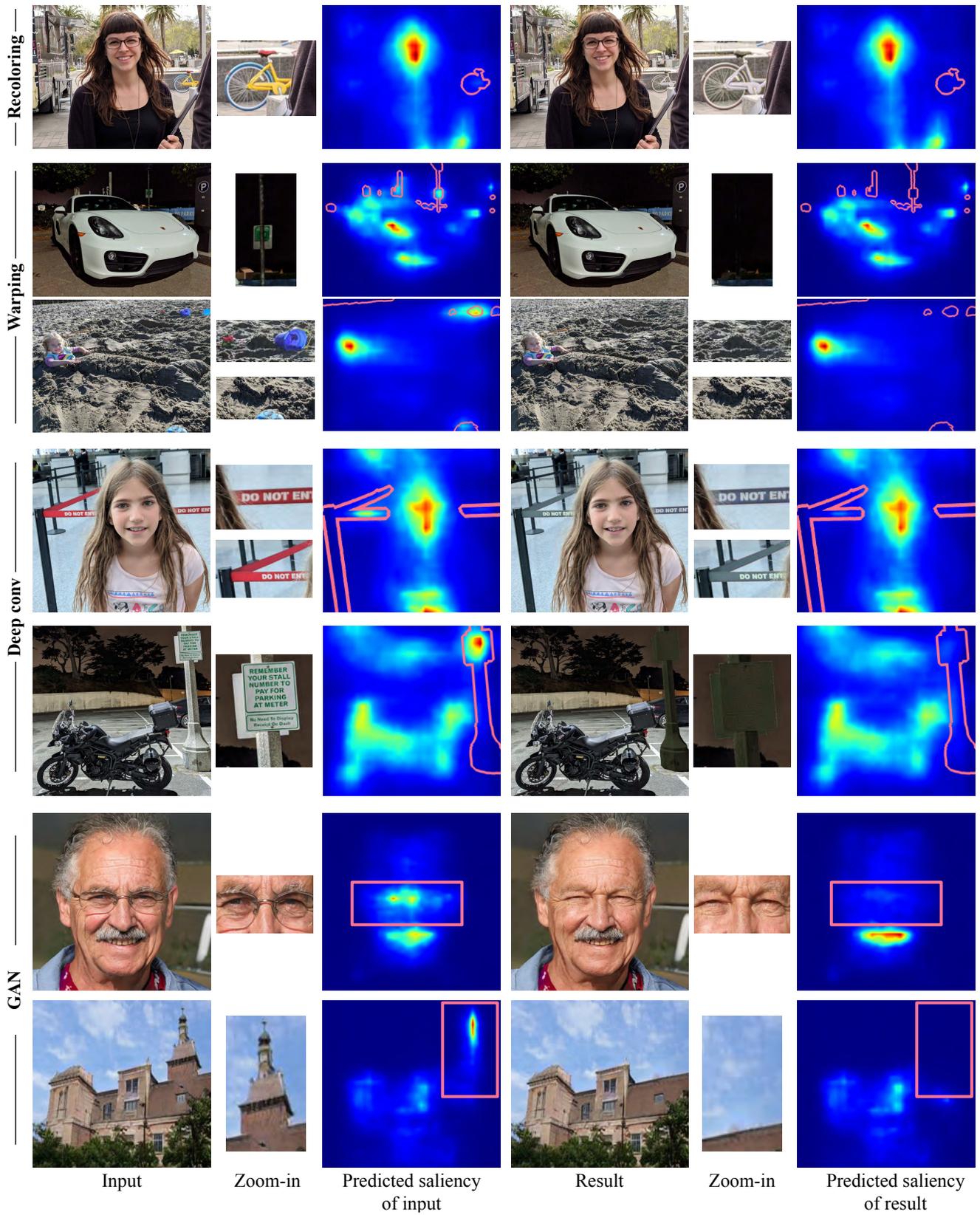


Figure 7. Additional results of reducing visual distractions, guided by the saliency model with several operators. The region of interest is marked on top of the saliency map (red border) in each example. More results are available in the SM html (Sec. 2).

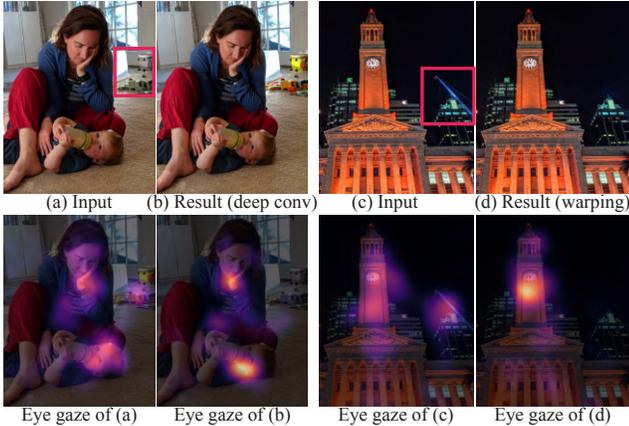


Figure 8. Examples of real eye-gaze saliency maps measured in our perceptual study, involving 20 subjects and 31 images. Top row: each pair shows an original image (left) with a region of interest (red border) and our result (right). Bottom row: the corresponding average eye-gaze maps across participants in the study.

while maintaining the overall “atmosphere” of the subject’s environment.

