Guide Your Eyes: Learning Image Manipulation under Saliency Guidance

Yen-Chung Chen*[1]
yenc.cs06g@nctu.edu.tw
Keng-Jui Chang*[1]
adplz53.cs06g@nctu.edu.tw
Yi-Hsuan Tsai[2]
ytsai@nec-labs.com
Yu-Chiang Frank Wang[3]
ycwang@ntu.edu.tw
Wei-Chen Chiu[1]
walon@cs.nctu.edu.tw

[1] National Chiao Tung University
[2] NEC Laboratories America
[3] National Taiwan University

Abstract

In this paper, we tackle the problem of saliency-guided image manipulation for adjusting the saliency distribution over image regions. Conventional approaches ordinarily utilize explicit operations on altering the low-level features based on the selected saliency computation. However, it is difficult to generalize such methods for various saliency estimations. To address this issue, we propose a deep learning-based model that bridges between any differentiable saliency estimation methods and a neural network which applies image manipulation. Thus, the manipulation is directly optimized in order to satisfy saliency-guidance. Extensive experiments verify the capacity of our model in saliency-driven image editing and show favorable performance against numerous baselines.

1 Introduction

Saliency estimation, which predicts eye-catching regions over the image for capturing the underlying characteristics of human visual system, has long been an important problem in computer vision and cognitive science. Knowing where in image attracts human attention, which is usually represented as a *saliency map*, is fundamental and beneficial in a wide range of applications, such as object detection and image segmentation. Apart from directly leveraging saliency map as an informative cue for various vision tasks, recently there are works [8, 17, 19, 20, 26] in turn perform image manipulation with being conditioned on the constraints in image saliency, which is referred to as *guiding saliency map* in this paper.

Figure 1 presents one example, where the couple attracts more attention then the giraffe in the original image. Given a guiding saliency map that aims to make the giraffe more eye catching, one targets to modify the original image such that manipulated output satisfies this guiding saliency condition. In real-world applications, one can apply *saliency-guided image manipulation* to many practical scenarios, e.g., human-computer interaction [6], autonomous driving, or advertisement with needs for highlighting the specific regions or objects.

The symbol * indicates equal contribution.

Figure 1: Example for saliency-guided image manipulation. The **(b)** saliency map of **(a)** original image indicates the human couple as the most salient object. Upon being conditioned on **(c)** the guiding saliency map which aims to attend more on giraffe, our proposed method edits the image accordingly to get **(d)** the manipulated output, with its corresponding saliency map shown in **(e)**. Note here that we visualize the saliency map (which is single channel) by stacking it upon color images in order to provide their spatial correspondence.

The existing related works actually have large dependency on the corresponding algorithms of saliency estimation. To be detailed, these works need to first fully understand the properties (e.g. which feature cues are utilized) of the saliency estimation algorithm, and then explicitly design the closely-related objectives to manipulate the image output. However, this requirement limits the flexibility of using different saliency estimation approaches within the same framework. Furthermore, as saliency estimation could aggregate multiple features in a bottom-up manner, the relationship between various features might be quite complicated thus make it hard to manually derive a proper objective for manipulation.

In this paper, we propose a learning-based model which seamlessly combines the image manipulation and saliency estimation into a unified framework, and accordingly resolves the limitation described above. We leverage two main ideas in the proposed model. First, we choose to use the deep-learning-based saliency estimation approach, where both the feature extraction and final saliency prediction are learned jointly from data. Compared to conventional methods that rely on hand-crafted features, we take advantages of the differentiable property of neural networks to gain prior knowledge on how the saliency is predicted through back-propagation. We also note that, the proposed method is not tied to any specific saliency estimation framework but supports arbitrary off-the-shelf architectures once they are end-to-end differentiable. Second, the proposed manipulation network learns to take an image and a guiding saliency map as the input to generate the manipulated output. In particular, the output image should preserve the content of original image, be realistic, and have its saliency map (estimated by the saliency estimation network) matched with the guiding one.

We evaluate the proposed model on the MS-COCO dataset [15], make qualitative and quantitative comparison to numerous baselines under various scenarios, and demonstrate favorable performance against state-of-the-art algorithms. In addition, we adapt our method to perform memorability-guided image manipulation, where the image is edited to be likely or unlikely more memorable according to the guided memorability measurement [10]. The extension shows the potential usage and generalizability of our model across different tasks.

