Look here! A parametric learning based approach to redirect visual attention

Youssef A. Mejjati[1]*, Celso F. Gomez[2], Kwang In Kim[3],
Eli Shechtman[2], and Zoya Bylinskii[2]

[1] University of Bath, UK
[2] Adobe Research
[3] UNIST

Abstract. Across photography, marketing, and website design, being able to direct the viewer's attention is a powerful tool. Motivated by professional workflows, we introduce an automatic method to make an image region more attention-capturing via subtle image edits that maintain realism and fidelity to the original. From an input image and a user-provided mask, our GazeShiftNet model predicts a distinct set of global parametric transformations to be applied to the foreground and background image regions separately. We present the results of quantitative and qualitative experiments that demonstrate improvements over prior state-of-the-art. In contrast to existing attention shifting algorithms, our global parametric approach better preserves image semantics and avoids typical generative artifacts. Our edits enable inference at interactive rates on any image size, and easily generalize to videos. Extensions of our model allow for multi-style edits and the ability to both increase and attenuate attention in an image region. Furthermore, users can customize the edited images by dialing the edits up or down via interpolations in parameter space. This paper presents a practical tool that can simplify future image editing pipelines.

Keywords: automatic image editing, visual attention, adversarial networks

1 Introduction

Photographers, advertisers, and educators seek to control the attention of their audiences, redirecting it to the content that matters most. Professionals working with images accomplish this via subtle adjustments to the contrast, tone, color, etc., of the relevant image regions to make them "pop-out". Motivated by professional workflows, we propose an automated learning based approach that predicts a set of global parametric edits to apply to an image to redirect a viewer's attention towards (or away from) a specific image region. Importantly, our approach is constrained, via an adversarial module, to produce realistic edits that remain faithful to the photographer's intentions and original image semantics. To ensure that the edited image successfully redirects attention, we use a state of the art saliency model during training.

While a glowing red arrow in an image would surely attract attention, this would not be practical for many use cases. Our GazeShiftNet model is specifically designed to be a

* Work done while Youssef was interning at Adobe Research.

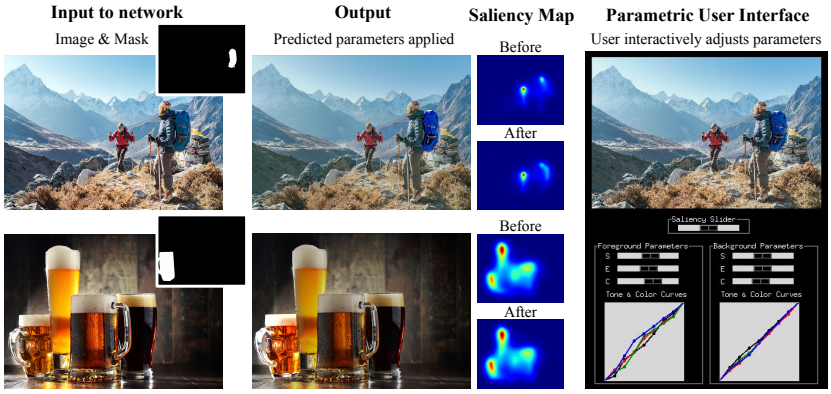

Fig. 1: GazeShiftNet takes an image and binary mask as input and predicts a set of parameters (sharpening, exposure, contrast, tone, and color curves) that are sequentially applied to the image to produce the output. The transformed image *subtly* redirects visual attention towards the mask region, seen from the saliency maps. A user can then tune the edits up or down (as shown on the right) at interactive rates, using the saliency slider.

practical solution according to the following criteria: (1) the model predictably redirects attention towards (or away from) image pixels denoted by a user-provided binary mask, (2) the proposed edits conserve the image semantics to maintain realism, (3) the model performs consistently and robustly across a variety of image content (e.g., objects and people), and (4) image edits are predicted at interactive rates, for use within applications.

Our solution involves applying global parametric transformations - sharpening, exposure, contrast, tone, and color - to the image, and employing an adversarial training strategy [15], rather than modifying each pixel separately [9, 14]. Our network predicts two sets of parameters to apply to the foreground and background of the image, demarcated by the input mask. The goal is to create a "pop-out" effect, which cannot be achieved by a single global transformation to the entire image. In sum, our model takes the form of a parametric generator followed by a cascade of differentiable layers that apply common photo editing transformations, mimicking a professional workflow. The choice and implementation of these transformations is motivated by related work on image enhancement [3, 18]. We train GazeShiftNet on a subset of the MS-COCO dataset [26], and present quantitative and qualitative results, including from user studies, on three different datasets. Compared to existing attention shifting methods [9, 14, 16, 28, 30], our approach successfully shifts viewers' attention, while achieving more realistic results that conserve the original image semantics and do not suffer from local color and texture artifacts.

We demonstrate the advantages of our global parametric approach for future applications. In particular, inference and editing take place at interactive rates, independent of image resolution, as the global transformations are predicted on a low resolution image using a feed-forward pass (rather than an optimization procedure [16, 28, 30]) and can be seamlessly applied to a high resolution version. Users can tweak the final result using photo editing sliders that control the parameters applied, since the interpolation takes place in parameter space. Our method also works for video data by applying the parameters predicted for the first frame to the rest of the video, thereby ensuring temporal

continuity. Finally, we present extensions of our model to produce image edits of different styles, and the option to both increase and attenuate visual attention in an image region.

2 Related Work

While image editing has been a popular research topic for many years [2, 8, 17, 37], recent breakthroughs have been made possible by generative adversarial networks [15], including in image generation [21, 35], image-to-image translation [31, 49], and style transfer [13, 19]. However, very few image editing methods exist for redirecting the viewer's attention towards a specific image region.

Among such methods, Gatys et al. [14] use an encoder-decoder model that takes an image and target saliency map as inputs, and generates a new image satisfying the target saliency map. This approach has a number of weaknesses: (1) operating directly in pixel space often produces artifacts in the final image, (2) a target saliency map is required as input, which is not straightforward for a user to provide, and (3) the edits from the generator are limited to a fixed image resolution. Chen et al. [9] similarly use an encoder-decoder model, but with an additional cyclic loss to stabilize the training procedure and reduce artifacts. However, this approach still suffers from the last two issues above, making it inconvenient for practical editing scenarios.