Evaluating Changes in Eye-Gaze 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 [41]. 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.² 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 [31]. Fig. 8 shows two examples (original and edited) and their average gaze maps. It can be seen that the subjects’ gaze saliency is reduced within the selected regions (red box) 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. We further show the changes of two other gaze metrics in Table 1 (b): consecutive gaze duration within the mask, and first time gaze intersects the masked region. These metrics show that after editing users spend less time looking at the distracting regions, and it takes them longer to notice the distractor. In the SM pdf (Sec. 3), we show that the change of each gaze metric is statistically significant using a paired samples T-Test.

²Our study obtained approval from an oversight panel within our organization, following strict institutional policies. All participants provided explicit and informed consent to participate in the study, and could opt out of the study at any time (with data wiped out) without any penalty.

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

(a) Average gaze saliency reduction within masked regions.

| | Duration (ms) | First gaze (ms) | Gaze saliency |
|----------|---------------|-----------------|---------------|
| Original | 215.5 | 4111.2 | 0.2 |
| Edited | 116.3 (−46%) | 4502.4 (+9.5%) | 0.08 (−60%) |

(b) Change in gaze metrics between original and edited images.

Table 1. Gaze metrics extracted from perceptual study using real eye gaze tracking.

Realism Modifying image saliency does not guarantee that the output image is realistic. Hence, we asked 32 users to tell whether a given image looks natural to them. Each user saw 16 arbitrary images, 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 our outputs as realistic, implying that our method preserves realism well.

Comparison to state-of-the art We compare our method to previous attention retargeting approaches (WSR [44], SDIM [33] and “look-here!” [34]) using the distractor attenuation dataset of Mechrez et al. [37]. Each method aims at attention retargeting with different restricted properties (e.g., Mechrez et al. [33] is limited to reusing colors and textures from the same image, and “look-here!” [34] tries to maintain high-fidelity to the original image), so a direct side-by-side comparison is not straightforward. To make the comparison fair, we selected our deep conv operator which is optimized to also maintain similarity within the mask as explained in Section 3. Table 2 summarizes the average saliency drop within the masked region of each approach and Figure 9 depicts a few representative results. Our results demonstrate a more significant saliency decrease both qualitatively (color and texture are blended better with the background) and quantitatively. More results are shown in the SM html (Sec. 5).

Since “look-here!” [34] is the most related approach to ours (both methods output parameters of image editing operators), but with a strict setting that limits the output effect to be subtle, we conducted a user-study to learn what kind of effects users prefer for the task of saliency reduction. 32 users were asked to look at 16 images with a marked region of interest, together with two outputs, ours (various effects) and “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 3(b) reports the breakdown of user selections between our method and “look-here!” [34]. 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 SM html Sec. 6), and Table 3(a) reports the percentage of saliency reduction (compared to the original image) for each of our effects and “look-here!”. Our method enables larger reduction in saliency compared to

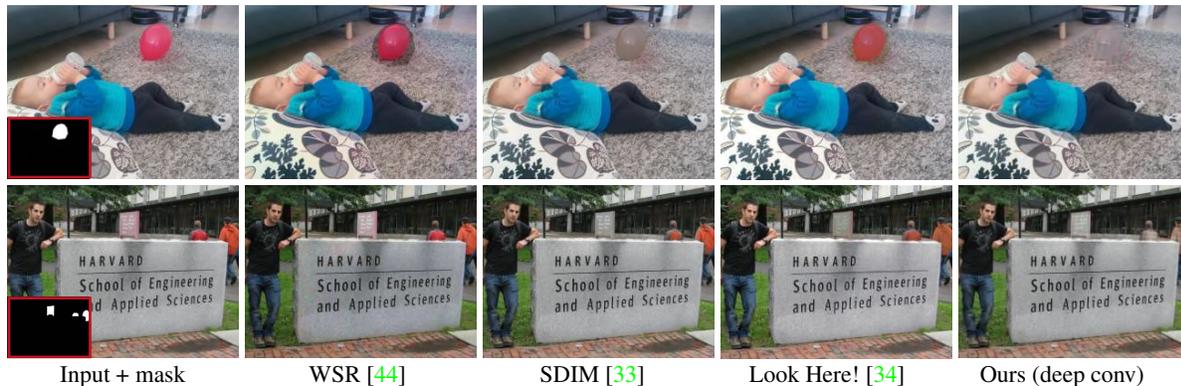


Figure 9. Comparison with previous attention retargeting methods on the Mechrez dataset [37]. The results of Look Here! [34] are generated with authors’ code, while the results of the other methods (no code available) are taken from [37]. More in the SM html (Sec. 5).