2 Related Works

Saliency-Guidance Image Manipulation. As described previously, most of the existing research works [6, 17, 18, 19, 20, 26] in saliency-guidance image manipulation require first discovering the feature cues used in saliency estimation, in which these features are used to perform image editing. In other words, actually the saliency estimation and image manipulation parts are two individual steps for these algorithms. In [6], the saliency map is computed

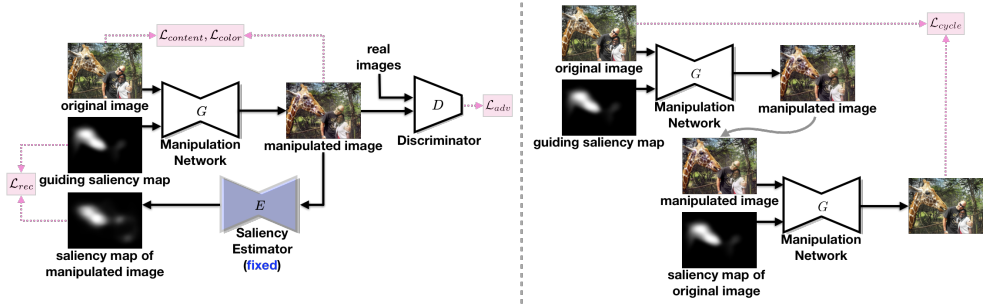

Figure 2: Overview of our Saliency-Guidance Image Manipulation (SaGIM) model. The manipulation network takes an image and a guiding saliency map as input, and produces a realistic manipulated output that has the saliency map consistent with the guiding one. The right side of figure visualizes the cycle consistency uniquely introduced in our model.

by using intensity and color features, and the authors discover that the point variation, which accounts for the degree of visual features changes in a local image area, determines how much a feature impacts the salience of a certain location in an image. The point variation is therefore used to guide the image manipulation. [17] utilizes a similar idea, while they keep both chromaticity and intensity unchanged, but manipulate the image to maximize the dissimilarity of hue distribution of target area from the neighborhood. [26] increases average luminance, color saturation, and sharpness of the target region to enhance its salience.

The most recent work from [19] first extracts from the input image to form two groups of image patches with high and low saliency, and then edits the image such that the target region reaches similar high-saliency patches in color channels while non-target regions are closer to low-saliency ones, respectively. However, this approach is still based on a predefined feature to drive the manipulation. More detailed review of related works can be found in [18].

Deep-Learning-Based Saliency Estimation. We review some of the works in supervised deep-learning-based saliency estimation. [12] utilizes the AlexNet architecture [11] pre-trained on Imagenet dataset [1], and learns to linearly combine the feature maps across network layers in order to obtain the prediction of saliency map. [16] divides images under different resolutions into patches and train a convolutional neural network for classifying the fixation and non-fixation image patches, in which during the testing time the saliency is estimated on the patch level. [14] proposes a two-stream network to combine the pixel-level saliency map with the superpixel-wise features that better model the discontinuities along object boundaries in the final saliency prediction. [21] directly uses a fully convolutional network to map the input image into saliency map estimation, while employing the adversarial loss, which is originally proposed in generative adversarial networks [5], in order to improve the quality of output saliency map and make it more realistic.

3 Proposed Method

The objective of our proposed method, *Saliency-Guidance Image Manipulation* (**SaGIM**), is to edit an input image such that the saliency estimation of manipulated output agrees with a given saliency-guidance (as depicted in Figure 2). Our SaGIM model consists of 3 major components: manipulation network G, saliency estimation network E, and discriminator D, in which we are going to detail in the following subsections together with the loss functions. Note that the details of network architecture are provided in the supplementary material.

3.1 Saliency Estimator

We use the state-of-the-art SalGAN [21] as our saliency estimation network E, which is pre-trained beforehand and kept fixed (not updated) during the learning of proposed model. We follow SALICON [7] to use a large-scale dataset for benchmarking visual saliency prediction as the training/validation data for the saliency estimation network E. In addition, as the images used in SALICON are collected from the MS-COCO dataset [15], which is exactly the same dataset we carry out experiments, the potential issue of domain-shift for saliency estimation is eliminated. Please note again that our model is not limited to SalGAN but supports any differentiable saliency estimation approaches.