Su et al. [39] use smoothened power maps with steerable pyramids to equalize texture. Mechrez et al. [30] use patches from the same image to increase the saliency of a given region. The space of possible edits is naturally limited to appearances of other pixels in the image. Furthermore, this approach is very time consuming and requires an optimization per image, with compute time scaling with image resolution.

Related to our approach Hagiwara et al. [16] predict color and intensity changes in RGB space and add them pixel-wise to the original image. Mateescu et al. [28] use LAB space and adjust the hue within the object mask to maximize the KL divergence between the color distributions inside and outside the mask. Both methods require computationally intensive optimization at test time. Additionally, while these algorithms are based on heuristics that successfully redirect human attention, they do not preserve the semantics of the image, and can generate color anomalies affecting image realism [28].

Wong et al. [43] also use an optimization strategy to predict parameters such as saturation, sharpness and brightness to apply on different segments of an image, which require significant human supervision to generate. Compared to other parametric methods [16, 28], this method struggles to effectively redirect the viewer's attention [30]. In contrast, our method uses a deep neural network to predict global parametric transformations like exposure and contrast to apply to the image, all while preserving the semantics of the image, avoiding artifacts, and significantly reducing computation time.

Similar parametric transformations as ours are used for other image processing tasks, including for photo enhancement [7] and for recovering an image from its raw counterpart by emulating the image processing pipeline [18]. Tsai et al. [40] tackle a related problem, to make the composition of image foreground and background as natural as possible, by enhancing the foreground with respect to its background using an encoder-decoder network. We rather use an encoder-decoder network to enhance the foreground with the goal of redirecting the viewer's attention.

Attention modeling has also seen significant progress in the past few years thanks to deep neural networks [4, 6]. Attention models are becoming increasingly more accurate

and practical for applications including object detection [41, 42], object recognition [25, 33], content aware image re-targeting [1, 50], graphic design [5, 38, 47], image captioning [10, 48], and action recognition [24, 29]. While the application of such models has proven prolific, relatively little effort has been dedicated towards automatically manipulating attention, i.e. creating new image content that satisfies a given attention objective.

3 Method

Motivation: Given an image and region of interest, our goal is to automatically edit the image to make the selected region more attention-capturing. To learn how professionals complete this task, we ran an exploratory study on www.usertesting.com, providing participants with image-mask pairs, and asking them to edit the images in Adobe's Photoshop to make the masked regions more attention-capturing. We collected edits from 5 different participants on 30 high-resolution stock photographs (see Supplemental Material). Lessons learned: (1) image semantics are preserved, (2) edits are mostly restricted to parametric transformations, and (3) different operations are applied to image pixels inside and outside the mask. We designed our computational approach accordingly. Our approach (1) strives to maintain fidelity/faithfulness to the input image, while (2) applying global parametric transformations to the image, such that, (3) one set of transformations is applied to the mask (foreground), and a separate set is applied to the background.

Computational approach: Given an input image I and a binary mask $\mathbf{m}$, our goal is to generate a new image I' redirecting viewer attention towards the image region specified by $\mathbf{m}$. We generate I' by sequentially applying a set of parametric transformations commonly used in photo editing software and by computational photography algorithms [7, 18, 22, 44]. We apply the following ordered sequence of parameters: *sharpening*, *exposure*, *contrast*, *tone adjustment*, and *color adjustment* [18]. We discuss the order of operations in Sec. 4.4, and provide mathematical definitions in the Supplemental Material.

One possible approach is to train an encoder to predict a set of parameters to apply to the foreground region within a mask $\mathbf{m}$. However, we found that predicting a set of parameters both for the foreground $\mathbf{p}_f$ and for the background $\mathbf{p}_b$, is especially effective in cluttered scenes. Where multiple regions in the input image may have high saliency, attenuating the saliency of the background can be easier than increasing the saliency of the foreground. Formally, our encoder G_E is a neural network with a shared feature extractor and two specialized heads, predicting two sets of parameters: $G_E(I,\mathbf{m}) = (\mathbf{p}_f,\mathbf{p}_b)$.

Model architecture: Our generator G is composed of an encoder G_E and a decoder G_D, where G_E predicts global parametric transformations conditioned on I and $\mathbf{m}$, and G_D applies these transformations to the image to produce the final edited image. The pipeline is as follows (Fig. 2): we concatenate I with the mask $\mathbf{m}$ and feed this to a down-sampling convolutional neural network. Providing the mask as input allows the network to focus on the region to be enhanced. Furthermore, we reinforce the mask conditioning by applying the concatenation layer-wise throughout the convolutional part of the network. We bilinearly down-sample the mask before concatenating on each hidden layer to fit the corresponding input dimensions. The resulting representation is then fed to a series of fully connected layers via global average pooling. Global average pooling provides global information about high level features in the image, which is useful for predicting global parametric transformations. The foreground and background parameters are then predicted by two fully connected network heads.

The convolutional part of G_E encodes the semantics of the image and the region specified by the mask, and is the shared part of our network. The subsequent fully connected networks are the specialized heads that leverage this shared knowledge to predict separate parameters for the foreground and background image regions.

The decoder, G_D, applies the predicted parameters to the input image. G_D consists of a sequence of fixed (non-trainable) differentiable functions applied separately to the foreground and background image regions, demarcated by the mask $\mathbf{m}$. Specifically, we sequentially generate a pair of intermediate images $I'_f(i)$ and $I'_b(i)$ from $\mathbf{p}_f$ and $\mathbf{p}_b$, respectively, based on a fixed order of operations. At iteration i:

$$I'_f(i) = G_D(I'_f(i-1), \mathbf{p}_f(i)) \circ \mathbf{m} + G_D(I'_b(i-1), \mathbf{p}_b(i)) \circ (1-\mathbf{m}), \tag{1}$$

$$I'_b(i) = G_D(I'_b(i-1), \mathbf{p}_b(i)) \circ \mathbf{m} + G_D(I'_f(i-1), \mathbf{p}_f(i)) \circ (1-\mathbf{m}), \tag{2}$$

where $\circ$ refers to element-wise multiplication, and $I'_f(0) = I'_b(0) = I$. This decoding process is illustrated on the right side of Fig. 2. The final products of this sequential editing process process, I'_f and I'_b, are blended to synthesize the final image I', using $\mathbf{m}$:

