Supplemental Material of Paper 4152

Anonymous CVPR 2021 submission

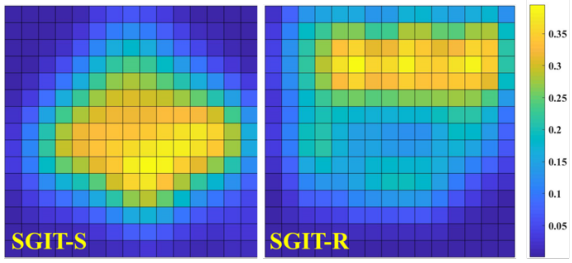
Paper ID 4152

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. 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.

Table 1. 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

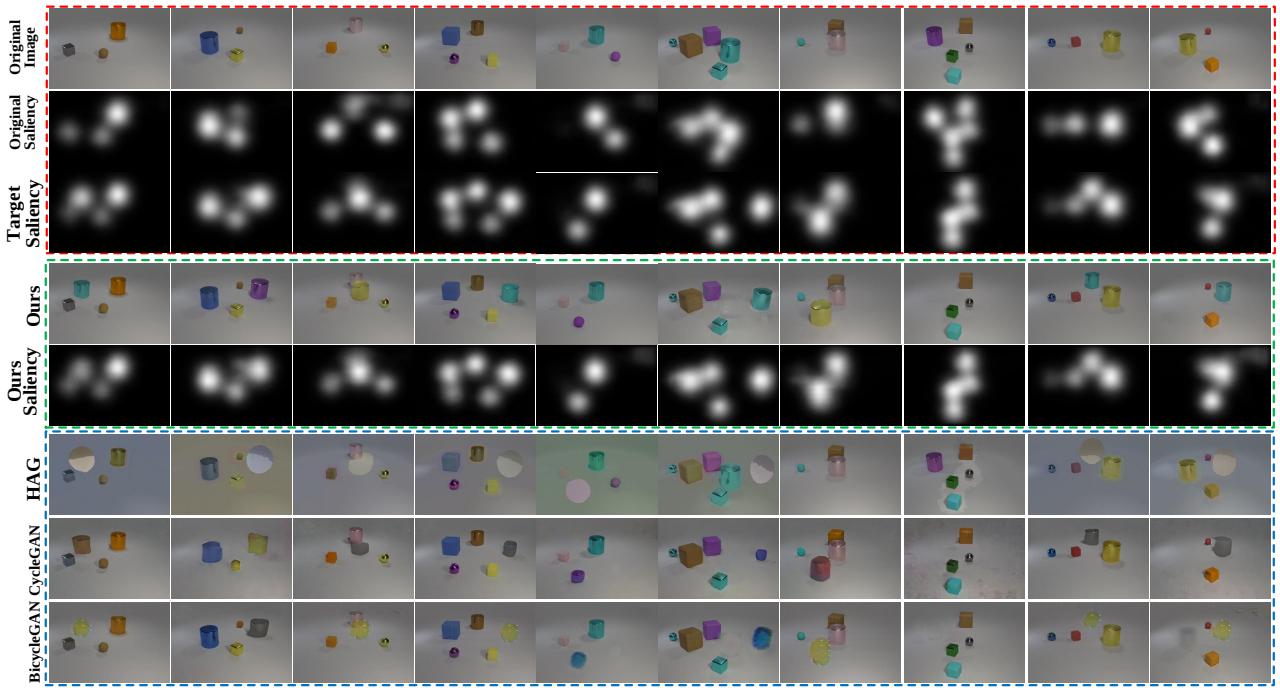


Figure 2. Saliency-guided translation examples from SGIT-S obtained using our SalG-GAN and three baseline models.

