Saliency-Guided Image Translation

Lai Jiang[1,2], Mai Xu[1]*, Xiaofei Wang[1], Leonid Sigal[2]
[1] School of Electronic and Information Engineering, Beihang University, Beijing, China
[2] Department of Computer Science, University of British Columbia, Vancouver, BC Canada
{jianglai.china, MaiXu, xfwang}@buaa.edu.cn, lsigal@cs.ubc.ca

Abstract

In this paper, we propose a novel task for saliency-guided image translation, with the goal of image-to-image translation conditioned on the user specified saliency map. To address this problem, we develop a novel Generative Adversarial Network (GAN)-based model, called SalG-GAN. Given the original image and target saliency map, SalG-GAN can generate a translated image that satisfies the target saliency map. In SalG-GAN, a disentangled representation framework is proposed to encourage the model to learn diverse translations for the same target saliency condition. A saliency-based attention module is introduced as a special attention mechanism for facilitating the developed structures of saliency-guided generator, saliency cue encoder and saliency-guided global and local discriminators. Furthermore, we build a synthetic dataset and a real-world dataset with labeled visual attention for training and evaluating our SalG-GAN. The experimental results over both datasets verify the effectiveness of our model for saliency-guided image translation.

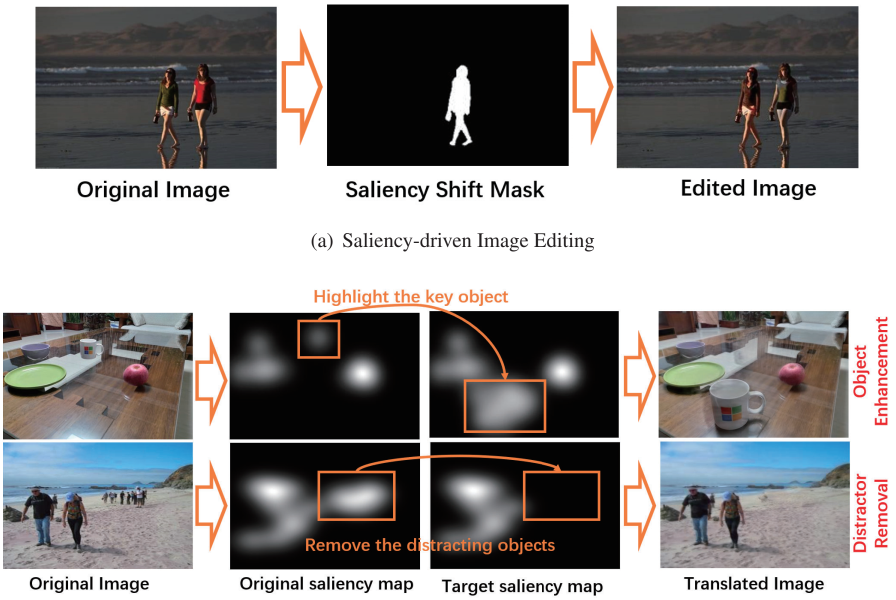
1. Introduction

Conditional image generation has gained significant attention in recent years, especially in light of the progress in Generative Adversarial Network (GAN)-based, and, to a lesser extent, Variational Auto Encoder (VAE)-based generative methods. Impressive results have been achieved in generating high-quality images from different (conditioning) information such as text [6, 11], sketches [13], layouts [38], facial attributes [5, 34] and scene graphs [18]. Image-to-image translation [13, 20, 33, 40] has been a particularly successful sub-class of these methods. Image-to-image translation focuses on producing images that are structurally similar to the original inputs but deviate in stylistic [20] or texture detail [33]. This allows models such as CycleGAN [40] and alternatives to produce images of zebras from horses, or Picasso painting renditions from everyday photographs. More recent models [1] also provide ability to modify the image more structurally by, for example, adding specific objects. This of course requires a user to select and place an object in a desired location. None of these methods, however, allow for the ability to model more abstract translations or image modifications that alter the way in which the original image is *perceived*.

Consider someone taking a photo of a person outdoors. In addition to the person, the image may contain other background or foreground objects (*e.g.*, cars, motorcycle) that distract attention of the viewer. How can this be mitigated in “post-production”? Many techniques and strategies can be employed. For example, distracting objects may simply be removed, using image inpainting techniques [36] on the object regions. Alternatively, good bokeh (good quality blur) could be computationally applied to all pixels but those belonging to the main subject, effectively modeling shallow depth of field which is a common technique in professional photography. Note that a bad bokeh (a distracting blur) may actually have an adverse effect toward the desired goal. Further, a color pallet of either distracting objects or the subject itself maybe altered to make the subject more distinctive. These are just some of the multitude of ways that an image maybe altered to achieve a desired effect. Lets consider what all of these strategies have in common, in effect they are trying to modify the saliency distribution of the input image, by modifying image itself, to achieve a certain visual effect. We posit that ability to manipulate an image to achieve a desired saliency distribution is a core task for a variety of high-level applications including image retargeting [26], object enhancement [27], distractor removal [7] and intelligent advertisement [31]. To this end we propose a novel task of *saliency-guided image translation* and corresponding benchmark datasets.

The goal of *saliency-guided image translation* is to perform image-to-image translation conditioned on the (user specified) *target* image saliency map. Some examples of *saliency-guided image translation* are shown in Figure 1. Despite long history of saliency in computer vision [14], few approaches exist that carry ability to perform saliency-driven image adjustments; most focus on

[*] Corresponding author.