$$I' = I'_f \circ \mathbf{m} + I'_b \circ (1-\mathbf{m}). \tag{3}$$

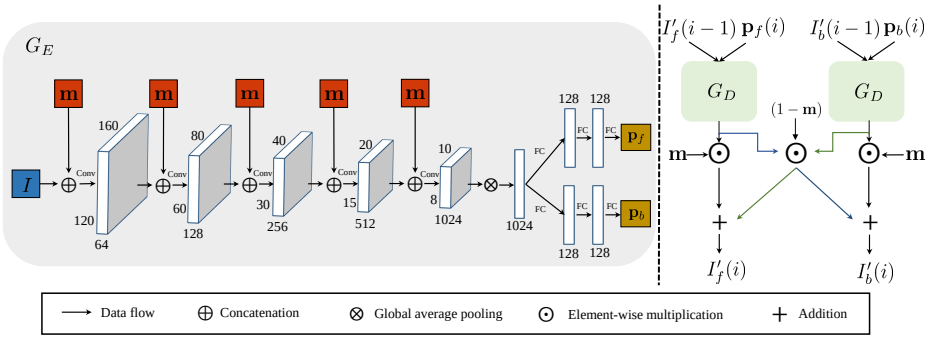

Fig. 2: GazeShiftNet architecture. Left: G_E takes the image I and the mask $\mathbf{m}$ as inputs and encodes them through a series of convolutional layers, then predicts the foreground and background parameters using fully connected network heads. Right: G_D applies a series of differentiable functions sequentially using the predicted parameters from G_E and creates the intermediate images $I'_f(i)$ and $I'_b(i)$ at iteration i.

Losses: We constrain our image transformations to respect the following properties: (1) I' should be a realistic image, without deviating significantly from the input I, (2) viewer attention should be redirected towards the image region specified by $\mathbf{m}$. To ensure the first property, we use a critic D, which differentiates between real and generated images. We train D adversarially with G_E using a hinge GAN loss known for its stability [20, 32, 45]:

$$\begin{aligned} \mathcal{L}_D(\Theta_D) &= -\mathbb{E}_{I,\mathbf{m}}\big[\min\{0, -D(I')-1\}\big] - \mathbb{E}_I\big[\min\{0, D(I)-1\}\big], \\ \mathcal{L}_G(\Theta_G) &= -\mathbb{E}_{I,\mathbf{m}}\big[D(I')\big], \end{aligned} \tag{4}$$

where Θ_D and Θ_G are the learnable weights of D and G_E, respectively.

To ensure the second property, we use a state-of-the-art deep saliency model [11], termed $\mathbf{S}$ throughout the paper, as a proxy of viewer attention. We use $\mathbf{S}$ to compute the saliency map of the output image, $S_{I'}=\mathbf{S}(I',\Theta_S)$, where Θ_S are the parameters of $\mathbf{S}$, and calculate the attention loss in the mask as:

$$\mathcal{L}_{att}(S_{I'},\mathbf{m})=-\frac{1}{\sum_{i,j}\mathbf{m}_{[ij]}}\sum_{(i,j)\in\mathbf{m}} S_{I'[ij]}\mathbf{m}_{[ij]}, \tag{5}$$

by iterating over all mask pixels (i,j). Normalization by the area of the masked region gives equal importance to regions of any size. The saliency maps obtained from $\mathbf{S}$ are normalized via softmax such that they sum to one. This ensures that increasing the saliency of the foreground necessarily decreases the saliency of the background.

Minimizing Eq. 5 alone would lead to unrealistic results, but in combination with the adversarial loss in Eq. 4, the algorithm tries to redirect viewer attention as much as possible while maintaining the realism of the output I'. The overall loss becomes:

$$\mathcal{L}(\Theta_G,\Theta_D)=\mathcal{L}_G(\Theta_G)+\mathcal{L}_D(\Theta_D)+\lambda_s\mathcal{L}_{att}(S_{I'},\mathbf{m}), \tag{6}$$

where λ_s controls the weight of the attention loss, which we empirically set to 2.5×10^4. Attention loss values are very small due to softmax normalization of the saliency maps.

Training dataset: We selected the MS-COCO dataset [26] because of the semantic diversity it contains, and the object segmentations that could be used as input masks to our algorithm. We curated the dataset to produce a set of image-mask pairs appropriate for our task. Specifically, we used the saliency model $\mathbf{S}$ to select images where the masked region is not already too salient, so that we can train our algorithm to shift attention to these regions. In addition, we discarded images containing only one mask instance or where the mask was too small or too large. Shifting attention to very small regions would likely require unrealistic image edits. Instances that are too large are often already quite salient. For the same reason, images with only one object segment (e.g., a single plane in the sky) may have no other regions to shift attention towards/away from. If a training image contained multiple masks satisfying all these conditions, we randomly picked one and discarded the others to ensure a diverse training set without over-representing any images. Following this process, we ended up with 49,949 image-mask pairs for the training set and 6,519 pairs for the validation set. We call this curated dataset 'CoCoClutter' to emphasize that the images contain multiple objects, without being dominated by any one object. We trained our model with a batch size of 4 for 37,500 iterations, sufficient for the model to converge. We used a learning rate of 1e-5, linearly decayed towards 0 starting halfway through training.

4 Evaluation

In this section, we compare GazeShiftNet to competing approaches on two datasets, to measure the ability of each method to shift visual attention to the target image region, and to produce image edits that are realistic and faithful to the original image. We first consider representative examples from all the methods and discuss common behaviors. Next, we evaluate the methods using computational metrics that measure saliency shift and fidelity, and finally, we present the results of three user studies to measure attention shift, image realism, and fidelity. We conclude with runtime comparisons of all the methods, and ablation experiments run on our model.