3.2 Manipulation Network

Let $I \in \mathbb{R}^{H\times W\times 3}$ be an input image and $S^{guiding} \in \mathbb{R}^{H\times W\times 1}$ be a guiding saliency map where all the values in a saliency map are within the interval $[0,1]$, indicating the pixel-wise salience. The manipulation network G takes both the image I and guided saliency map $S^{guiding}$ as input and maps them into a manipulated output image $\tilde{I} = G(I, S^{guiding})$.

Reconstruction loss. The manipulated output $\tilde{I}$ ideally should have the saliency map which is consistent with the guiding saliency map $S^{guiding}$. Let S^{edited} be the saliency map of $\tilde{I}$ that is predicted by saliency estimation network E, i.e. $S^{edited} = E(\tilde{I})$, we penalize the difference between S^{edited} and the corresponding guiding saliency map $S^{guiding}$ based on the averaged binary cross entropy (BCE) loss over all pixels, as suggested in [21]:

$$\mathcal{L}_{rec} = -\frac{1}{N}\sum_{i,j} S^{guiding}_{i,j}\log(S^{edited}_{i,j}) + (1 - S^{guiding}_{i,j})(1 - \log(S^{edited}_{i,j}), \tag{1}$$

where i, j index pixel positions and $N = W \times H$ is the total number of pixels.

Content Loss. The image manipulation solely based on the constraints from saliency-guidance might disrupt the structure of the original image, which is undesirable. In order to preserve the entire structure and content of the input image, we impose the content loss $\mathcal{L}_{content}$ as utilized in the work of neural style transfer [4] (similar idea can be also found in the perceptual loss proposed by [8]). Basically, we penalize the mean squared error (MSE) between the deep features extracted from I and $\tilde{I}$ respectively. In our model, we take feature maps of the VGG network [25] as the deep features.

$$\mathcal{L}_{content} = \frac{1}{N_l}\sum_{l} \text{MSE}(\phi^l(\tilde{I}), \phi^l(I)), \tag{2}$$

where $\phi^l(\cdot)$ denotes the feature representation obtained from the l-th layer of VGG network, and the total number of VGG layers considered in this content loss N_l is set to 5.

Color Loss. Moreover, in order to preclude drastic changes on the color tone, especially on the regions that the guiding saliency map aims to enhance, we add another color loss term $\mathcal{L}_{color}$. It encourages the color consistency between corresponding local regions from original image I and its manipulated output $\tilde{I}$. Let $M(I)$ denote the response map obtained by applying Gaussian filter on each color channel of an image I, which derives averaged colors locally, the color loss $\mathcal{L}_{color}$ is defined as:

$$\mathcal{L}_{color} = \frac{1}{N}\sum S^{guiding} * |M(I) - M(\tilde{I})|, \tag{3}$$

where $*$ is element-wise multiplication, and the size of Gaussian filter is set to 21.

Cycle Consistency Loss. We further adopt the idea of cycle consistency proposed in [28] to not only enforce the stability of our manipulation network G but also benefit to

boost the overall performance without seeing additional images. Basically, as illustrated in the right half of Figure 2, after having manipulated output image $\tilde{I} = G(I, S^{guiding})$ based on the original input image I and guiding saliency map $S^{guiding}$, we can use G to map the input pair of $\{\tilde{I}, E(I)\}$ to a new output image $\check{I}$, where now the saliency map $E(I)$ of original image is treated as the guiding saliency for $\tilde{I}$. This procedure is analogous to an inverse mapping for performing de-manipulation on $\tilde{I}$ (i.e., recover the original image by using its saliency map as guidance), therefore the new output image $\check{I}$ should be similar to the original input image I. We utilize the same metric used in the content loss $\mathcal{L}_{content}$ to measure the distance between $\check{I}$ and I, where the cycle consistency loss is formulated as:

$$\mathcal{L}_{cycle} = \frac{1}{N_l}\sum_l \text{MSE}(\phi^l(I), \phi^l(\check{I})) = \frac{1}{N_l}\sum_l \text{MSE}(\phi^l(\tilde{I}), \phi^l(G(\tilde{I}, E(I)))) \tag{4}$$

3.3 Discriminator

Inspired by the adversarial learning scheme [5], we adopt the adversarial loss function $\mathcal{L}_{adv}$ to improve the quality of the manipulated images and make them more realistic, such that the data distribution $P_{\tilde{I}}$ of manipulated outputs is close to the one P_I of real images. The objective is formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}_{I\sim P_I} log D(I) + \mathbb{E}_{\tilde{I}\sim P_{\tilde{I}}} log(1 - D(\tilde{I})), \tag{5}$$

where discriminator D distinguishes between real images and manipulated ones. In adversarial learning, we minimize $\mathcal{L}_{adv}$ w.r.t. D while maximizing the second term to update our manipulation network G, where $\tilde{I}$ is produced by $G(I, S^{guiding})$. Please note that, although our overall framework is alike to bidirectional-GAN [2], we do not consider the joint distribution over images and saliency maps as the input to D, since the guiding saliency map used in our experiments is manually defined and thus it is different from the real ones.

