Supplemental Material of Paper Saliency-Guided Image Translation

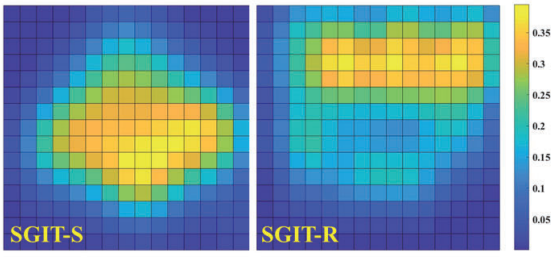
Lai Jiang[1,2], Mai Xu[1]*, Xiaofei Wang[1], Leonid Sigal[2]
[1] School of Electronic and Information Engineering, Beihang University, Beijing, China
[2] Department of Computer Science, University of British Columbia, Vancouver, BC Canada
{jianglai.china, MaiXu, xfwang}@buaa.edu.cn, lsigal@cs.ubc.ca

1. Statistics of the datasets

Here, we provide some statistics of SGIT-S and SGIT-R. Specifically, we count the object numbers of each image in SGIT-S and SGIT-R. As shown in Figure 1-(a), most images in SGIT-S and SGIT-R contain 3 to 5 objects, while in some images the object number can be more than 8. Besides, the averaged saliency distribution of SGIT-S and SGIT-R are shown in Figure 1-(b). We can see from this figure that, SGIT-S has an obvious center-bias, which is quite common in other saliency datasets. It is interesting to find that, in SGIT-R, the salient regions are more likely to be on the top part of the images. This indicates a long-shot photographer bias in SGIT-R.

	2	3	4
SGIT-S	6,601	12,852	18,085
SGIT-R	40	173	267
	5	>= 6	Total
SGIT-S	13,108	2,354	53,000
SGIT-R	200	40	720

(a) Statistic of object number



(b) Averaged saliency distribution

Figure 1. Statistics of proposed SGIT-S and SGIT-R datasets.

2. Additional qualitative results

Here we represent more qualitative results on SGIT-S, SGIT-R and SGIT-C datasets. Specifically, Figures 2, 3 and 4 show the translated images of our and three baseline models over SGIT-S, SGIT-R and SGIT-C, respectively. As can be seen from the figure, our model learns to translate the images according to the target saliency maps. Meanwhile, our model achieve generating higher quality images than all baseline models.

3. Ablation study

Here, we conduct ablation experiments by removing each single loss, developed components and attention mechanism and in our SalG-GAN. All ablation models in Table 1 are trained in the same setting, and evaluated in terms of FID, local DS and SalD-KLD. Recall that $\mathcal{L}_{\text{cont}}$, $\mathcal{L}_1^{\text{img}}$, $\mathcal{L}_1^{\text{local}}$, $\mathcal{L}_{\text{KL}}$, $\mathcal{L}_{\text{sal}}$, $\mathcal{L}_1^{\text{cycle}}$ and $\mathcal{L}_1^{\text{cue}}$ denote the content loss, image reconstruction loss, local reconstruction loss, latent saliency cue KL loss, saliency consistency loss, cycle loss and latent saliency cue regression loss. The definition of the other ablation models are as follows.

w/o S-path indicates the model without the supervised path.

w/o U-path is the model without the unsupervised path.

*Corresponding author.