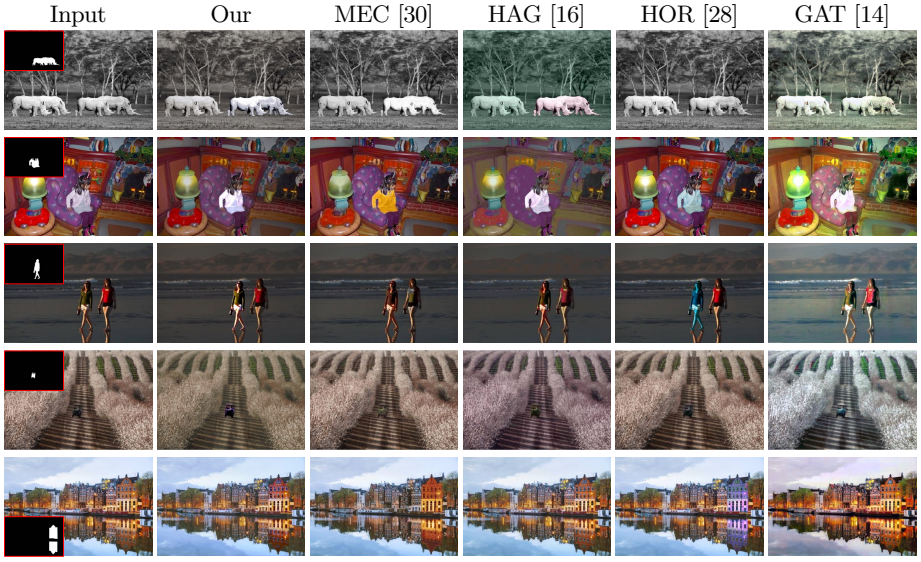

Fig. 3: Model comparisons on the Mechrez dataset. More in the Supplemental Material.

4.1 Qualitative comparisons

We present visual results obtained from our approach and competing methods on the Mechrez and CoCoClutter datasets in Figs. 3 and 4, respectively. Here we discuss the key properties of each method.

MEC (Mechrez et al. [30][4]): Being patch-based, this approach is limited to reusing colors and textures from the same image. This approach does not always preserve the authenticity of the original photograph (note the change in shirt colors in Fig. 3, rows 2 and 3). Changing image semantics may be unwanted behavior for some applications. MEC is not included in Fig. 4 because the code they provided did not reproduce their results, and led to significantly subpar quality images.

HAG (Hagiwara et al. [16]): This approach alters the intensity and color values in an image, most often relying on changing the background, with only slight modifications to the foreground. This can result in big changes to the image content and low fidelity to the original (note the change of background image hue in Fig. 3, rows 1-4, which has removed background texture in row 2).

HOR (Mateescu et al. [28]): This approach modifies the masked region only to maximize the distribution separation between colors of the foreground and background, which often results in substantial color anomalies and loss of realism (striking examples can be found in Fig. 3, rows 3 and 5, and Fig. 4, all rows).

GAT (Gatys et al. [14]): This approach is trained using a probabilistic attention model, and requires a full target saliency map as input, which is not practical for a user to produce. To make this approach comparable with ours, we used their attention

[4] Because the provided code could not reproduce the high quality results presented in their paper, for favorable comparison, we directly used images from their project page: https://webee.technion.ac.il/labs/cgm/Computer-Graphics-Multimedia/Software/saliencyManipulation/

model [14] to compute a map for our generated images, which we then used as the input target map for their method. From the last column of Figs. 3 and 4, we see that GAT both struggles with making objects more salient and produces many artifacts.

In comparison to these approaches, our method does not change the hue of objects or semantics of the image, remaining faithful to the input image while succeeding to emphasize the desired image region. In the next section, we will quantify these observations.

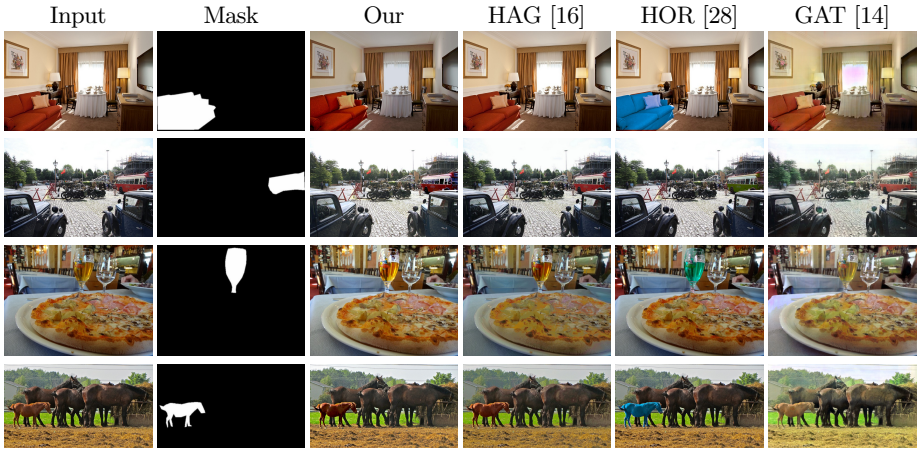

Fig. 4: Results on CoCoClutter dataset. More in the Supplemental Material.

4.2 Quantitative comparisons

We compare our approach to competing methods based on two criteria: (1) whether a given method successfully shifts visual attention towards desired image regions, and (2) whether the generated images remain realistic and maintain fidelity/faithfulness to the original photograph. We first performed a set of analyses using computational measures - i.e., with a computational model of saliency as a proxy for visual attention, and the LPIPS similarity metric as a measure of fidelity. Next, we ran a set of user studies on the top-performing models to (1) validate whether human attention indeed shifts towards the desired region, and (2) whether humans rate the generated images as both being realistic and having high fidelity to the original photographs.

We performed all quantitative comparisons on two evaluation datasets: a subset of images from the Mechrez dataset [30] and images from our CoCoClutter validation set. We sampled 64 images from the Mechrez dataset, corresponding to 46 images from their *object enhancement* collection and 18 images from their *saliency shift* collection. We selected images containing multiple objects, in order to evaluate the ability of a given algorithm to shift attention to different image regions. For the computational measures, we used the entire CoCoClutter validation set, while for the user studies, we randomly sampled a set of 50 images with clean masks (well-segmented objects).

Computational evaluation: To evaluate the ability of algorithms to successfully shift attention, we first use the initial and final saliency maps, corresponding to the

original and edited images, to measure the mean increase of attention inside the mask area (Tables 1a,1b, *Saliency increase - Absolute*). We also measure the relative increase of attention by normalizing by the initial saliency map (*Saliency increase - Relative*[5]). Second, we use measures previously used to evaluate attention shift [30], the Pearson correlation and weighed F-beta [27] between the final saliency map and the binary mask (*Similarity to mask*). Finally, as a measure of fidelity of the edited image to the original image, we use the LPIPS metric [46]. To have more insight about the behaviour of each algorithm, we compute LPIPS on the entire image (*Full*), on the background only (*BG*) and on the foreground only (*FG*). Note that Mechrez et al. [30] does not appear in Table 1b as their code did not produce usable results on this dataset.