3.4 Total Loss

Overall, the total objective of our SaGIM model is the sum of the aforementioned loss terms:

$$\mathcal{L}(\theta_G, \theta_D) = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{content} + \lambda_3\mathcal{L}_{color} + \lambda_4\mathcal{L}_{cycle} + \lambda_5\mathcal{L}_{adv}, \tag{6}$$

where θ_G and θ_D are the network parameters of manipulation network G and discriminator D. Again, we note that the saliency estimation network E is pre-trained and stays fixed in our SaGIM model. The hyperparameters λ control the balance between each loss function and are set to be $\lambda_1 : \lambda_2 : \lambda_3 : \lambda_4 : \lambda_5 = 5 : 5 : 1 : 9 : 5$ in our experiments. We use Adam optimizer with learning rate of 10^{-3} and train for 100 epochs.

4 Experiments

In this section, we describe various experimental settings and results for evaluating the performance of our proposed method. We not only compare our model with respect to several baselines for shifting the saliency distribution but also perform analysis from the perspective of adversarial attacks. Finally, we show an extension of our method into the task of memorability-guided image manipulation.

4.1 Data Preparation

Dataset. Based on the training and validation sets of MS-COCO [15], we sample 15,686 images to construct a SAliency Manipulation dataset (**SAM**) used in our experiments. Every sampled image contains more than 2 objects, and each object covers 10% to 70% of area of

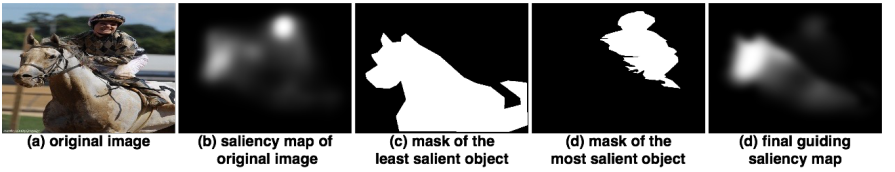

Figure 3: Example of generating guiding saliency map. The most salient object is changed from the rider to the horse.

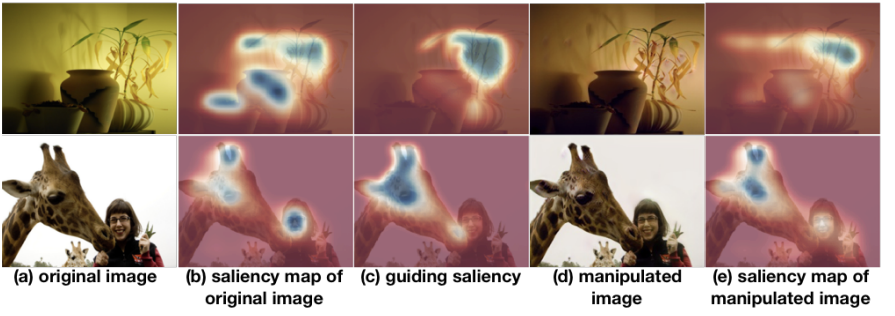

Figure 4: Example results of saliency-guided image manipulation by our SaGIM model.

entire image, in order to obviate tiny or overwhelming ones. We partition our dataset into training and testing sets of $4,000$ and $11,686$ images respectively. We note that, although the SalGAN is pre-trained on SALICON [7] which could potentially overlap with our SAM dataset, it is kept fixed as an off-the-shelf saliency estimator in our model training. In addition, the images fed to SalGAN in our model are already manipulated ones, which is already different from SALICON in appearance. Our SAM dataset, source code, and models are released at https://github.com/YenchungChen/GuideYourEyes

Guiding Saliency Maps. The construction of the corresponding *guiding saliency maps* for our dataset is based on the following procedure. First, we use SalGAN to estimate the saliency maps for all images, and compute the average saliency of each object hinge upon its object mask provided by MS-COCO annotations. Second, for makinig changes on the distribution of saliency, we increase the saliency of the least salient object and decrease the saliency of the most salient object respectively by random factors which can lead to re-ordering of objects' saliency. Last, we normalize the modified saliency map into $[0,1]$ and apply Gaussian filter to smooth out sharp edges. An example is demonstrated in Figure 3.