(a) Saliency-driven Image Editing

(b) Saliency-Guided Image Trnslation

Figure 1. In traditional **saliency-driven image editing**, the modification is pixel aligned; while for our **saliency-guided image translation**, the composition of the image itself can be changed, allowing spatial transformations or shifts, addition, removal of objects as a whole. Meanwhile, instead of an accurate mask, our image translation method is directly guided by the fixation map, which can be easily acquired by mouse-contingent tool or eye-tracker. In (b), we present two potential applications of saliency-guided image translation: object enhancement (first row) and object removal (second row). In the first row, the mug in the original image attracts little human attention, mainly because it is far away from the camera. We can make the mug more focal by a suitable target saliency map. The saliency-guided translated image can be seen on the right. Similarly, in the second row, the distracting objects can be removed by giving the target saliency map. Note that the results in (a) and (b) are from [27] and our proposed method, respectively.

saliency prediction. Saliency-driven image editing methods [4, 7, 27, 31, 35, 8, 28], that come closest, are a special case of the proposed, and much more broadly defined, *saliency-guided image translation*. Saliency driven image manipulation approaches are limited to low-level pixel modifications such as color, luminance, saturation and sharpness; while our task also allows for object removal, creation and even motion within the image. As shown in Figure 1, saliency-driven image editing methods are limited to low-level pixel modifications such as color, luminance, saturation and sharpness; while our task also allows for object removal, creation and even motion within the image. while our task also allows for object removal, creation and even motion within the image. Meanwhile, instead of an accurate mask, our image translation method is directly guided by the fixation map, which can be easily acquired by mouse-contingent tool or eye-tracker. Thus, beyond saliency-driven image editing, saliency-guided image translation offers more flexible and vast potential real-world applications, such as the go-to tools for product designers, market researchers and consumer behavior modeling, including in advertising.

Compared to traditional image-to-image translation tasks, the *saliency-guided image translation* is much more challenging. Impoverished content and ambiguity of the saliency are the core challenges. For example, saliency is object and content agnostic, meaning same added level of saliency in a given image location can be achieved by inserting a variety of objects that adhere to the correct proportions. Also, there are multiple conceptual solutions that can satisfy the same saliency change. For example, the saliency of a object can be enhanced by changing its appearance or removing other salient objects around it. Last, the saliency of same object can be different across images, due to the influence of surrounding objects. Thus, the models for *saliency-guided image translation* should be inherently capable of both generating real images and understanding of human attention and how it can be manipulated.

Contributions: In this paper, we take the first step towards the saliency-guided image translation, by proposing a novel GAN-based model, namely SalG-GAN. To address the challenge of saliency ambiguity, *a disentangled representation framework* is developed in SalG-GAN, in order to encourage the model to learn diverse translations for the same target saliency map. Besides, a *saliency-based attention module* is introduced as a special attention mechanism for facilitating the developed structures of *saliency-guided generator*, *saliency cue encoder* and *saliency-guided global and local discriminators*. Additionally, a light but effective *saliency detector* is developed as part of the framework, to help the generator understand and modify human attention. For training and testing our SalG-GAN, we build a synthetic dataset (SGIT-S) consisting of 53,000 images and a real-world dataset (SGIT-R). Both datasets are labeled by 7 subjects with attention; datasets will be released. The experiments over these two datasets show the effectiveness of our method for saliency-guided image generation.

2. Related work

Saliency-driven Image Editing. Many existing image editing tasks [4, 9, 27, 31, 35, 8, 28] use saliency cues as guidance. For instance, Nguyen *et al.* [31] proposed a Markov Random Field (MRF) based method for retargeting the human attention to certain parts in an image, by recoloring surrounding super-pixels. In [35], Wong *et al.* improved image aesthetic by modifying low-level properties of the visually dominant subjects. Similarly, advanced image editing algorithms were developed in [29] and [2], to direct human attention to the advertisements/important information in Mixed Reality and computer games, respectively. More recently, Mejjati *et al.* [28] proposed a practical image editing pipeline for increasing or attenuating attention in an image region, based on a encoder-decoder network. However, all the above methods are pixel aligned. They mainly focus on pixel-by-pixel manipulation of the saliency related properties (such as color, luminance and sharpness) of a certain object/region. In contrast, for saliency-guided image translation, the composition of the image itself can be changed, allowing spatial transformations or shifts, addition, removal of objects as a whole.

Conditional Image Generation. Recently, conditional image generation methods have shown great success in generating high-quality images from different conditions such as text [11, 6], sketches [13], layout [38], facial attribute [5, 34] and scene graph [18]. However, these works can not be simply applied for saliency-guided image translation, due to the saliency ambiguity. For example, the same salient region of an expected saliency map could be occupied by a different object, as long as that object can draw similar level of saliency. Also, there are multiple solutions to satisfy the same saliency change, *e.g.*, the saliency of a object can be enhanced by changing its appearance or removing other salient objects around it. Besides, the saliency of the same object can be different across images, influenced by other surrounding objects. Therefore, the models for saliency-guided image translation should be able to both generate real images and understand human attention.

3. Methodology

3.1. Framework of SalG-GAN

The overall training pipeline of our proposed SalG-GAN is illustrated in Figure 2. As shown in the figure, SalG-GAN consists of five components: (i) *saliency-based attention module*, (ii) *saliency-guided generator*, (iii) *saliency cue encoder*, (iv) *saliency detector*, and (v) *saliency-guided global and local discriminators*.

Specifically, given the original image $\boldsymbol{X}$ and a target saliency map $\boldsymbol{S_y}$, our goal is to learn a model that can generate the translated image $\hat{\boldsymbol{Y}}$ with saliency $\boldsymbol{S_y}$. First, based on the original saliency map[1] $\boldsymbol{S_x}$ and target saliency map $\boldsymbol{S_y}$, additive attention map $\boldsymbol{S_p}$ and subtractive attention map $\boldsymbol{S_m}$ are extracted from *saliency-based attention module*. The two maps, separately, capture where attention is required to increase/decrease with respect to the source image. Second, taking $\boldsymbol{S_p}$, $\boldsymbol{S_m}$ and the original image $\boldsymbol{X}$ as inputs, *saliency-guided generator* generates the fake image $\hat{\boldsymbol{Y}}$ through a residual learning scheme, to fool the *saliency-guided global and local discriminators*.

Furthermore, due to the ambiguity of the saliency, there are multiple conceptual ways that can satisfy the same target saliency map. To this end, in addition to the attention maps, we also embed the latent saliency cue $\boldsymbol{z} \in \mathcal{Z}$ as the input for the generator to output diverse saliency-guided images that satisfy the same target saliency map $\boldsymbol{S_y}$. $\mathcal{Z}$ is the latent space of all saliency cues. In order to disentangle the representation of latent saliency cues, our SalG-GAN is developed in a supervised and an unsupervised paths, which are introduced as follows. Notably both paths share all five developed components. Similar ideas for disentangling representation can be found in [41, 24].