The three sets of computational measures used are complementary. *Saliency increase* evaluates the increase in saliency values for the foreground, independently of changes to the background. *Similarity to mask* considers the extreme case where the ground truth saliency map would be defined by the mask, hence taking into account changes in saliency both for the foreground and background. Finally, LPIPS measures how different the edited images are compared to the originals. An ideal model has high *Saliency increase* and *Similarity to mask*, and low LPIPS.

Our approach performs the best on all computational measures of saliency across Mechrez (Table 1a) and CoCoClutter (Table 1b) datasets. MEC [30] comes in second due to its ability to replace foreground patches (from the mask) with salient patches from the same image. However, this often comes at the cost of sacrificing the colors and semantics of the original image. HAG [16] is third overall, but suffers from high LPIPS scores (LPIPS *Full* and *BG*) as it often heavily modifies the original colors and details of the background. LPIPS scores for the foreground are generally very small, suggesting that HAG relies mostly on background modifications in its attention shifting pipeline. In contrast, HOR [28] achieves very high LPIPS scores on the foreground due to the severe color artifacts it creates. At the same time, this method achieves the lowest LPIPS scores when considering the entire image (LPIPS *Full*) because it leaves the background untouched, which usually covers the largest area of the image. Finally GAT [14] performs the worst in terms of shifting attention, while simultaneously heavily modifying the image, resulting in the highest LPIPS scores and lowest *Saliency increase* and *Similarity to mask* scores. Overall, our algorithm is most successful at shifting computational attention, all while conserving original image properties, including, colors, textures, and semantics. It is in the top two models when considering LPIPS scores.

User studies: We reinforce the findings from the computational measures by collecting human data using Amazon's Mechanical Turk for three tasks: visual attention, image fidelity, and image realism. Our first set of studies measured shifts in human attention, when images are modified by the various methods compared. We used the same crowd-sourced gaze tracking method, *CodeCharts* [34], that was used to collect training data for the saliency model [11] we adapted in GazeShiftNet. In CodeCharts, participants are asked to look at an image for a few seconds, then a jittered grid of alphanumeric triplets ("codes") is flashed for a brief interval, and the participant is subsequently prompted to enter the code seen last. This task design captures the area where a participant was looking at the moment when the image disappeared. CodeCharts has been shown to

[5] Relative saliency increases can grow large when the corresponding instance has an average initial saliency value near zero.

(a)

Model	LPIPS ↓			Saliency increase ↑		Similarity to mask ↑	
	Full	BG	FG	Absolute	Relative	WFB	CC
Our	5.96	5.10	0.81	3.80	35.77	11.30	30.43
MEC [30]	8.95	7.65	1.29	3.42	28.39	10.42	26.25
HOR [28]	1.60	0	1.41	1.74	16.29	9.85	24.43
HAG [16]	11.08	10.68	0.37	2.24	21.75	10.33	26.12
GAT [14]	25.64	24.72	1.11	0.34	3.05	9.59	22.98

(b)

Model	LPIPS ↓			Saliency increase ↑		Similarity to mask ↑	
	Full	BG	FG	Absolute	Relative	WFB	CC
Our	4.87	2.84	1.95	1.99	25957.92	11.81	20.65
HAG [16]	7.37	6.58	0.69	1.30	24419.15	11.22	18.89
HOR [28]	4.00	0	3.61	1.27	11065.47	11.27	18.98
GAT [30]	30.07	27.08	3.61	-0.05	2920.75	10.23	15.85

(c)

Model	Avg. run time
Our	8.87s
HOR [28]	31.82s
HAG [16]	4743.54s
MEC [30]	>1 day

Model	Avg. run time
Our	1.54s
GAT [14]	4.34s

(d)

Model	LPIPS ↓	Saliency increase ↑	
	Full	Absolute	Relative
Our	5.96	3.80	35.77
Parameter ablations			
tone + color	10.97	5.33	56.19
sharp + exp + cont	1.76	1.47	13.00
color	9.95	3.50	37.25
our + saturation	9.70	4.84	48.76
Fixed parameters	2.28	1.95	17.23
fg/bg ablations			
bg-only	2.31	0.54	3.50
fg-only	1.19	2.53	29.40
order of operations ablations			
col,ton,con,exp,sha	6.74	4.01	35.36
ton,col,sha,exp,con	8.04	3.73	32.47
con,exp,sha,ton,col	6.55	3.65	36.98
ton,col,sha,con,exp	7.87	3.72	30.49

Table 1: (a) Computational evaluation on Mechrez dataset. (b) Computational evaluation on CoCoClutter dataset. The top two performing models according to each metric are highlighted in green (darker is better). (c) Left: Run time averaged over 30 high resolution images; Right: As GAT is unable to run directly on high resolution images, we use low resolution versions of the same images. (d) Ablation studies on Mechrez dataset. Our chosen method involves applying sharpening, exposure, contrast, tone adjustment, and color adjustment (in that order) to both foreground and background. Ablations consider a subset of parameters and different orderings, as well as application to either one of foreground/background. Darker green colors indicate better scores, darker red colors indicate worse scores. All metrics except run time are multiplied by 100 for legibility.

approximate human eye movements collected using an eye tracker [34], which we use to evaluate the ability of algorithms to shift human attention by modifying images.

We used CodeCharts to collect human attention data on 64 images from the Mechrez dataset and 50 from the CoCoClutter dataset. We collected attention data on the original images and on the edited images produced by each method. We obtained gaze points from an average of 50 participants per image, that we converted into an attention heatmap for the image. We then compared the attention heatmaps of the original images to those of the edited images, to evaluate the relative attention increase achievable by each method. These values are plotted on the x-axes of Figs. 5a and b (left) for the CoCoClutter and Mechrez datasets, respectively. The table of values is also available in the Supplemental Material. Based on these study results, GazeShiftNet achieves a greater attention shift than HAG, a smaller one than HOR, and a comparable one to MEC. These results are unsurprising given the algorithm behaviors: HAG mostly modifies the background which leads to less noticeable change to the masked objects, whereas HOR often drastically changes the color of the foreground object making it significantly stand out from the rest of the image. From Fig. 5, we can see that our method achieves a balance between shifting attention and maintaining image fidelity, as described below.