4.2 Saliency Manipulation by Guiding Saliency

Our SaGIM model is trained and tested on the SAM dataset. Figure 4 shows some example results of our method for performing saliency-guided image manipulation. We can observe that the original images are mapped into the manipulated outputs with their saliency maps satisfying the given guiding saliency maps. Furthermore, our model learns to utilize different manipulation operations to produce the required changes in saliency, as two examples visualized in Figure 5. This verifies the advantage of our proposed method w.r.t. related works (e.g., [6, 17, 19, 26]) that we do not need to specify a certain feature cue for manipulation, and thus our model is more general.

User Study. The guiding saliency maps in our SAM data are generated to have changes on the ordering among the saliency of object instances and we perform a user study to evaluate the performance of the proposed method. We select from the testing set with 50 pairs

Figure 5: Two examples for visualizing different manipulation operations used in our model. In each example, the first and second columns show the original image and the manipulated one, while the third and forth columns provide the zoom-in views of the annotated regions in the first two columns respectively. We can see that the man on the left gets less salient by being blurred while the shoe on the right gets more salient by having higher saturation.

	Guidance Consistency (in percentage)				Image Captioning		Object Detection
	DeepGaze w/ C1	SalGAN w/ C1	DeepGaze w/ C2	SalGAN w/ C2	BLEU	C3	MSE
Ours	**23.1**	**70.2**	36.5	**86.9**	**0.309**	19.0	0.06
OHA	15.7	12.0	32.9	42.7	0.310	17.4	0.12
HAG	18.9	11.3	34.9	47.8	0.322	**19.4**	**0.04**
WSR	17.4	10.6	**36.7**	40.6	0.319	18.8	0.07

Table 1: Quantitative evaluations of our proposed method with respect to several baselines in various schemes.

of an original image and its corresponding manipulated output, which is produced by our SaGIM model. We construct a questionnaire that consists of 3 questions for each image pair: (1) which is the most salient object in the original image. (2) which is the most salient object in the manipulated output (where several candidate image regions are given for selection in first two questions) and (3) Do you perceive that the most salient objects on two images are different ones? Our user study includes in total 24 participants with roughly equal proportion of females and males, and we obtain the statistics as follows. The results show that the most salient objects are accurately selected by the participants for 63.50% of original images and 58.92% of manipulated outputs. Additionally, conditioned on the case of answering correctly for original images, 62.25% of questions for manipulated images are simultaneously correctly answered. On the other hand, users see that on average 58.17% of image pairs have the most salient object varying across original and manipulated images. The high consistency of SalGAN w.r.t. human perception, cf. question (1), verifies the design choice of taking it as our saliency estimation network. Most importantly, the results show that our saliency-guided image manipulation does change the saliency distribution and match guiding saliency map to a certain extent.

4.3 Quantitative Comparisons

Saliency Enhancement based on Object Masks. Based on our SAM dataset, here we have quantitative comparisons w.r.t. several baseline methods, including HAG [6], WSR [26], and OHA [17], which are identified as top performers in [18]. In order to have fair comparisons, we take *binary guiding saliency maps* as used in these approaches, for applying saliency-driven manipulation. In each image, we denote the most salient object as O_{high} and the least one as O_{low}, and the binary guiding saliency map is exactly the object mask of O_{low}, which indicates that image manipulation is to simultaneously enhance O_{low} and de-emphasis O_{high}.

Here, we define a **Guidance Consistency** metric with two criteria **C1, C2** for evaluation: **(C1)** a manipulation is effective when the average saliency of O_{low} is higher than the one of O_{high} in the manipulated output; **(C2)** a manipulation is effective when the average saliency of O_{low}/O_{high} in the manipulated output is higher/lower than the one of O_{low}/O_{high} in original input image. It is worth noting that C1 is a stricter criterion than C2. Furthermore, as our model is optimized for SalGAN, in addition to using SalGAN for saliency computation on

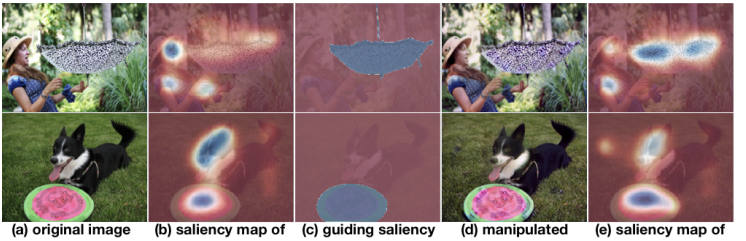

Figure 6: Example results of our SaGIM model with using binary guiding saliency maps.