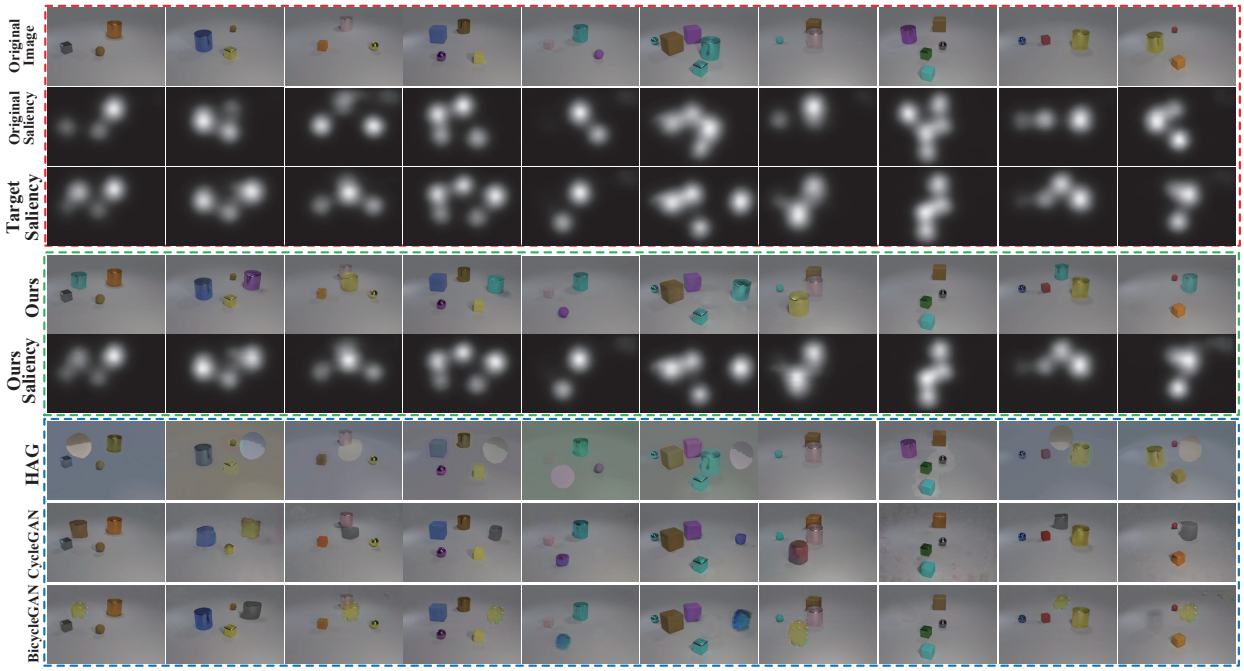


Figure 2. Saliency-guided translation examples from SGIT-S obtained using our SalG-GAN and three baseline models.

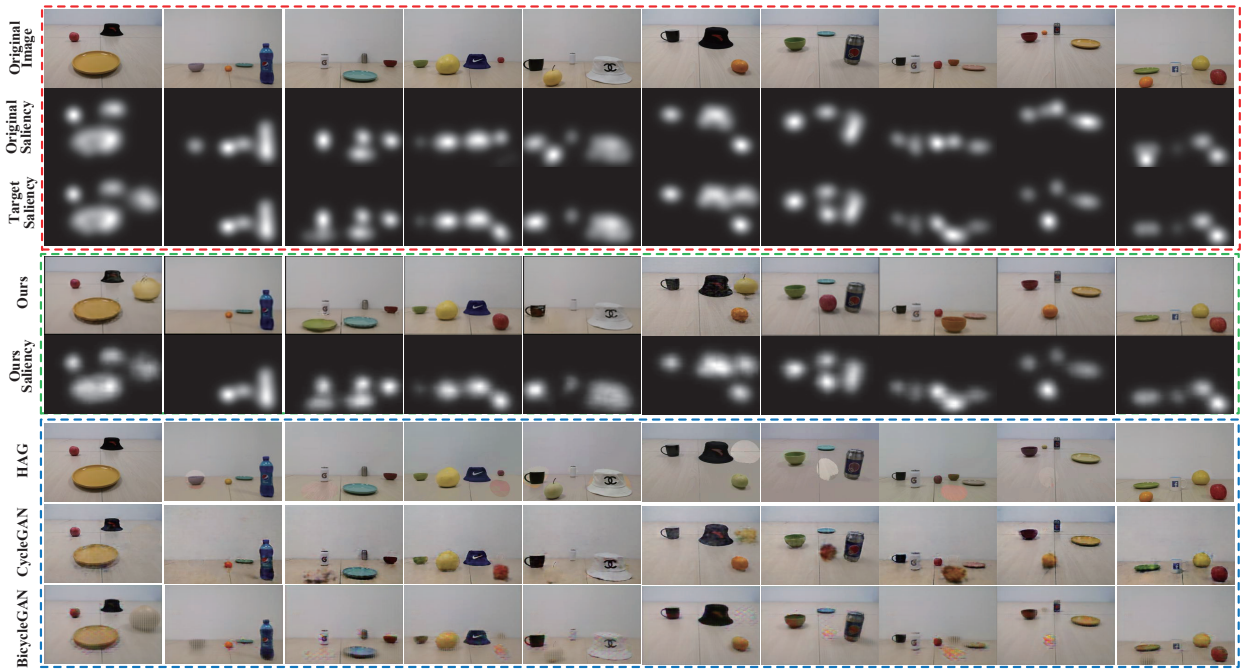


Figure 3. Saliency-guided translation examples from SGIT-R obtained using our SalG-GAN and three baseline models.

w/o D_L indicates the model without the local discriminator.
w/o SalD indicates the model without the saliency detector.