The next set of user studies measured image fidelity. We asked an average of 25 participants to evaluate each edited image compared to its original using the following options: not edited (3), slightly edited (2), moderately edited (1), definitely edited (0). In parentheses are the fidelity scores we assigned to each answer (higher is better). Fidelity score distributions across images, averaged over study participants, are visualized as box plots in Fig. 5a and b (middle) for the CoCoClutter and Mechrez datasets, respectively. Both GazeShiftNet and HAG achieve fidelity scores that are statistically

significantly higher than HOR on the CoCoClutter dataset, and higher than MEC on the Mechrez dataset ($p < 0.001$). No other pairwise comparisons were significant. The human fidelity judgements are intended to evaluate the same aspect of models as the LPIPS computational metric. Because the fidelity scores of the models correspond most to their LPIPS FG scores from Tables 1a and 1b, we suspect that study participants focused more on the foreground regions of the edited images when judging fidelity.

While fidelity measured how similar an edited image is compared to its original, we also ran a user study to evaluate the realism of the edited images in isolation. We asked an average of 25 participants to evaluate whether each image is: definitely not edited (3), probably not edited (2), probably edited (1), definitely edited (0). In parentheses are the realism scores we assigned to each answer (higher is better). Realism score distributions across images, averaged over study participants, can be found in Fig. 5a and b (right) for the CoCoClutter and Mechrez image datasets, respectively. On Mechrez, all methods perform comparably in terms of realism. This supports the findings from Mechrez et al. [30], who similarly found that when evaluated using a realism user study, the algorithms performed comparably. However, on the CoCoClutter dataset, differences across methods are more pronounced. GazeShiftNet and HAG obtained statistically significantly higher realism scores than HOR ($p < 0.001$). The reason for these differences is that the Mechrez dataset masks often cover only part of an object, which conserves realism when edits are applied (e.g., an edited blue shirt is no less realistic than the original gray one; though fidelity suffers in this case), but the CoCoClutter dataset masks include full objects, which exposes significant failure methods of other approaches (such as when the HOR method turns entire people blue).

Across fidelity and realism, our approach achieves the smallest standard deviations in scores, and the lower range of scores starts higher (Fig. 5). This shows that GazeShiftNet behaves more consistently across a variety of image types, with fewer catastrophic failures than some of the other methods, making it a more practical method overall.

As additional validation of our approach, we compared its performance to that of professional users on the 30 high-resolution images from our exploratory studies (Sec. 3). The results show that while users produce edits that are judged to have more realism and fidelity (and correspondingly lower LPIPS *FG*) scores, our model is able to achieve both higher saliency and attention shift scores than users on average (although some users produce quite effective attention shifts in photos). We note that our model achieves the highest increase in saliency/attention than HAG and HOR on this image dataset. Detailed study results can be found in the Supplemental Material.

4.3 Run time comparisons

We timed each algorithm on 30 high-resolution images ($[648,1332] \times 1500$ pixels) obtained from Adobe Stock and manually annotated to include a masked object per image. Average run times over the 30 images are listed in Table 1c. Our algorithm is significantly faster than HAG and MEC, which are iterative optimization-based algorithms. In fact, MEC took more than one day to process the 30 images. HOR is fairly quick but still takes more than a second per image. GAT cannot run on high resolution images, so for this comparison we used the same 30 images in resolutions corresponding to the required neural network input sizes. Because their network is deeper and their method additionally requires computing a target saliency map as an input, their computation time is nearly 3 times ours.

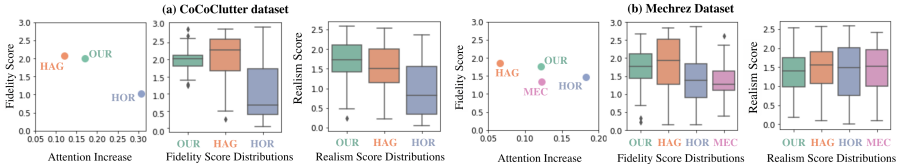

Fig. 5: Results of user studies on three separate crowdsourcing tasks - measuring image fidelity, realism, and human attention - on two datasets: (a) CoCoClutter and (b) Mechrez. Different methods trade-off average image fidelity for average attention increase, and vice versa, whereas our approach achieves a good balance of both (left). Box plots of the fidelity score distributions (middle) and realism score distributions (right) demonstrate the variability in the performance of each method. Ours performs most consistently, with the smallest range of scores, i.e., fewer failure modes. All methods perform similarly in realism on the Mechrez dataset because of the nature of the objects selected for emphasis (see text for details).

4.4 Ablation studies

Table 1d includes a summary of the ablation tests evaluating our design choices: (1) the set of parametric transformations, (2) whether to apply parametric transformations to foreground, background, or both, (3) the ordering of transformations applied to images. As it is not immediately obvious how to balance the trade-off between realism and attention shift across our different ablations, we identify the five best models according to three metrics (LPIPS, absolute and relative saliency increase), and choose a high-performing model across all criteria. In Table 1d the best performing models are highlighted in green (darker is better), and the worst performing models are highlighted in red (darker is worst). Our final model selection achieves good performance across the metrics. A detailed discussion of the ablation experiments can be found in the Supplemental Material.

5 Extensions

Our method could be extended further, making it even more flexible. We discuss additional use cases and extensions below.

Application to videos: GazeShiftNet can be seamlessly generalized to videos by predicting parameters on the first frame, and applying them to subsequent video frames (provided we have a common, segmented object across frames). We find this produces good results on the rest of the frames while avoiding flickering. In contrast, most competing attention shifting algorithms [14, 16, 30] would need to be run on each frame separately, as their transformations cannot be transferred across images. We show in Fig. 6 snapshots from a video from the DAVIS dataset [36].