Supervised Path. For the supervised path, the latent saliency cue $\boldsymbol{z_s}$ is sampled from the posterior distribution $Q(\boldsymbol{z_s}|\boldsymbol{Y})$, which is estimated from our *saliency cue encoder* applied to the ground-truth image $\boldsymbol{Y}$. By encouraging the translated image $\hat{\boldsymbol{Y}}_{\boldsymbol{s}}$ to reconstruct the ground-truth image, the network can learn how to encode latent saliency cue codes effectively. Further, $Q(\boldsymbol{z_s}|\boldsymbol{Y})$ is regularized to approach standard normal distribution $\mathcal{N}(0, 1)$ in the supervised path, in order to perform sampling at test time.

Unsupervised Path. For the unsupervised path, the latent saliency cue $\boldsymbol{z_u}$ is sampled from a normal prior distribution $\mathcal{N}(0, 1)$. In order to emphasize the role of saliency cue during image generation, $\boldsymbol{z_u}$ should be re-predicted by the *saliency cue encoder*, from the translated image $\hat{\boldsymbol{Y}}_{\boldsymbol{u}}$. This would help generate diverse results by mitigating many-to-one mapping problem, which is consistent with [38]. Further, since there is no ground-truth image to supervise $\hat{\boldsymbol{Y}}_{\boldsymbol{u}}$,

[1] At test time, the original saliency map is generated by the saliency detector in SalG-GAN.

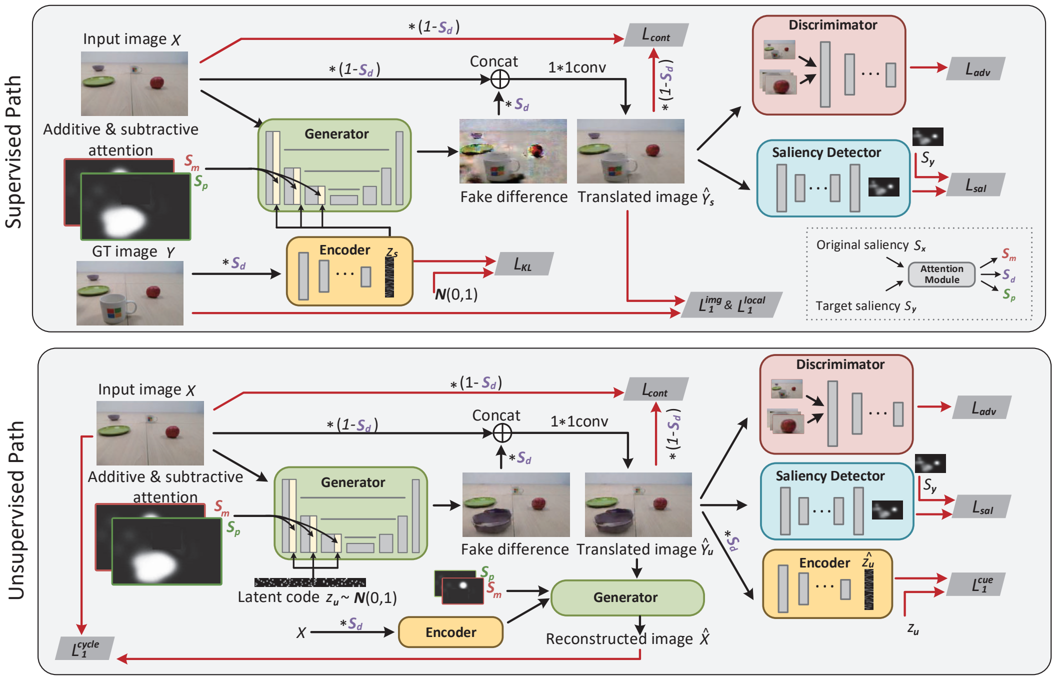
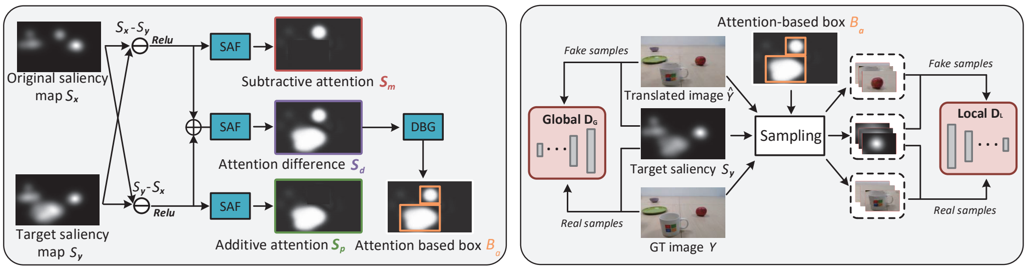


Figure 2. **Training pipeline of our SalG-GAN.** Given the original image, latent saliency cue, and corresponding attention maps, the fake image is generated by the generator to fool the discriminators. In the supervised path, the latent saliency cue is extracted from the ground-truth image, and thus the generated fake image is supposed to close to the ground-truth. In the unsupervised path, the latent saliency cue is sampled from a normal distribution, and then this latent code needs to be re-predicted from the generated fake image by the encoder.


Original saliency map S_x
S_x-S_y
Relu
SAF
Subtractive attention S_m
SAF
Attention difference S_d
DBG
S_y-S_x
Relu
SAF
Additive attention S_p
Attention based box B_a
Target saliency map S_y
Attention-based box B_a
Fake samples
Translated image $\hat{Y}$
Global D_G
Target saliency S_y
Sampling
Real samples
GT image Y
Fake samples
Local D_L
Real samples

(a) Saliency-based attention module

(b) Saliency-guided global and local discriminators

Figure 3. **(a) The details about saliency-based attention module.** Based on original and target saliency maps, the attention module generates the additive attention map, subtractive attention map, attention difference map and attention-based bounding box for further use. Note that SAF and DBG are the saliency adjustment and density-based bounding box generation functions. **(b) The details about saliency-guided global and local discriminators.** As shown, the global discriminator is used to classify real and fake images along with corresponding saliency map, while local discriminator focuses on regions/objects with high saliency difference.

we use $\hat{\boldsymbol{Y}}_{\boldsymbol{u}}$ to reconstruct the original image $\boldsymbol{X}$, with corresponding loss functions.