Figure 10. Comparison with “look-here!” [34]. More in the SM html (Sec. 6). In Table 3 we compare numerically with [34] the effective change to the saliency maps and the users’ preferences as found in our user study.

| WSR [44] | SDIM [33] | Look-Here! [34] | Ours (deep conv) |
|----------|-----------|-----------------|------------------|
| -12.38% | -29.80% | -21.51% | -40.71% |

Table 2. Quantitative results of the gaze saliency reduction achieved by our method and previous attention retargeting approaches.

“look-here!”, as expected from the more dramatic effects we design it for.

5. Discussion and 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. One notable limitation of our approach is that some of the effects, like recoloring and camouflage, require accurate masks. However, as we show in the SM html (Sec. 8), state-of-the-art tools for instance segmentation [39] can be used to reduce the level of expertise needed to annotate masks, while maintaining the quality of the results in most of the cases.

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

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

| ”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% |

Table 3. Comparison of our effects to “look-here!” [34]. Top: Reduction of average predicted saliency. Bottom: User study results. We show representative qualitative comparisons to [34] in Fig. 10, and more are available in the SM.

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.

Acknowledgments We thank Jon Barron, Phillip Isola, Tali Dekel, Dilip Krishnan and Bill Freeman for their valuable feedback, and thank Ethan Steinberg, Zhoude Liu, Nachiappan Valliappan and Na Dai for their contribution to saliency models and help with the Gaze user study application.

References

- [1] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 2019. [2](#)
- [2] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. <http://saliency.mit.edu>, 2012. [2](#), [5](#)
- [3] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. [5](#)
- [4] 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. [2](#)
- [5] 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. [2](#)
- [6] Jiawen Chen, Andrew Adams, Neal Wadhwa, and Samuel W Hasinoff. Bilateral guided upsampling. *ACM Transactions on Graphics (TOG)*, 35(6):1–8, 2016. [3](#)
- [7] 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. [2](#), [3](#)
- [8] 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. [4](#)
- [9] 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. [2](#)
- [10] 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. [2](#), [5](#)
- [11] 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. [5](#)
- [12] Simone Frinot, 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. [2](#), [3](#), [5](#)
- [13] Yosef Gandelsman, Assaf Shocher, and Michal Irani. ”double-dip”: Unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [4](#)
- [14] 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. [3](#)
- [15] Leon A Gatys, Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Guiding human gaze with convolutional neural networks. *arXiv preprint arXiv:1712.06492*, 2017. [1](#), [2](#), [3](#)
- [16] 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. [2](#)
- [17] 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. [2](#)
- [18] 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. [2](#)
- [19] 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. [2](#)
- [20] Laurent Itti. Visual salience. [doi:10.4249/scholarpedia.3327](https://doi.org/10.4249/scholarpedia.3327), 2007. [2](#), [5](#)
- [21] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001. [2](#)
- [22] Sen Jia and Neil DB Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020. [2](#), [3](#), [5](#)
- [23] Lai Jiang, Mai Xu, Xiaofei Wang, and Leonid Sigal. Saliency-guided image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16509–16518, 2021. [3](#)
- [24] 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. [2](#), [3](#)
- [25] 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. [3](#)
- [26] 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. [4](#), [5](#)
- [27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [5](#)
- [28] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987. [2](#)
- [29] 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. [2](#)
- [30] Matthias Kümmerer, 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. [2](#)
- [31] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013. [7](#)

- [32] 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. 2
- [33] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Saliency driven image manipulation. *Machine Vision and Applications*, 30(2):189–202, 2019. 2, 7, 8
- [34] 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. 1, 2, 3, 7, 8
- [35] 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. 2
- [36] Navyasri Reddy, Samyak Jain, Pradeep Yarlagadda, and Vineet Gandhi. Tidying deep saliency prediction architectures. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10241–10247. IEEE, 2020. 2, 5
- [37] Eli Shechtman Roey Mechrez and Lihi Zelnik-Manor. Mechrez distractors attenuation data set. <https://cg.technion.ac.il/Computer-Graphics-Multimedia/Software/saliencyManipulation>, 2012. 7, 8
- [38] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021. 5
- [39] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer, 2020. 5, 8
- [40] 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. 4
- [41] 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. 7
- [42] 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. 2, 5
- [43] Jeremy M Wolfe and Todd S Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):1–8, 2017. 2, 3, 5
- [44] 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. 2, 7, 8
- [45] Jonas Wulff and Antonio Torralba. Improving inversion and generation diversity in stylegan using a gaussianized latent space. *arXiv preprint arXiv:2009.06529*, 2020. 5
- [46] 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. 5
- [47] 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. 2