Deep Saliency Prior for Reducing Visual Distraction - Supplementary Material

July 1, 2021

1 Image regularization as a baseline

In the paper, we've demonstrated how a naive and direct optimization on the region-of-interest's pixels may lead to an out-of-distribution result. Here, we show that adding a regularization term to equation (1) in the paper is insufficient. To apply the regularization, we first expand the main loss function to be

$$\mathcal{L}_{\text{sal}}\left(\tilde{\mathbf{I}}\right) + \beta\mathcal{L}_{\text{sim}}\left(\tilde{\mathbf{I}}\right) + \gamma\mathcal{L}_{\text{reg}}\left(\tilde{\mathbf{I}}\right),$$

where $\mathcal{L}_{\text{reg}}\left(\tilde{\mathbf{I}}\right)$ is a regularization loss on the pixels of $\tilde{\mathbf{I}}$, e.g., image total variation

$$\mathcal{L}_{\text{reg}}\left(\tilde{\mathbf{I}}\right) = M \circ (\|\nabla_x \tilde{\mathbf{I}}\|_1 + \|\nabla_y \tilde{\mathbf{I}}\|_1),$$

where ∇_x and ∇_y represent the gradients with respect to the horizontal/vertical axes x and y, respectively.

Figure 1 shows a couple of results with different values of the hyper-parameter γ. We can see that simple regularization is insufficient for avoiding the out-of-distribution examples, or for generating realistic images.

2 Implementation Details

First, recall that our loss functions is

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

where $\tilde{\mathbf{I}} = O_\theta(\mathbf{I})$.

2.1 $\mathcal{L}_{\text{sal}}$

In (1), $\mathcal{L}_{\text{sal}}$ is defined as $\mathcal{L}_{\text{sal}}\left(\tilde{\mathbf{I}}\right) = \left\|\mathbf{M} \circ \left(S(\tilde{\mathbf{I}}) - \mathbf{T}\right)\right\|^2$. The target map $\mathbf{T}$ can be an arbitrary map, where the saliency can be controlled. For example,

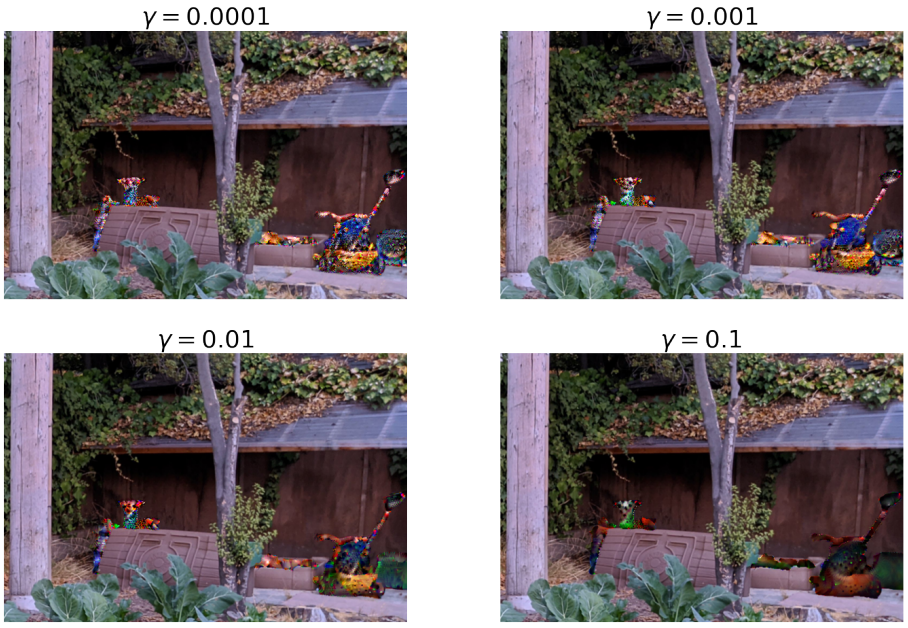

Figure 1: A naive baseline that applies a direct optimization to the region's pixels together with a total variation regularization yields out-of-distribution, unnatural results. γ is the regularization term weight described in equation (1) in the original paper.

Figure 2 shows an example where the saliency level can be controlled by a single scalar with $\mathbf{T} = \alpha S(\mathbf{I})$, for various values of α that are indicated in the paper.