3.2. Detailed Structures

Saliency-based Attention Module. As illustrated in Figure 3-(a), based on original saliency map $\boldsymbol{S_x}$ and the target saliency map $\boldsymbol{S_y}$, we introduce a *saliency-based attention module* to obtain additive attention map $\boldsymbol{S_p}$, subtractive attention map $\boldsymbol{S_m}$ and attention difference map $\boldsymbol{S_d}$, which indicate the regions of saliency increase, saliency decrease and absolute saliency change, respectively. Before obtaining the attention maps, a saliency adjustment function $\mathrm{SAF}(\cdot)$ is developed to adjust the sparsity of a saliency map $\boldsymbol{S}$:

$$\mathrm{SAF}(\boldsymbol{S}) = \mathrm{Norm}(\frac{1}{1+\exp(-\theta_\alpha \cdot (\boldsymbol{S} - \theta_\beta))}). \quad (1)$$

In (1), $\mathrm{Norm}(\cdot)$ is 0 to 1 normalization, while θ_α and θ_β are the scaling and shifting hyper-parameters. Given the attention difference map $\boldsymbol{S_d}$, a density-based bounding box generation function is also developed to extract the bounding box B_a of each salient region for further useage.

Saliency-guided Generator. In the task of saliency-guided image translation, most parts of the input image are intended to stay consistent. Therefore, as illustrated in Figure 2, a residual learning scheme is introduced in our *saliency-guided generator* G, which can also help improve the training efficiency. Specifically, based on original image $\boldsymbol{X}$, additive attention map $\boldsymbol{S_p}$, subtractive attention map $\boldsymbol{S_m}$, latent saliency cue $\boldsymbol{z}$ and attention difference map $\boldsymbol{S_d}$, the generated fake image $\hat{\boldsymbol{Y}}$ can be represented as follows

$$\begin{aligned} \hat{\boldsymbol{Y}} &= \mathrm{G}(\boldsymbol{X}, \boldsymbol{S_d}, \boldsymbol{S_p}, \boldsymbol{S_m}, \boldsymbol{z}) \\ &= \mathrm{C}_{1\mathrm{x}1}\left(\boldsymbol{S_d} \cdot \mathrm{Unet}(\boldsymbol{X}, \boldsymbol{S_p}, \boldsymbol{S_m}, \boldsymbol{z}) \oplus (1 - \boldsymbol{S_d}) \cdot \boldsymbol{X}\right), \end{aligned} \tag{2}$$

where $\mathrm{C}_{1\mathrm{x}1}$ is the 1×1 convolutional layer, and $\oplus$ is channel wise concatenation. In (2), $\mathrm{Unet}(\cdot)$ is the U-shaped structure, including 8 pairs encoder and decoder blocks with symmetric skip connections. It is worth noting that, instead input the target saliency map, the generator G is separately feed with $\boldsymbol{S_p}$ and $\boldsymbol{S_m}$. That helps the generator directly learn how to increase or decrease the saliency of certain regions during image translation. Further, the attention difference map $\boldsymbol{S_d}$ in (2) is used to encourage the generator to focus on generating the regions with high saliency change.

Saliency Cue Encoder. Inspired by the idea of VAE-GAN [23], the *saliency cue encoder* E is developed to estimate the mean (μ) and variance (σ) of the posterior distribution for each input image, then the corresponding latent saliency cue can be sampled from this posterior. For example, in the supervised path, $\boldsymbol{z_s}$ is encoded by E, as the saliency cue of ground-truth image $\boldsymbol{Y}$:

$$\begin{aligned} \boldsymbol{z_s} \sim Q(\boldsymbol{z_s}|\boldsymbol{Y}) &= \mathcal{N}(\mu_y, \sigma_y), \\ \text{where} \qquad \mu_y, \sigma_y &= \mathrm{E}(\boldsymbol{Y} \cdot \boldsymbol{S_d}). \end{aligned} \tag{3}$$

In (3), $\boldsymbol{S_d}$ is the attention difference map from our *saliency-based attention module*, which helps the encoder focus on the regions with high saliency change. The *saliency cue encoder* consists of 5 convolutional layers, followed by 2 FC layers for estimating mean and variance, respectively.

Saliency Detector. In our SalG-GAN, a light but effective *saliency detector* is developed to predict saliency map from image. Specifically, the saliency detector consists of 3 dense blocks [12] followed by 3 deconvolutional blocks. Between dense and deconvolutional blocks, an Atrous Spatial Pyramid Pooling (ASPP) [3] is added to extract multi-scale features for saliency prediction.

Saliency-guided Global and Local Discriminators. As illustrated in 3-(b), we adopt a global discriminator $\mathrm{D_G}$ and a local discriminator $\mathrm{D_L}$ to judge the realism of the translated images. As shown, $\mathrm{D_G}$ is used to discriminate the realism of the whole input image, while $\mathrm{D_L}$ works on the image patches sampled by the attention based bounding boxes B_a from our *saliency-based attention module*. In addition to the fake ($\hat{\boldsymbol{Y}}$) or real ($\boldsymbol{Y}$) image, the target saliency map $\boldsymbol{S_y}$ is also input to discriminators as the conditional information. The experimental results show that this helps to avoid mode collapse problems. Consequently, an LSGAN [25] based objective can be formulated as

$$\begin{aligned} \mathcal{L}_{\mathrm{adv}} = & \mathop{\mathbb{E}}_{\hat{\boldsymbol{Y}}, \boldsymbol{S_y} \sim p_f(\hat{\boldsymbol{Y}}, \boldsymbol{S_y})} \left(||\mathrm{D_G}(\hat{\boldsymbol{Y}}, \boldsymbol{S_y})||^2 + ||\mathrm{D_L}(\hat{\boldsymbol{Y}}, \boldsymbol{S_y})||^2 \right) \qquad (4) \\ + & \mathop{\mathbb{E}}_{\boldsymbol{Y}, \boldsymbol{S_y} \sim p_r(\boldsymbol{Y}, \boldsymbol{S_y})} \left(||1 - \mathrm{D_G}(\boldsymbol{Y}, \boldsymbol{S_y})||^2 + ||1 - \mathrm{D_L}(\boldsymbol{Y}, \boldsymbol{S_y})||^2 \right), \end{aligned}$$