Interactive image editing: GazeShiftNet, being parametric, makes it easy to hand control of the edited image back to the user. By interpolating between the predicted parameters and the set of parameters where $I' = I$, we can let a user dial the edits up or down. We built a prototype of an interface where a user can interact with a single saliency slider that interpolates the parameters at interactive rates (see Fig. 1 and the Supplemental Material). This form of interpolation results in smooth, artifact-free image transformations. The user can also adjust each of the parameter sliders separately. Importantly, while parameters can be predicted on smaller image sizes, final

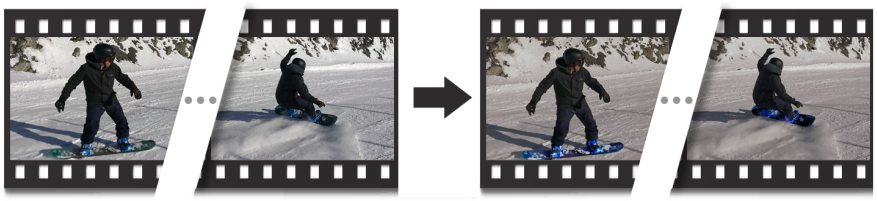

Fig. 6: Our method can be seamlessly applied to video frames. We predict parameters for the first frame and transfer them to all subsequent frames containing the same segmented object (in this case, the snowboard).

transformations can be applied to professional high-resolution photography at interactive rates. These properties do not hold for non-parameteric methods [14, 30]. Optimization-based methods [16, 28] do not allow for interactive and artifact-free post-processing edits.

Stochastic parameter generation: There isn't a single way to enhance an object in an image, and different users may choose to apply different sets of transformations when editing an image (two far right columns in Fig. 7). Motivated by such natural variability, GazeShiftNet can be extended to produce results in 'multiple styles' by predicting different parameter values for the transformations. We introduce stochasticity using a latent vector $z \in \mathbb{R}^{10}$, randomly sampled from a Gaussian distribution. This vector z is first tiled to match the mask dimensions, and then concatenated as an input, similarly to how the mask is handled. In fact, the architecture is the same as in Fig. 2, but replacing $\mathbf{m}$ with $\mathbf{mz}$ (where $\mathbf{mz}$ is the concatenation of $\mathbf{m}$ and tiled z: $\mathbf{z} \in \mathbf{R}^{h \times w \times 10}$). To force the network to actually use $\mathbf{z}$ in producing the output parameters $\mathbf{p}_f$ and $\mathbf{p}_b$, we add an additional loss encouraging z to be reconstructed from $\mathbf{p}$ (the concatenation of $\mathbf{p}_f$ and $\mathbf{p}_b$): $\mathcal{L}_r(z, \mathbf{p}_f, \mathbf{p}_b) = \frac{1}{10}\|z - \mathrm{ENC}(\mathbf{p})\|_1$, where ENC is an encoder formed by a series of fully connected layers. We can then randomly sample the latent vector z to produce diverse edits (Fig. 7).

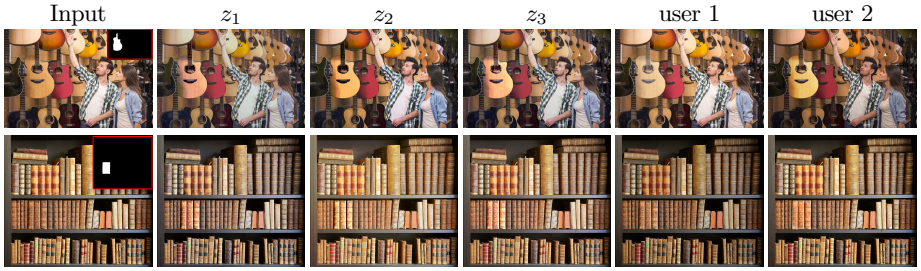

Fig. 7: Sampling different latent z_i in our model results in stochastic variations, all of which achieve the same saliency objective, but with different 'styles' of edits. We include two sample edits done by professional users from our exploratory studies.

Decreasing human attention: While GazeShiftNet has been trained to shift viewer attention *towards* a specific region in the image, it can also be trained to achieve the opposite: shifting attention *away* from an image region. This can be useful for distractor attenuation [12, 23], e.g., reducing the prominence of passerbys in personal

photo collections. Towards this goal, we extend our network to have two additional heads to output the parameters $\mathbf{p}_f^{dec}$ and $\mathbf{p}_b^{dec}$. We then generate two images, I'_{dec} using $\mathbf{p}_f^{dec}$ and $\mathbf{p}_b^{dec}$ and I'_{inc} using $\mathbf{p}_f^{inc}$ and $\mathbf{p}_b^{inc}$. The new attention loss (Eq. 5) becomes: $\mathcal{L}'_{att}(S_{I'_{inc}}, S_{I'_{dec}}, \mathbf{m}) = \mathcal{L}_{att}(S_{I'_{inc}}, \mathbf{m}) - \mathcal{L}_{att}(S_{I'_{dec}}, \mathbf{m})$. To successfully train this model, we had to discard from the training set all masks where the average computational saliency was too small. Such masks were not suitable for the objective of decreasing saliency. In addition, we trained this model for double the amount of time, with hyper-parameter $\lambda_s = 2 \times 10^4$. Results from this network are visualized in Fig. 8.

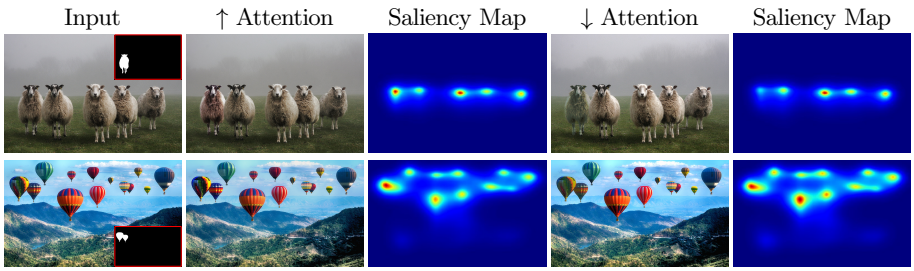

Fig. 8: Using the input image and corresponding mask, we generate two images, one to shift visual attention towards the mask (col. 2) as seen from the saliency map in col. 3, and one to shift attention away from the mask (col. 4, saliency map in col. 5).

6 Conclusion

We presented a practical method for automatically editing images (with extensions to videos) in a subtle way, while effectively redirecting viewer attention. We demonstrated that our method achieves a good balance between shifting the attention - of saliency models and human participants alike - while maintaining the realism and fidelity of the original image (i.e., the photographer's intent, semantics/colors of the original image). Having a practical method depends on balancing these objectives, and achieving consistent and robust results across a variety of images. We also showed significant improvements in computation time over past approaches. Most importantly, our global parametric approach allows running our method on high-resolution images and videos at interactive rates, as well as interpolating in parameter space to hand control of the edits back to a user within an editing interface. Such an effective and practical approach to image editing can benefit many applications, including website design, effective marketing campaigns, and image enhancement. These tasks are typically completed by professionals using advanced image editing software. Our automated approach can simplify, and has the potential to replace, some professional image editing workflows.