However, all other examples in the paper are generated by setting $\mathbf{T} \equiv 0$ to minimize the saliency within the region of interest (or $\mathbf{T} \equiv 1$ to maximize the saliency, for the GAN operator). When $\mathbf{T} \equiv 0$ or 1, $\mathcal{L}_{\text{sal}}\left(\tilde{\mathbf{I}}\right)$ is equivalent to $\left\|\mathbf{M} \circ S(\tilde{\mathbf{I}})\right\|^2$ for saliency decreasing or $-\left\|\mathbf{M} \circ S(\tilde{\mathbf{I}})\right\|^2$ for saliency increasing, respectively.

2.2 $\mathcal{L}_{\text{sim}}$

In equation (1), the similarity term in the loss is defined as

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

Another way to preserve similarity of pixels outside of the mask is using a hard constraint such as $(1 - \mathbf{M}) \circ \left(\tilde{\mathbf{I}} - \mathbf{I}\right) = 0$. We can enforce this constraint by copying pixels outside of mask $\mathbf{M}$ in $\mathbf{I}$ to $\tilde{\mathbf{I}}$ in each iteration step. Our recolor and warp operators adopt this hard constraint approach, while "Learning

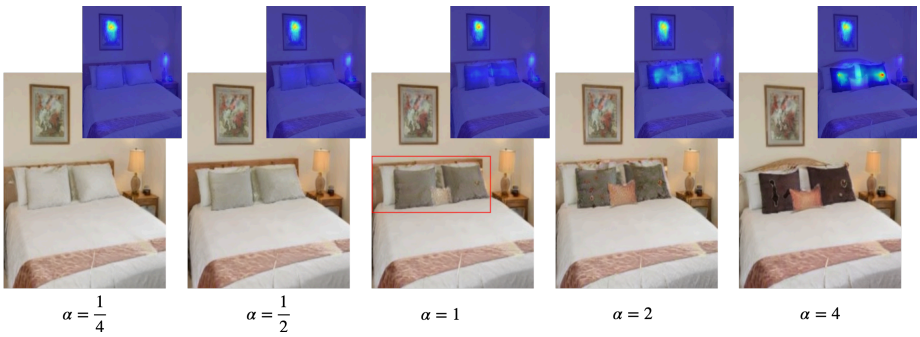

Figure 2: Varying the target saliency level. Given an image (middle) and a region of interest (marked by a red box) our approach can control the saliency level in the specified region of interest. The target saliency is α times the original saliency (within the mask), i.e, $\mathbf{T} = \alpha S(\mathbf{I})$.

convolutional networks" and GAN operator adopt $\mathcal{L}_{\text{sim}}$ as equation (1). Such a hard constraint enables us to crop the image around the distracting region when the region of interest does not have enough saliency in the original image. Namely, we can crop the image to a smaller one that still contains the region of interest - a step which enables to gain high-saliency, which can be removed more effectively. Since pixels outside of the mask will be exactly the same, the crop will not create any artifacts around the cropping boundary. This is helpful when the region of interest is too small or not very salient in the original, large image, that contains more attention grabbing regions.

2.3 Hyper parameters

- For the "recolor" operator, the hyper parameters include weight parameter $\gamma = 0.01$, learning rate $= 10$, number of iterations $= 800$.
- For the "warp" operator,, the hyper parameters include weight parameter $\gamma = 0.1$, learning rate $= 1$, number of iterations $= 500$.
- For the "deep conv" operator, the hyper parameters include weight parameters $\beta = 100$, $\gamma = 0$, learning rate $= 0.1$, number of iterations $= 1000$.
- For "GAN" operator, we set $\gamma = 0$, $\beta = 0.03$, learning rate $= 0.01$, number of iterations $= 200$.

2.4 GAN operator on real images

In order to apply our approach to real images, the image should be first embedded in latent codes, e.g., with a GAN inversion technique. Existing works [8, 9] show that in-domain images (i.e., images similar to the GAN training

set) can be reconstructed effectively within different subspaces of StyleGAN. For example, Figure 3 shows some examples of reconstructed images with GAN inversion techniques [9]. As can be seen, the technique achieves reasonable reconstruction quality on those in-domain examples, ,however, reconstruction of some facial, fine grained details which are crucial for identity preservation need to be improved. Moreover, GAN inversion for out-of-domain images is even a more challenging task. However, recently, some works [4] show promising results in reconstructing out-of-domain real images too. Exploring GAN inversion techniques is beyond the scope of this work, and we assume that a latent code corresponding to the input image is given, so our method is focused on the modification of the latent codes to modify the content (and saliency) of the output image.