where p_f (p_r) represents the joint distribution of all fake (real) images and corresponding saliency maps. In our SalG-GAN, p_f includes the translated images ($\hat{\boldsymbol{Y}}_{\boldsymbol{s}}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{u}}$) from the supervised and unsupervised paths, and p_r includes the original images $\boldsymbol{X}$ and the ground-truth images $\boldsymbol{Y}$. The structure of $\mathrm{D_G}$ is based on PatchGAN [13] with 3 scales, while $\mathrm{D_L}$ is a single-scale discriminator.

3.3. Loss Functions

The proposed SalG-GAN is trained in an end-to-end and adversarial manners. Besides the adversarial loss $\mathcal{L}_{\mathrm{adv}}$ in (4), the following 7 losses are also introduced.

(i) *Content Loss* $\mathcal{L}_{\mathrm{cont}}$ is introduced to preserve content consistency of regions without attention change, between the translated image $\hat{\boldsymbol{Y}}$ and original image $\boldsymbol{X}$:

$$\mathcal{L}_{\mathrm{cont}} = \mathrm{D_{VGG}}\left((1 - \boldsymbol{S_d}) \cdot \hat{\boldsymbol{Y}}, (1 - \boldsymbol{S_d}) \cdot \boldsymbol{X}\right), \tag{5}$$

where $\boldsymbol{S_d}$ is attention difference map, and $\mathrm{D_{VGG}}(\cdot)$ is the VGG-based feature-wise distance in [17].

(ii) *Image Reconstruction Loss* $\mathcal{L}_1^{\mathrm{img}} = ||\hat{\boldsymbol{Y}}_{\boldsymbol{s}} - \boldsymbol{Y}||_1$ penalizes the L1 difference between the ground-truth image $\boldsymbol{Y}$ and the translated image $\hat{\boldsymbol{Y}}_{\boldsymbol{s}}$ in the supervised path.

(iii) *Local Reconstruction Loss* $\mathcal{L}_1^{\mathrm{local}}$ is the L1 distance inside the attention based bounding boxes B_a, between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{s}}$ in the supervised path.

(iv) *Latent Saliency Cue KL Loss* $\mathcal{L}_{\mathrm{KL}}$ is applied to penalizes the posterior distribution $Q(\boldsymbol{z_s}|\boldsymbol{Y})$ in the supervised path to be close to the standard normal distribution $\mathcal{N}(0, 1)$, by measuring KL divergence $\mathrm{D_{KL}}(\cdot)$:

$$\begin{aligned} & \mathcal{L}_{\mathrm{KL}} = \mathbb{E}[\mathrm{D_{KL}}\left(Q(\boldsymbol{z_s}|\boldsymbol{Y})\,||\,\mathcal{N}(0, 1)\right)], \\ \text{where} \quad & \mathrm{D_{KL}}(p||q) = -\int \mathrm{p(z)} \log \frac{\mathrm{p(z)}}{\mathrm{q(z)}} \mathrm{dz}. \end{aligned} \tag{6}$$

(v) *Saliency Consistency Loss* $\mathcal{L}_{\mathrm{sal}}$ penalizes the distribution difference between the translated image $\hat{\boldsymbol{Y}}$ and the target saliency map $\boldsymbol{S_y}$ in the terms of KL divergence $\mathrm{D_{KL}}(\cdot)$:

$$\mathcal{L}_{\mathrm{sal}} = \mathrm{D_{KL}}\left(\mathrm{SalD}(\hat{\boldsymbol{Y}})\,||\,\boldsymbol{S_y}\right), \tag{7}$$

where $\mathrm{SalD}(\cdot)$ is the saliency detector in our SalG-GAN.

(vi) *Cycle Loss* $\mathcal{L}_1^{\text{cycle}}$ is applied in the unsupervised path to ensure that the translated image $\hat{\boldsymbol{Y}_{\boldsymbol{u}}}$ can further reconstruct the original image $\boldsymbol{X}$:

$$\mathcal{L}_1^{\text{cycle}} = ||\boldsymbol{X} - \mathrm{G}(\hat{\boldsymbol{Y}_{\boldsymbol{u}}}, \boldsymbol{S_d}, \boldsymbol{S_m}, \boldsymbol{S_p}, \boldsymbol{z}'_{\boldsymbol{u}})||_1, \quad \text{where} \quad \hat{\boldsymbol{Y}_{\boldsymbol{u}}} = \mathrm{G}(\boldsymbol{X}, \boldsymbol{S_d}, \boldsymbol{S_p}, \boldsymbol{S_m}, \boldsymbol{z_u}). \tag{8}$$

In (8), $\mathrm{G}(\cdot)$ is the proposed generator formulated in (2). $\boldsymbol{S_d}$, $\boldsymbol{S_p}$ and $\boldsymbol{S_m}$ are the additive attention, subtractive attention and attention difference maps. Besides, $\boldsymbol{z}'_{\boldsymbol{u}}$ and $\boldsymbol{z_u}$ are latent saliency cues, sampled from $Q(\boldsymbol{z}'_{\boldsymbol{u}}|\boldsymbol{X})$ and $\mathcal{N}(0,1)$.

(vii) *Latent Saliency Cue Regression Loss* $\mathcal{L}_1^{\text{cue}} = ||\hat{\boldsymbol{z_u}} - \boldsymbol{z_u}||_1$ penalizes the L1 difference between the randomly sampled $\boldsymbol{z_u}$ and the re-estimated $\hat{\boldsymbol{z_u}}$ from the translated image $\hat{\boldsymbol{Y}_{\boldsymbol{u}}}$ in the unsupervised path.