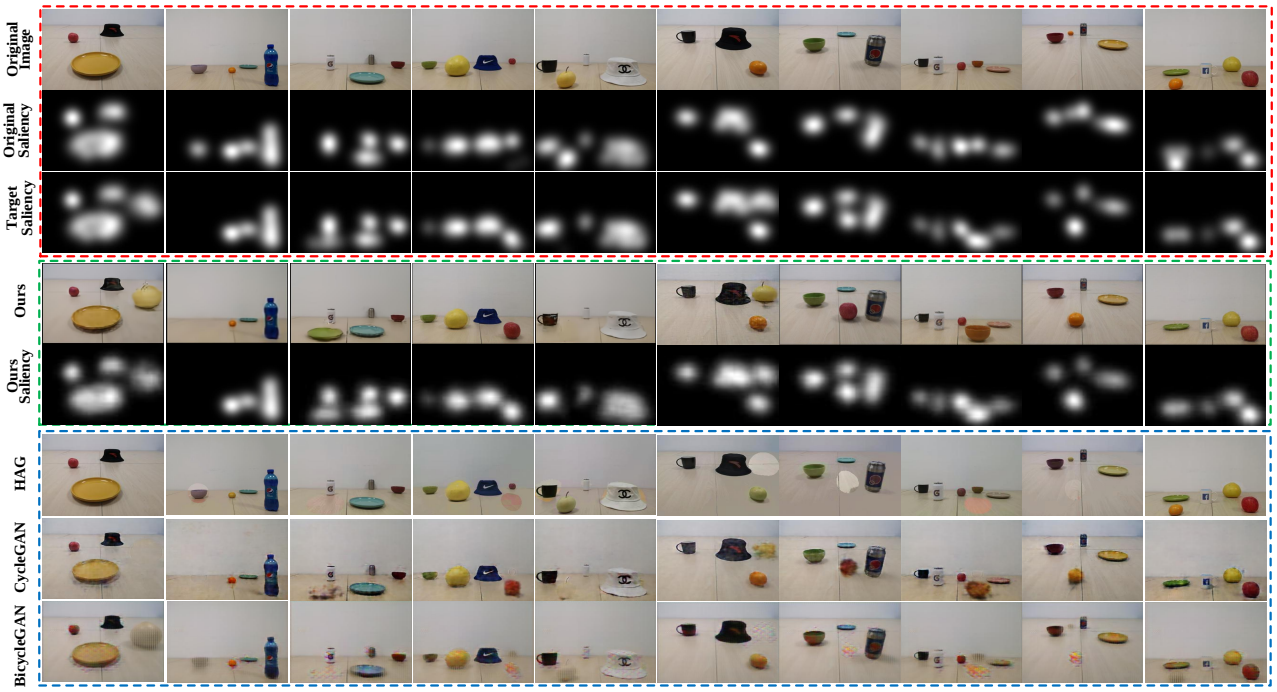
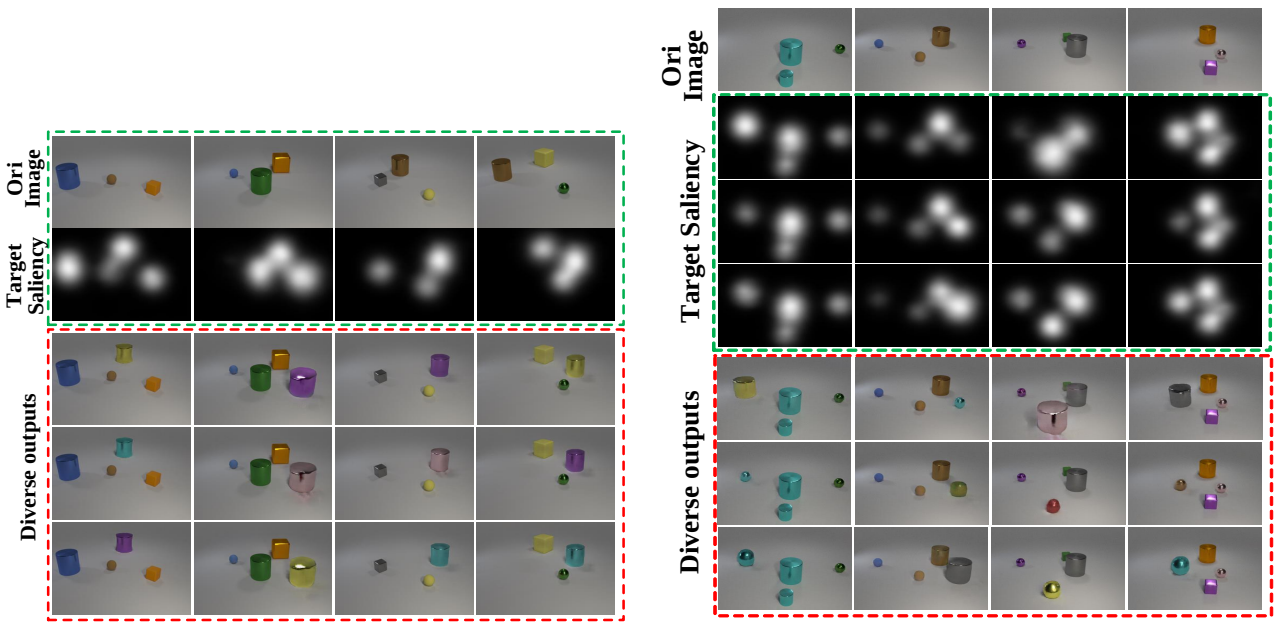


Figure 3. Saliency-guided translation examples from SGIT-R obtained using our SalG-GAN and three baseline models.

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



(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.

4. 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 1 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.