Figure 4. Saliency-guided translation examples from SGIT-C obtained using our SalG-GAN and three baseline models.

E w/o sal indicates a variant where encoder E, in SalG-GAN, does not use the attention difference map $\boldsymbol{S_d}$.
D w/o sal indicates a variant where the local and golbal discriminators ($\mathrm{D_G}$ and $\mathrm{D_G}$) in SalG-GAN do not use the target saliency map $\boldsymbol{S_y}$ as a condition.
G w/o sal is a variant where generator G in SalG-GAN use target saliency map as the condition rather than additive attention map $\boldsymbol{S_p}$ and subttractive attention map $\boldsymbol{S_m}$.

Table 1. The ablation of our model's components on SGIT-S.

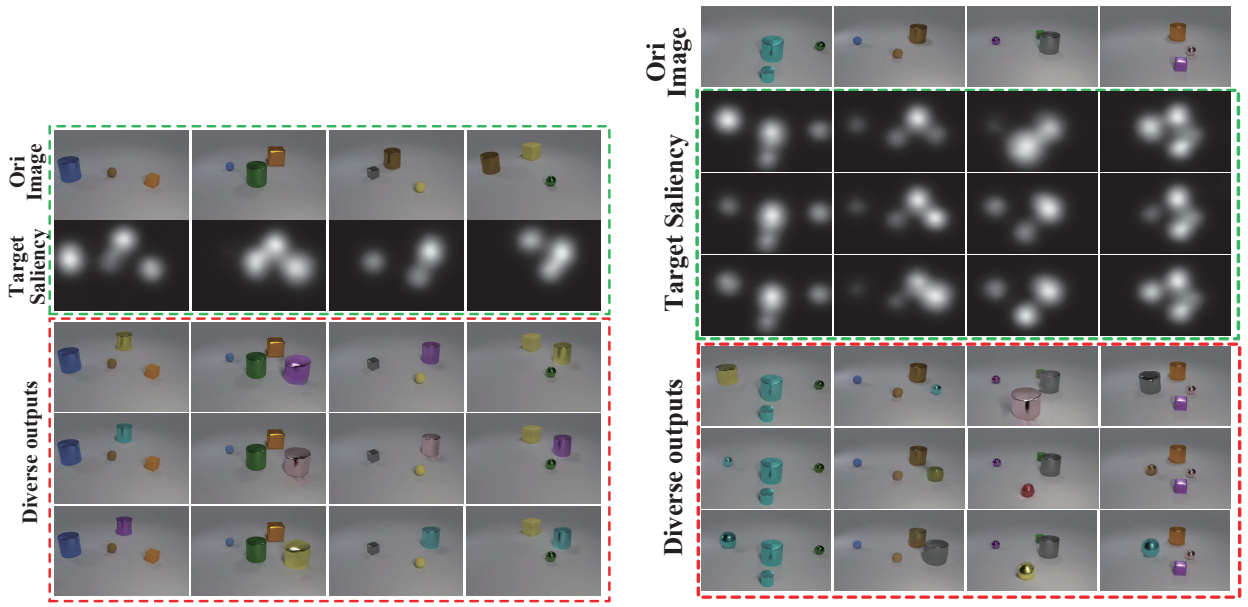
		FID	Local DS	SalD-KLD
Loss	w/o $\mathcal{L}_{\mathrm{cont}}$	45.42	0.30 ± 0.12	0.03 ± 0.01
	w/o $\mathcal{L}_1^{\mathrm{img}}$	60.74	0.36 ± 0.13	0.15 ± 0.09
	w/o $\mathcal{L}_1^{\mathrm{local}}$	47.25	0.24 ± 0.11	0.05 ± 0.01
	w/o $\mathcal{L}_{\mathrm{KL}}$	34.26	0.01 ± 0.01	0.02 ± 0.01
	w/o $\mathcal{L}_{\mathrm{sal}}$	57.54	0.24 ± 0.09	1.22 ± 0.52
	w/o $\mathcal{L}_1^{\mathrm{cycle}}$	40.45	0.28 ± 0.11	0.02 ± 0.01
	w/o $\mathcal{L}_1^{\mathrm{cue}}$	35.45	0.31 ± 0.11	0.02 ± 0.02
Components	w/o S-path	58.42	0.01 ± 0.01	0.03 ± 0.01
	w/o U-path	54.26	0.01 ± 0.01	0.02 ± 0.01
	w/o D_L	72.05	$\mathbf{0.50} \pm 0.19$	0.03 ± 0.01
	w/o SalD	57.54	0.24 ± 0.22	1.22 ± 0.58
Attention	E w/o sal	41.20	0.11 ± 0.10	0.02 ± 0.01
	D w/o sal	43.94	0.01 ± 0.01	0.02 ± 0.01
	G w/o sal	54.52	0.25 ± 0.14	0.02 ± 0.01
SalG-GAN (Full model)		**30.51**	0.31 ± 0.17	$\mathbf{0.02} \pm 0.01$