Overall objective. Combining all above losses, the overall objective function of our SalG-GAN is formulated as

$$\min_{\mathrm{G,E,SalD}} \max_{\mathrm{D_G,D_L}} \mathcal{L}_{\text{adv}} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}} + \lambda_{\text{img}}\mathcal{L}_1^{\text{img}} + \lambda_{\text{local}}\mathcal{L}_1^{\text{local}} \\ + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{sal}}\mathcal{L}_{\text{sal}} + \lambda_{\text{cycle}}\mathcal{L}_1^{\text{cycle}} + \lambda_{\text{cue}}\mathcal{L}_1^{\text{cue}}, \tag{9}$$

where λ_{cont}, λ_{img}, λ_{local}, λ_{KL}, λ_{sal}, λ_{cycle} and λ_{cue} are the hyper-parameters to balance the effect of each single loss.

4. Datasets establishment

Since there is no existing dataset for saliency-guided image translation, we build a synthetic and a real-world dataset for training our SalG-GAN, called SGIT-S and SGIT-R, respectively. In practise, we manually edit the original image $\boldsymbol{X}$, to obtain the ground-truth translated image $\boldsymbol{Y}$. Then, a mouse-contingent based experiment is conducted to record the visual attention over both $\boldsymbol{X}$ and $\boldsymbol{Y}$, for generating corresponding saliency maps of $\boldsymbol{S_x}$ and $\boldsymbol{S_y}$. Additionally, considering the practical applications, a small dataset with more complex background, SGIT-C is collected as the test set (without ground-truth). Some examples of above datasets can be found in Figure 4, and the statistics of datasets are introduced in the supplementary material.

Mouse-contingent experiment. Inspired by [16, 21], we conduct mouse-contingent experiments to collect the clicks over images to represent the human attention. In this way, the saliency map can be obtained by several mouse clicks. Specifically, in our experiments, each image is first blurred by a Gaussian filter. Then, the subject is asked to click anywhere on image to reveal a small region at the original resolution. The location of each click is recorded as the proxy of ”fixation”. Finally, similar to [15], the saliency map is generated by convolving the fixations with Gaussian mask. Note that the mouse-contingent is much easier than the eye-tracking experiment, and it is convenient in practical use.

SGIT-S. Our synthetic dataset for saliency-guided image translation (SGIT-S) is built on the top of open-source project of CLEVR [19], where users can synthesize images with objects depending on the pre-set attributes of location, shape, color, material and size. First, we generate around 60,000 synthetic images with random attributes, as original images $\boldsymbol{X}$. Then, for each $\boldsymbol{X}$, we randomly conduct one of the following actions to generate the edited image $\boldsymbol{Y}$. 1) Add one or two objects. 2) Randomly remove one or two existing objects. 3) Randomly move one or two existing objects. 4) Randomly change the attribute of one or two existing object. Then, the mouse-contingent experiments are conducted on both $\boldsymbol{X}$ and $\boldsymbol{Y}$ by 7 subjects. After that, we further remove the samples with small saliency changes, based on a KL threshold of 0.2. Finally, the SGIT-S consists of 50,000 training, 1,500 validation and 1,500 testing samples (106,000 images and saliency maps in total).

SGIT-R. In addition to the synthetic dataset, it is more interesting to generate saliency-guided real-world images. To this end, we further build a real-world image dataset SGIT-R, including around 30 different objects. First, we take photos with randomly-selected objects as original images $\boldsymbol{X}$. Similar to SGIT-S, for each $\boldsymbol{X}$, we also randomly add, remove, move and replace the existing objects, and then take a new photo as the edited image $\boldsymbol{Y}$. Note that the camera is mounted on a tripod, to keep the same view. The mouse-contingent experiments are conducted on all these images to collect visual attention from 7 subjects. After removing the pairs of images with small saliency change, we obtain 600 training, 40 validation and 80 testing samples of original and “changed” images, as well as their saliency maps (1,440 images and saliency maps in total).

SGIT-C. In order to evaluate the proposed method in practical scenario, we further collect a test set of around 300 images with more complex background, namely SGIT-C. Specifically, SGIT-C is collected via two ways: 1) Similar to SGIT-S, we take photos with randomly-selected objects, but in the scenes with more complex background. 2) Meanwhile, we randomly select the images from the test set of Place2 [39]. Given a collected image $\boldsymbol{X}$, we conduct the mouse-contingent experiment to get the saliency map $\boldsymbol{S_x}$ of $\boldsymbol{X}$. Then, $\boldsymbol{S_x}$ is manually modified to be the target saliency map $\boldsymbol{S_y}$ via another mouse-contingent experiment. Note that the images in SGIT-C don’t have the ground-truth translated image, and are only used on the test stage.

5. Experiments

5.1. Implementation Details

For stable training, we apply spectral normalization [30] in all generator and discriminators of our SalG-GAN. Instance normalization [32] and leaky RelU are used as the normalization and activation functions in SalG-GAN. The resolution of input and output images are set to 256×256, and the dimension of latent saliency cue is 16. As the hyper-parameters, λ_{cont}, λ_{img}, λ_{local}, λ_{KL}, λ_{sal}, λ_{cycle} and λ_{cue} are set to 5, 10, 10, 10^{-5}, 15, 5 and 10, respectively. Dur-