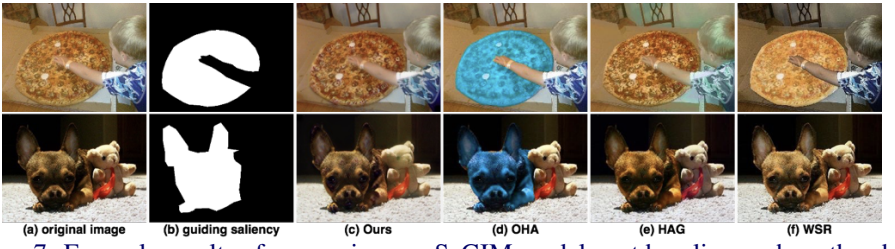

Figure 7: Example results of comparing our SaGIM model w.r.t baselines, where the object mask of O_{low} (the least salient object in the original image) is now taken as guiding saliency.

O_{low} and O_{high}, we further introduce DeepGaze [13] as an unbiased saliency estimator.

The quantitative results shown in Table 1 demonstrate that our proposed method has better or comparative performance w.r.t. baselines in various evaluation settings. Some example results based on our SaGIM are provided in Figure 6, while Figure 7 visualizes comparisons on qualitative results between our method and the baselines. We observe that OHA [17] and WSR [26] usually add unnatural color and contrast to the image, while HAG [6] can not perform saliency enhancement and reduction simultaneously.

As we can observe from the examples in Figure 1, 4, 5, 6, and 7, the image modifications happen mostly on the local salient regions of original image and guided saliency map, thus the manipulated output is not globally different from its original image and still with similar content. Therefore, we consider that the proposed framework could be treated as a way of finding *adversarial examples*, where it tries to keep the structure/content of the input image ($\mathcal{L}_{content}$) but now the objective of attack is not a specific classification posterior but instead guided by the saliency estimation ($\mathcal{L}_{rec}$), which is similar to the targeted attack scenario. To be more detailed, our model tackles not only the targeted attack but also conditional generation of adversarial examples based on the proposed manipulation network. Tackling these two difficulties is novel in adversary attack area, especially that our target network (i.e. saliency estimator) produces higher dimensional output than simple image classification, and thus it is much harder to have successful attack. Here we propose to perform two quantitative evaluations from the perspective of adversarial attack.

Adversarial Attack on Image Captioning. The image captioning network in [21] utilizes the attention mechanism which we hypothesize to have correlation with saliency such that the saliency-guided image manipulation would result in an attack to change the output of captioning. We first evaluate the difference between generated captions from input image and its manipulated output, based on BLEU [22], a widely-used evaluation metric for machine translation. In addition, we define another metric **C3** (shown in percentage), such that a

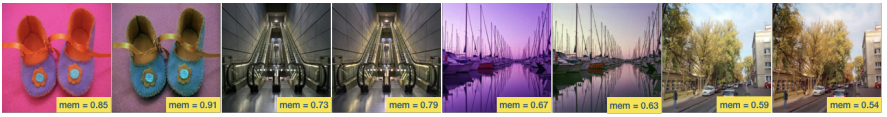

Figure 8: Examples of memorability-guided manipulation with corresponding memorability values in yellow boxes.

Guidance Consistency (%)	**Ours**	w/o $\mathcal{L}_{cycle}$	w/o $\mathcal{L}_{adv}$	w/o $\mathcal{L}_{color}$
SalGAN w/ C1	**45.0**	37.4	24.3	25.3
SalGAN w/ C2	**66.4**	63.3	61.7	62.6

Table 2: Guidance Consistency performance for different variants of our SaGIM model.

successful attack happens when the caption of the manipulated output must simultaneously exclude O_{high} and include O_{low} of the original image. Table 1 shows that our method achieves better or competitive attack for both metrics (larger the better).