2.5 Saliency model

For all the experiments in this paper we use the saliency prediction model EML-Net of [6], with minor modifications. More specifically, we use NasNet [15] only, without DenseNet [5] in the EML-Net architecture for simplicity, since adding DenseNet will only slightly increase the accuracy (as reported in Table 1 in [6]). Moreover, in EML-Net, the loss function is the combination of Normalized Scanpath Saliency(NSS), Kullback-Leibler Divergence (KLD) and Pearson's Correlation Coefficient (CC) [6, 2]). We instead use NSS and cross-Entropy loss (between the predicted saliency and ground-truth gaze fixation points), since we found the pixel-wise cross-entropy loss generates better results, while not requiring any additional parameter (like Gaussian kernel size in KLD or CC).

2.6 Data sets

The data set used to train our saliency model is the Salicon data set [7]. The data sets to train domain-specific GAN are FFHQ dataset [8] for faces, the LSUN dataset [14] for churches, towers and bedrooms, and AFHQ [3] for animals. All these data sets are used with appropriate citation and license permission.

3 More results/analysis on eye-gaze user study

Besides the 31 images to evaluate decreasing saliency as reported in the paper, we also include 13 images from our GAN operator for increasing the saliency. Due to space limitations, this is reported here in the supplementary materials instead of the main paper. Example images and their gaze saliency maps for increasing saliency cases can be found in the supplementary html. The average gaze duration is increased from 985.09 ms to 1548.23 ms, and average first gaze time is decreased from 1555.28 ms to 1210.44 ms. We have also conducted the Paired Samples T-Test for the three gaze features on these 13 examples, as in Table 1. We can see that we have achieved the statistical significance with these

Figure 3: Reconstruction of real images by GAN inversion. The reconstructed images are then manipulated by our proposed method with the GAN operator.

Table 1: Paired Samples T-Test of manipulated images vs original ones, with three gaze features.

Gaze saliency	**Duration**	**First gaze**
t=4.66	t=3.86	t=-1.96
p=0.00055	p=0.0022	p=0.073

small number of examples (except for first gaze time, which is a noisier feature essentially).

4 Implementation details for experiments

4.1 Details of eye-gaze user study

We developed a user study app on Android phones to measure users' gaze/attention on the original and manipulate images, based on the mobile eye tracking techniques introduced in [13]. The app collects calibration data by asking participants to look at moving dots on the screen, to enable high eye tracking accuracy comparable to state-of-the-art eye trackers with special hardware [13]. Example screen shots of the user study app are shown in Figure 4. The study was conducted on internal participants with explicit and fully informed consent, and with the option for participants to opt-out of the study or delete the data anytime during or after the study.

Generate gaze saliency The gaze saliency map is computed following the procedure in [10]. More specifically, we first convolve each gaze point with a Gaussian kernel, then take sums of all Gaussian kernels for the gaze points on the image, and normalize by the number of participants. The kernel width is chosen 1/30 of the minimum of image width and height.

First gaze time First gaze time is defined as the time that gaze visits the region of interest for $>$ 50ms. Note that a threshold is needed so that only gaze visit on purpose to the regions within the mask will count, ruling out accidental gaze visit. If there is no gaze within mask, first gaze time will be set as 5s, the duration one image is shown to the user.

4.2 Details of the survey user study

As mentioned in the paper, we conducted a user study that contains two main questions. The goal of the first was to evaluate how natural our results seem, and the second to understand what kind of effects users prefer for the task of saliency reduction. The users were asked to look at various 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?". Figure 5 shows a couple of screenshots from the study.