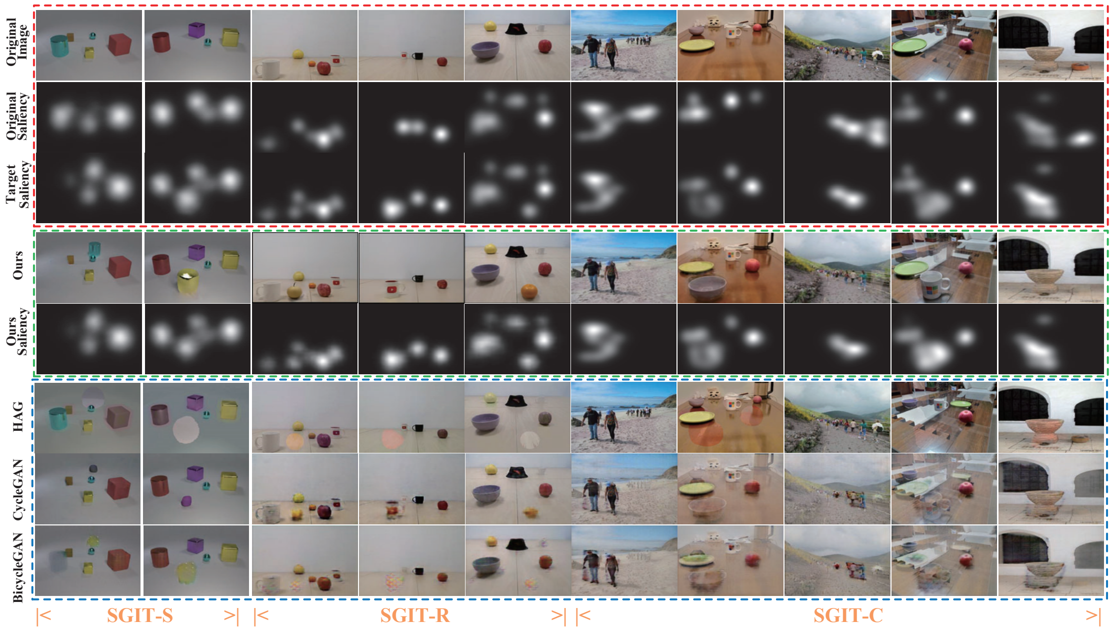


Figure 4. **Examples of our and baseline methods on SGIT-S, SGIT-R and SGIT-G.** In the figure, from the top to bottom rows are: original images, original saliency maps, target saliency maps, the translated images from our method and the baselines.

ing training, Adam optimizer [22] is applied in SalG-GAN with initial learning rate of 2×10^{-4} and batch size of 16. The training process over SGIT-S takes about 30 hours on a single GTX 1080Ti GPU for 30 epochs. For SGIT-R, we exchange the original and ground-truth images for data augmentation, and the whole training takes around 12 hours on the same device for 240 epochs. The pre-trained model of SGIT-R is directly applied on SGIT-C for evaluation.

5.2. Baseline models

Since there is no existing method can be directly used for saliency-guided image translation, we take a saliency-driven image editing method HAG [9] and state-of-the-art conditional image generation methods, CycleGAN [40] and BicycleGAN [41] as the baseline models. For HAG, the target saliency map is input for editing the original image. For CycleGAN and BicycleGAN, the saliency maps are concatenated with the input, as the conditional information to guide translation of the input images. Meanwhile, same saliency detector and saliency-related loss in our SalG-GAN, are also added in CycleGAN and BicycleGAN. The models of CycleGAN and BicycleGAN are retrained over SGIT-S and SGIT-R for fair comparison.

5.3. Metrics

The translated images should be realistic, diverse and satisfying the target saliency distribution. Thus, we use 4 evaluation metrics, Frechet Inception Distance (FID), local Diversity Score (local DS) and KL divergence between saliency map (SalD-KLD and F-KLD). The lower FID, SalD-KLD and F-KLD mean better performance, while higher local DS indicates that results are more diverse.

FID. FID [10] is a robust metric to evaluate the realism of generated images, based on the 2nd order similarity of the final layer of the inception model.

Local DS. In [37] authors propose DS by measuring perceptual similarity between two images in deep feature space, to evaluate the diversity of the generated images from the same input. However, for the task of saliency-guided image translation, the non-salient regions of the translated images are supposed to be consistent with input. Thus, we conduct DS on the image patches with attention difference, based on the saliency based bounding boxes B_a.

SalD-KLD and F-KLD. In order to evaluate the saliency of the translated image, we calculate KLD to measure the distribution difference between the target saliency and the saliency map of the translated image. Specifically, we apply the pre-trained saliency detector to generate the saliency map of each translated image, the KLD results based on which are denoted as SalD-KLD. Simliarly, we further collect eye-movement data on translated images by conducting mouse-contingent experiments with 7 subjects. Then, the KLD results based on fixation maps are denoted as F-KLD.

Table 1. Performance of ours and baseline methods on SGIT-S, SGIT-R and SGIT-C, in terms of FID, local DS, SalD-KLD and F-KLD.

	SGIT-S				SGIT-R				SGIT-C			
	FID	Local DS	SalD-KLD	F-KLD	FID	Local DS	SalD-KLD	F-KLD	FID	Local DS	SalD-KLD	S-KLD
HAG	60.31	–	0.26	0.47	69.63	–	0.65	0.35	**45.21**	–	0.73	1.84
CycleGAN	34.81	–	0.03	0.28	106.45	–	0.03	0.18	114.58	–	0.43	1.17
BicycleGAN	113.92	–	0.06	0.47	122.03	–	0.08	0.30	118.67	–	0.38	1.47
SalG-GAN(Ours)	**30.51**	**0.31**	**0.02**	**0.27**	**48.59**	**0.11**	**0.02**	**0.14**	53.22	**0.08**	**0.12**	**0.70**

Table 2. **User study.** Preference (in %) between results obtained using our and baseline methods.

	SGIT-S			SGIT-R			SGIT-C		
	Realism	Saliency	Content	Realism	Saliency	Content	Realism	Saliency	Content
HAG	8.2%	4.7%	18.3%	25.2%	9.7%	21.6%	47.7%	2.0%	43.7%
CycleGAN	21.0%	21.8%	10.2%	12.8%	16.7%	16.5%	2.1%	13.1%	2.1%
BicycleGAN	15.7%	18.8%	10.1%	8.8%	10.3%	14.1%	2.0%	13.0%	2.1%
Ours	**55.1%**	**54.7%**	**61.4%**	**53.2%**	**63.3%**	**47.8%**	**48.2%**	**71.9%**	**52.1%**

5.4. Qualitative results

Figure 4 shows the results obtained by our and baseline methods on SGIT-C and the test sets of SGIT-S and SGIT-R. As seen from this figure, we show the original images randomly selected from SGIT-S, SGIT-R and SGIT-C, as well as their original and target saliency maps. Then, the translated images of our and other 3 baseline models (*i.e.*, HAG, CycleGAN and BicycleGAN) are also presented in Figure 4. It is clear that our method can generate higher quality images than all baseline models. Moreover, the saliency maps of our translated images are close to the target saliency maps, which verifies our model is able to achieve image translation that perfectly satisfy the target saliency map. On the contrary, in most cases, the baseline models fail to generate realistic images, or fail to generate translations that satisfy the target saliency maps. For more image translation results of SGIT-S, SGIT-R and SGIT-C, please see our supplemental material. The supplemental material also demonstrates our model's ability to generate diverse results, by providing diverse translated images with the same target saliency map and original image as the inputs.