Adversarial Attack on Object Detection. Furthermore, we hypothesize that altering object saliency would also affect the results of object detection. Thus an evaluation of adversarial attack on object detector (YOLOv2 [24]) is performed to measure that objects should be either mislabeled or have confidence changes consistent with guiding saliency. We note that our method obtains similar success rates as baselines but with almost minimum perturbations on the image (measured by MSE as shown in Table 1).

4.4 Ablation study

We investigate the influences of different objectives in the proposed model based on the normalized saliency evaluation, and the results are shown in Table 2. Note that, we test the model variants without cycle consistency, adversarial, color losses, while both reconstruction and content losses are kept since they are the keys to fit the guiding saliency and maintain the image structure, respectively. Here are two observations that support our design of loss functions: (1) The inverse mapping procedure utilized in cycle consistency loss takes both the manipulated output and the saliency map of original image as input, therefore it enriches the data distribution that manipulation network sees. Removing it from our full model reduces the performance significantly. (2) Lacking of adversarial or color losses could unfavorably allows the manipulation network adding some unrealistic artifacts or having drastic color shift on the output image, which might be applicable to impact the saliency during training but not generalized well for test images. We provide some qualitative examples in the supplementary material.

4.5 Extension to Image Memorability

Numerous researches (e.g., [8, 9, 10, 23]) have been devoted to estimate the memorability of images. Here we extend our framework to the task of memorability-guided image manipulation, which is to manipulate the input image based on a preferable memorability score. This is achieved by replacing the input guiding saliency map and saliency estimator E by a guiding memorability value and the memorability estimator proposed in [10], respectively. We experiment on the LaMem dataset [10] and manipulate the images to have higher or lower memorability values than their original ones. Some example results are shown in Figure 8. In addition, with comparisons to the saliency map of an original image with respect to its difference from corresponding manipulated output, we find that the pixels with bigger difference are most likely located on the salient regions. This can be related to the observation

described in [10], where the pattern of human fixations on an image has a positive correlation with its memorability. Although this interesting fact is now out of the scope/focus of our work in this paper, we would like to have a further investigation as a future work.

5 Conclusions

We present a deep learning-based framework for tackling the task of saliency-guidance image manipulation. Our SaGIM model coordinates the image manipulation and saliency estimation into a unified framework and thus enables end-to-end optimization for learning to revise the input image conditioned on a guiding saliency map. We conduct comprehensive experiments and show that our method successfully achieves the target changes in saliency of the manipulated output, outperforming a series of baseline approaches in evaluation schemes such as adversarial attacks in object detection and image captioning, as well as supporting memorability-guided image editing.

Acknowledgement. This project is supported by the Ministry of Science and Technology of Taiwan under grant MOST-108-2636-E-009-001, MOST-108-2634-F-009-007, and MOST-108-2634-F-009-013, and we are grateful to the National Center for High-performance Computing of Taiwan for computer time and facilities.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv:1605.09782*, 2016.

[3] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. *arXiv:1804.03115*, 2018.

[4] 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 (CVPR)*, 2016.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[6] 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*, 2011.

[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 (CVPR)*, June 2015.

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[9] Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Memorability of image regions. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[10] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

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

[13] 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 (ICCV)*, 2017.

[14] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[16] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[17] 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*, 2014.

[18] Victor A Mateescu and Ivan V Bajic. Visual attention retargeting. *IEEE Transactions on Multimedia (TMM)*, 23(1):82–91, 2016.

[19] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Saliency driven image manipulation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[20] Tam V Nguyen, Bingbing Ni, Hairong Liu, Wei Xia, Jiebo Luo, Mohan Kankanhalli, and Shuicheng Yan. Image re-attentionizing. *IEEE Transactions on Multimedia (TMM)*, 15(8):1910–1919, 2013.

[21] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Cristian Canton Ferrer, Jordi Torres, Kevin McGuinness, and Noel E OâĂŹConnor. Salgan: Visual saliency prediction with adversarial networks. In *CVPR Scene Understanding Workshop*, 2017.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, 2002.

[23] Patrik Polatsek, Manuela Waldner, Ivan Viola, Peter Kapec, and Wanda Benesova. Exploring visual attention and saliency modeling for task-based visual analysis. *Computers & Graphics*, 72:26–38, 2018.

[24] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv:1612.08242*, 2017.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[26] Lai-Kuan Wong and Kok-Lim Low. Saliency retargeting: An approach to enhance image aesthetics. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2011.

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.