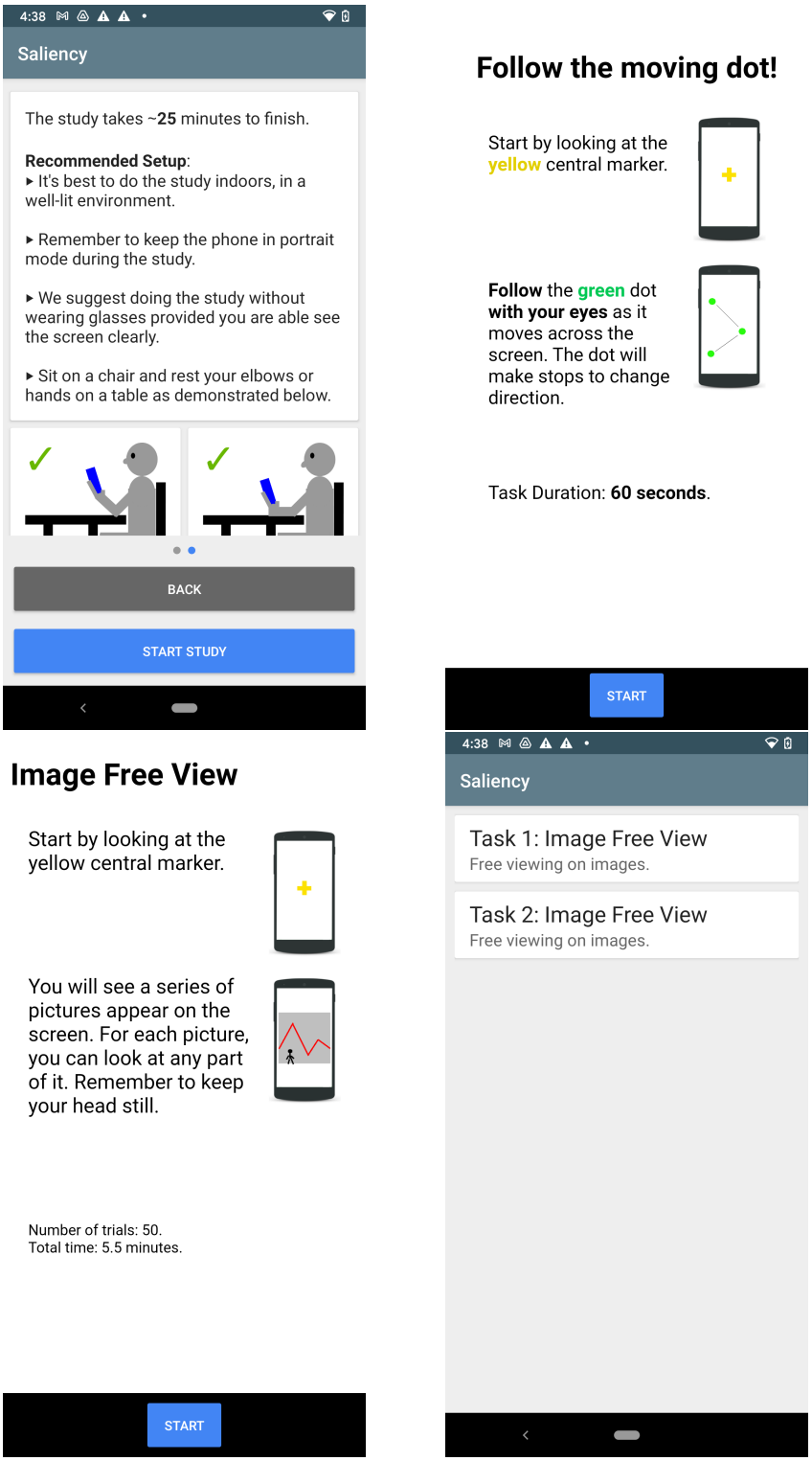

Figure 4: Example screenshots from the eye-gaze user study app.