5.5. Quantitative results

In addition to the qualitative results, Table 1 summarizes comparison results of FID, local DS, salD-KLD and F-KLD over our SalG-GAN and baseline models, *i.e.*, HAG, CycleGAN and BicycleGAN that mentioned in Section 5.2. As shown in Table 1, our proposed method significantly outperforms baselines over both datasets of SGIT-S, SGIT-R and SGIT-G, in the terms of FID. This indicates our method is able to generate more realistic than the baselines. Meanwhile, the generated images from our SalG-GAN can perfectly satisfy the target saliency map, with averaged SalD-KLD of only 0.02 and F-KLD of 0.27. Besides, our SalG-GAN is the only method succeeds to calculate DS. HAG is deterministic methods, so they generate single outputs. Interestingly, even though CycleGAN and BicycleGAN are conducted with the latent code, they fail to generate diverse results. This verifies the effectiveness of our disentangle representation in SalG-GAN.

Besides, we also evaluate results of images generated by our and other 3 baseline methods, *i.e.*, HAG [9], CycleGAN [40] and BicycleGAN [41]. The setting of our user study is similar to those in [41] and [24]. Specifically, given the input image and the target saliency, the translated images generated from our and other 3 methods are presented to 20 subjects. Then, they are asked: (1) "Which generated image is most realistic?", (2) "Which generated image satisfies target saliency best?" and (3) "Which generated image has the highest content consistency with the input?" For each question, the subject needs to pick up the "best" image. Table 2 lists the preference percentages over datasets SGIT-S, SGIT-R and SGIT-G, in terms of realism (Q1), saliency accuracy (Q2) and content consistency (Q3). It can be seen that our method performs better than others in the terms of all 3 subjective metrics. It is also worth noting that realism and consistency for HAG are high, but the saliency is much lower. That is because, in many cases, HAG just outputs the original image without any manipulation.

Additionally, the supplemental material provides the ablation results by removing each single loss, developed components and attention mechanism and in our SalG-GAN, which verify the effectiveness of the designs in SalG-GAN.

6. Conclusion

In this paper, we proposed a novel task of *saliency-guided image translation*, with the goal of image-to-image translation conditioned on the user specified saliency map. Also, we developed a novel SalG-GAN method for this task. Furthermore, we built a synthetic and a real-world datasets with labeled visual attention for training and evaluating our SalG-GAN. The experimental results over both datasets verified the effectiveness of our method.

Acknowledgments

This work was sponsored by NSFC under Grants 62050175, 62001016, 61876013, and 61922009, and Beijing Natural Science Foundation under Grant JQ20020.

References

[1] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018. 1

[2] M Bernhard, L Zhang, and Manuel Wimmer. Manipulating attention in computer games. In *2011 IEEE 10th IVMSP Workshop: Perception and Visual Signal Analysis*. IEEE, 2011. 3

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 5

[4] Yen-Chung Chen, Keng-Jui Chang, Yu-Chiang Frank Wang, Yi-Hsuan Tsai, and Wei-Chen Chiu. Guide your eyes: Learning image manipulation under saliency guidance. In *BMVC*, 2019. 2, 3

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018. 1, 3

[6] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *ICCV*, 2019. 1, 3

[7] Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. Finding distractors in images. In *CVPR*, 2015. 1, 2

[8] Leon A Gatys, Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Guiding human gaze with convolutional neural networks. *arXiv preprint arXiv:1712.06492*, 2017. 2, 3

[9] 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*. ACM, 2011. 3, 7, 8

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 7

[11] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *CVPR*, pages 7986–7994, 2018. 1, 3

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 5

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 3, 5

[14] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 1998. 1

[15] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. Deepvs: A deep learning based video saliency prediction approach. In *ECCV*. Springer, 2018. 6

[16] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015. 6

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*. Springer, 2016. 5

[18] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, pages 1219–1228, 2018. 1, 3

[19] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017. 6

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[21] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2017. 6

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[23] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 5

[24] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pages 35–51, 2018. 3, 8

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017. 5

[26] Victor A Mateescu and Ivan V Bajic. Visual attention retargeting. *IEEE Transactions on MultiMedia*, 2015. 1

[27] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Saliency driven image manipulation. *Machine Vision and Applications*, 2019. 1, 2, 3

[28] Youssef Alami Mejjati, Celso F Gomez, Kwang In Kim, Eli Shechtman, and Zoya Bylinskii. Look here! a parametric learning based approach to redirect visual attention. *ECCV*, 2020. 2, 3

[29] Erick Mendez, Steven Feiner, and Dieter Schmalstieg. Focus and context in mixed reality by modulating first order salient features. In *International Symposium on Smart Graphics*. Springer, 2010. 3

[30] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 6

[31] Tam V Nguyen, Bingbing Ni, Hairong Liu, Wei Xia, Jiebo Luo, Mohan Kankanhalli, and Shuicheng Yan. Image re-attentionizing. *IEEE Transactions on Multimedia*, 2013. 1, 2, 3

[32] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 6

[33] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 1

[34] Ziwei Liu Weidong Yin and Chen Change Loy. Instance level facial attributes transfer with geometry-aware flow. In *AAAI*, February 2019. 1, 3

[35] Lai-Kuan Wong and Kok-Lim Low. Saliency retargeting: An approach to enhance image aesthetics. In *WACV*. IEEE, 2011. 2, 3

[36] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, pages 4471–4480, 2019. 1

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[38] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *CVPR*, pages 8584–8593, 2019. 1, 3

[39] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 6

[40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 7, 8

[41] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, pages 465–476, 2017. 3, 7, 8