As shown in Table 1, removing either developed component or attention mechanism will decrease the realism of the translated images, especially when the local discriminator $\mathrm{D_L}$ is removed. However in the model of **w/o** $\boldsymbol{D_L}$, the DS is quite high. This is mainly because this model fails to generate realistic images and instead generates different random noise

patterns, leading to a high DS. Our model fail to generate diverse results if only single supervised/unspervised path is used (see **w/o S-path** and **w/o U-path**). Similar result can be also observed in **D w/o sal**, where we find that if the discriminators do not use saliency map as the conditions, the model can have mode collapse. As such, the variant **D w/o sal** can not generate diverse results. Finally, we find that all models can receive good results in the term of saliency KLD, except when the saliency detector is removed (**w/o SalD**). This shows the importance of our saliency detector for satisfying the target saliency condition.

4. Diversity of translated images

In addition to the image quality, we demonstrate our model's ability to generate diverse results in Figure 5-(a). As shown in Figure 5-(a), given different latent saliency cues, our model can generate diverse results, with the same target saliency map and original image as the inputs. However, we find in Figure 5-(a) that the shape variations are not captured well by the latent saliency cues. This is likely due to implicit correlation between the object shape and target saliency map. As such, we evaluate the shape variations by only slightly modifying the shape of a certain salient region in target saliency map. In this way, as shown in Figure 5-(b), our model can achieve shape variations.



(a) Diverse outputs with same target saliency

(b) Shape variations

Figure 5. (a) Diverse images generated from our method, with the same target saliency. (b) Shape variations from our method, by slightly modifying the shape of a certain salient region in target saliency. For each input, we present 3 translated images.

However, our model fails to generate diverse results in SGIT-R and SGIT-C as realistic as those in SGIT-S. That is probably because, the training data in SGIT-R and SGIT-C are not enough for learning the diversity. We take it as an interesting future work.

Table 2. Performance of our methods on the images with or without location bias, in the terms of FID, local DS and saliency KLD.

SGIT-S				SGIT-R			
	FID	Local DS	KLD		FID	Local DS	KLD
Center-bias	30.60	0.32	0.02	Top-bias	50.43	0.10	0.02
Non-bias	29.96	0.31	0.02	Non-bias	44.21	0.12	0.02

5. Evaluation on location bias

According to the statistics in Figure 1 of this supplementary, the saliency maps in our datasets exist location bias, *i.e.*, center-bias for SGIT-S and top-bias for SGIT-R. Considering that the model trained over our datasets may have generalization problem for the images without location bias, we conduct additional experiments to evaluate the performance of our methods on non-bias images. Specifically, we separate the test images in SGIT-S (or SGIT-R), depending on whether they have center-bias (or top-bias). As shown in Table 2 of this supplementary, our method performs even slightly better in non-bias images, over both SGIT-S and SGIT-R. Similar observation can be found in the qualitative results of our paper, where our method can translate non-bias images well.