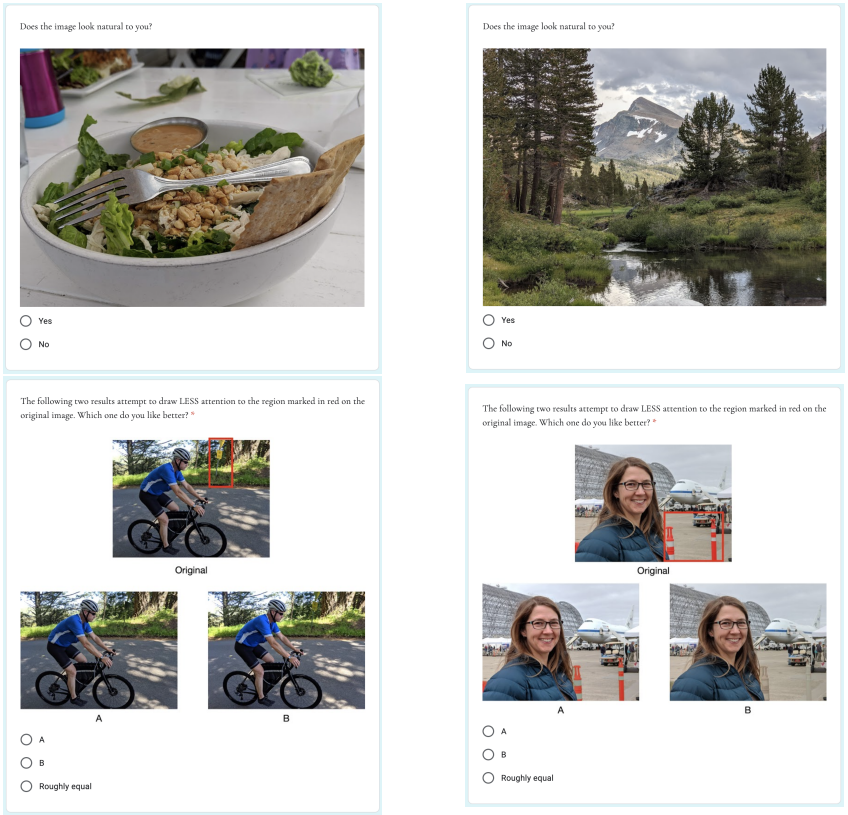

Figure 5: Example screenshots from our survey.

5 Examples for more operators

The proposed method is a general framework which can support other operators besides the four ones discussed in the paper, as long as the operator can constraint the solution space within the valid space of the saliency prediction model.

Another simple yet effective that can be used is a spectral decomposition and re-composition of the region of interest that modifies the saliency within the region. For example, we can project the image into different frequency bands and recompute the weights of each band in the reconstruction stage, such that the overall saliency in the region of interest is modified. We can use the multi-layer Laplacian decomposition introduced in [11], where original image $\mathbf{I}$ is decomposed to as $\mathbf{I} = \mathbf{I}_s + \sum_{k=1,...,K} \mathbf{I}_{r_k}$, where $\mathbf{I}_{r_k} = (\mathbf{1} - L)L^{k-1}\mathbf{I}$ is the k-th residual component, while $I_s = L^K\mathbf{I}$ is the smooth component, and L can be any smoothing or low frequency filters like bilateral filter [12], Gaussian filter, Nonlocal Means filter [1], etc. $\mathbf{1}$ is the identity matrix.

Here, O_θ is a recomposition operator, as

$$\tilde{\mathbf{I}} = \mathbf{I}_s + \sum_{k=1,...,m} \mathbf{I}_{r_k} + \sum_{k=m+1,...,K} \mathbf{I}_{r_k} * W_k, \tag{2}$$

where each weighting map W_k is upsampled from a low-resolution spatial grid θ_k, for $k = m+1, ..., K$. In other words, we will fix the smooth component and first m residual components, and only modify the remaining residual components, where m is a parameter that can be tuned depending how subtle we demand the effects to be. If we apply the above method to a RGB image directly, unnatural colors may be obtained sometimes. To avoid the issue, we can either apply the same W_k on all three R, G, B channels, or convert the image to L-ab format, and apply the decomposition and recomposition on the L-layer only. This operator can be combined with recoloring operator too, i.e, applying recomposition operator in L channel, and recoloring operator in ab channel.

Figure 6 shows a few examples that use the above spectral recomposition to increase saliency, where the operator tends to enhance detail/tone within the region of interest. Even though increasing saliency is not the major focus of this paper, we want to emphasize that the proposed framework is general enough to support different kinds of attention guiding effects with appropriate operators.