Acknowledgements: Y. A. Mejjati thanks the Marie Sklodowska-Curie grant No 665992, and the Centre for Doctoral Training in Digital Entertainment (CDE), EP/L016540/1. K. I. Kim thanks Institute of Information & communications Technology Planning Evaluation (IITP) grant (No.20200013360011001, Artificial Intelligence Graduate School support (UNIST)) funded by the Korea government (MSIT).

References

1. Achanta, R., Süsstrunk, S.: Saliency detection for content-aware image resizing. In: ICIP (2009)
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. In: TOG (2009)
3. Bianco, S., Cusano, C., Piccoli, F., Schettini, R.: Learning parametric functions for color image enhancement. In: International Workshop on Computational Color Imaging (2019)
4. Borji, A.: Saliency prediction in the deep learning era: Successes and limitations. TPAMI (2019)
5. Bylinskii, Z., Kim, N.W., O'Donovan, P., Alsheikh, S., Madan, S., Pfister, H., Durand, F., Russell, B., Hertzmann, A.: Learning visual importance for graphic designs and data visualizations. In: UIST (2017)
6. Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., Durand, F.: Where should saliency models look next? In: ECCV (2016)
7. Chandakkar, P.S., Li, B.: A structured approach to predicting image enhancement parameters. In: WACV (2016)
8. Chen, S.E., Williams, L.: View interpolation for image synthesis. In: SIGGRAPH (1993)
9. Chen, Y.C., Chang, K.J., Tsai, Y.H., Wang, Y.C.F., Chiu, W.C.: Guide your eyes: Learning image manipulation under saliency guidance. In: BMVC (2019)
10. Cornia, M., Baraldi, L., Cucchiara, R.: Show, control and tell: a framework for generating controllable and grounded captions. In: CVPR (2019)
11. Fosco, C., Newman, A., Sukhum, P., Zhang, Y.B., Zhao, N., Oliva, A., Bylinskii, Z.: How much time do you have? modeling multi-duration saliency. In: CVPR (2020)
12. Fried, O., Shechtman, E., Goldman, D.B., Finkelstein, A.: Finding distractors in images. In: CVPR (2015)
13. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016)
14. Gatys, L.A., Kümmerer, M., Wallis, T.S., Bethge, M.: Guiding human gaze with convolutional neural networks. arXiv preprint arXiv:1712.06492 (2017)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
16. Hagiwara, A., Sugimoto, A., Kawamoto, K.: Saliency-based image editing for guiding visual attention. In: Proceedings of the 1st international workshop on pervasive eye tracking & mobile eye-based interaction (2011)
17. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.H.: Image analogies. In: SIGGRAPH (2001)
18. Hu, Y., He, H., Xu, C., Wang, B., Lin, S.: Exposure: A white-box photo post-processing framework. SIGGRAPH (2018)
19. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
20. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard gan. In: ICLR (2019)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
22. Kaufman, L., Lischinski, D., Werman, M.: Content-aware automatic photo enhancement. In: Computer Graphics Forum (2012)
23. Kolkin, N.I., Shakhnarovich, G., Shechtman, E.: Training deep networks to be spatially sensitive. In: ICCV (2017)
24. Koutras, P., Maragos, P.: Susinet: See, understand and summarize it. In: CVPR Workshops (2019)

25. Li, N., Zhao, X., Yang, Y., Zou, X.: Objects classification by learning-based visual saliency model and convolutional neural network. Computational intelligence and neuroscience (2016)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
27. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: CVPR (2014)
28. Mateescu, V.A., Bajić, I.V.: Attention retargeting by color manipulation in images. In: Proceedings of the 1st International Workshop on Perception Inspired Video Processing (2014)
29. Mathe, S., Sminchisescu, C.: Dynamic eye movement datasets and learnt saliency models for visual action recognition. In: ECCV (2012)
30. Mechrez, R., Shechtman, E., Zelnik-Manor, L.: Saliency driven image manipulation. WACV (2019)
31. Mejjati, Y.A., Richardt, C., Tompkin, J., Cosker, D., Kim, K.I.: Unsupervised attention-guided image-to-image translation. In: NeurIPS (2018)
32. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018)
33. Moosmann, F., Larlus, D., Jurie, F.: Learning saliency maps for object categorization. In: ECCV (2006)
34. Newman, A., McNamara, B., Fosco, C., Zhang, Y.B., Sukhum, P., Tancik, M., Kim, N.W., Bylinskii, Z.: Turkeyes: A web-based toolbox for crowdsourcing attention data. In: ACM CHI Conference on Human Factors in Computing Systems (2020)
35. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Gaugan: semantic image synthesis with spatially adaptive normalization. In: SIGGRAPH (2019)
36. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
37. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. TOG (2003)
38. Shen, C., Zhao, Q.: Webpage saliency. In: ECCV (2014)
39. Su, S.L., Durand, F., Agrawala, M.: De-emphasis of distracting image regions using texture power maps. In: ICCV workshops (2005)
40. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: CVPR (2017)
41. Wang, W., Shen, J., Cheng, M.M., Shao, L.: An iterative and cooperative top-down and bottom-up inference network for salient object detection. In: CVPR (2019)
42. Wang, W., Zhao, S., Shen, J., Hoi, S.C.H., Borji, A.: Salient object detection with pyramid attention and salient edges. In: CVPR (2019)
43. Wong, L.K., Low, K.L.: Saliency retargeting: An approach to enhance image aesthetics. In: WACV-workshop (2011)
44. Yan, Z., Zhang, H., Wang, B., Paris, S., Yu, Y.: Automatic photo adjustment using deep neural networks. TOG (2016)
45. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. ICML (2019)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
47. Zheng, Q., Jiao, J., Cao, Y., Lau, R.W.: Task-driven webpage saliency. In: ECCV (2018)
48. Zhou, L., Zhang, Y., Jiang, Y., Zhang, T., Fan, W.: Re-caption: Saliency-enhanced image captioning through two-phase learning. IEEE Transactions on Image Processing (2020)
49. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
50. Zünd, F., Pritch, Y., Sorkine-Hornung, A., Mangold, S., Gross, T.: Content-aware compression using saliency-driven image retargeting. In: ICIP (2013)