6 Ethical Considerations

Our technology focuses on world-positive use cases and applications. Guiding visual attention in images through saliency models has a variety of beneficial and impactful uses, such as removing distractors from photos and video calls, or calling attention to specific areas of a poster or sign to improve the readability and understanding of its content, to name a few. However, we acknowledge the potential for misuse, given the use of generative models to edit images. We emphasize the importance of acting responsibly and taking ownership of

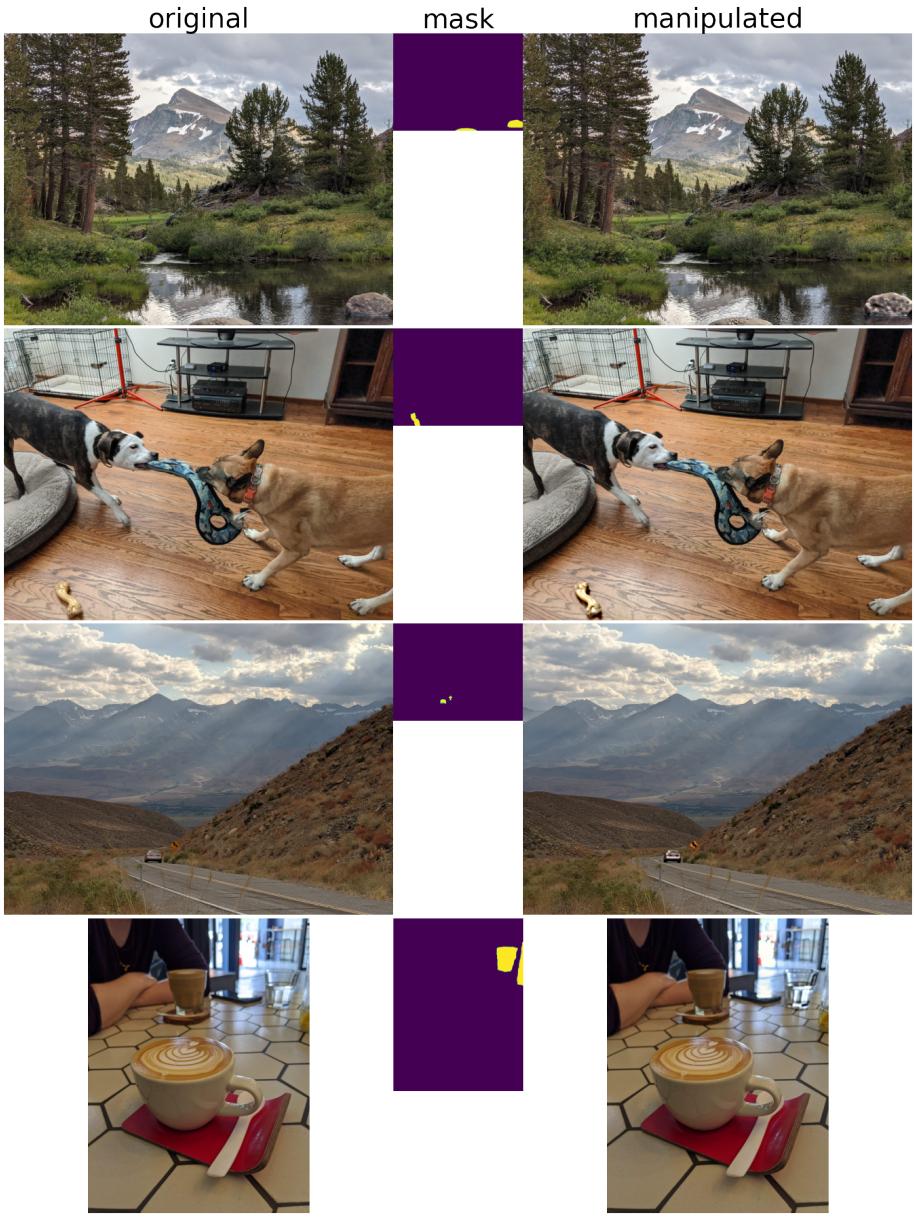

Figure 6: Increased saliency with spectral decomposition.

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.

References

[1] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, 4(2):490–530, 2005.

[2] 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.

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[4] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3012–3021, 2020.

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

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

[7] 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.

[8] 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.

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

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

[11] Hossein Talebi and Peyman Milanfar. Fast multilayer laplacian enhancement. *IEEE Transactions on Computational Imaging*, 2(4):496–509, 2016.

[12] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998.

[13] 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.

[14] 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.

[15] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.