DA-GAN: Supplementary Materials

Anonymous CVPR submission

Paper ID 1877

Implementation Details

The experimental settings for each task are listed in Table 1. '♯' denotes the number of attention regions that are predefined in each task, and 'Instances' denotes the attended level of the instances. The label Y and the distance metric $d(\cdot)$ are adopted in the optimization of Deep Attention Encoder (DAE) and the instance-level translation. Note that d is jointly trained from scratch with DA-GAN, where 'ResBlock' denotes a small classifier that consists of 9 residual blocks. The learned attention regions are adaptively controlled by the selection of Y, ♯ and $d(\cdot)$. For example, the instances we learned on tasks conducted on CUB-200-2011 are parts level (birds' four parts), and for task of Colorization and domain adaption, the attended instances are objects (flower and characters).

Datasets	Label Y	♯	Instances	$d(\cdot)$
MNIST & SVHN	10	1	object	ResBlock
CUB-200-2011	200	4	parts	VGG
FaceScrub	80	4	parts	Inception
Skeleton-cartoon	20	4	parts	VGG
CMP [2]	None	4	parts	L2
Colorization [3]	Binary	1	object	ResBlock

Table 1: Implementation Details.

Experiments on CUB-200-2011

More results generated by DA-GAN are shown in Figure 2. It can be seen that, given one description, the proposed DA-GAN is capable of generating diverse images according to the specific description. Comparing with existing text-to-image synthesis works, we train the DA-GAN by unpaired text-image data. Especially, because of our proposed **instance-level translation**, we can achieve high-resolution (256×256) images directly, which is more applicable than StackGAN (it needs two stages to achieve the same resolution). We also showed more results for Pose Morphing in Figure 4. Note that, the target should be bird breeds (image collections). Here we just random select one image to represent each bird breeds for reference.

Human Face to Animation Face Translation

In this experiments, we randomly select 80 celebrities which consists of 12k images for source human face images. We also showed fine-grained translation results in Figure 1. We can see that, with the same person, DA-GAN is capable of generating diverse images, while still remain the certain one's identity attributes, e.g. big round eyes, dark brown hairs, etc.

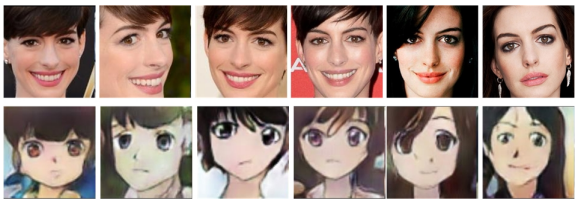

Figure 1: Fine-grained translation results.

Translation on Paired Datasets

We also conduct experiments on paired datasets. The image quality of ours results is comparable to those produced by the fully supervised approaches while our method learns the mapping without paired supervision. For the task of Skeleton to cartoon figure translation, we retrieved about 20 cartoon figures which consists of 1200 images on websites, and adopt Pose Estimator by [1] to generate skeletons for each image. The DA-GAN is trained by feeding into skeletons and generate cartoon images.

References

[1] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. 1

[2] R. Tyleček and R. Šára. *Spatial Pattern Templates for Recognition of Objects with Regular Structure*, pages 364–374. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. 1

[3] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. 1

Figure 2: Experimental Results of text-to-image synthesis.

Figure 3: Results of pose morphing. In each group, the first column is the source image, the second row is target images. The red dashed box labeled the generated images, which possess the target objects pose while remain the source objects appearance.

Figure 4: The first row is source images, and second row is target images. The translated images are placed in the third row, labeled by red dash box.

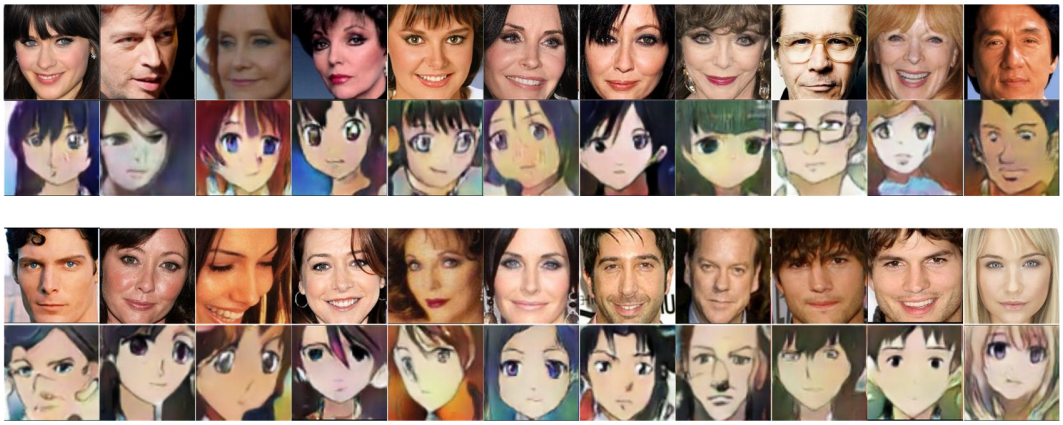

Figure 5: Results of human-to-animation faces translation. In each group, the first row is human faces, and the second row is translated animation faces.

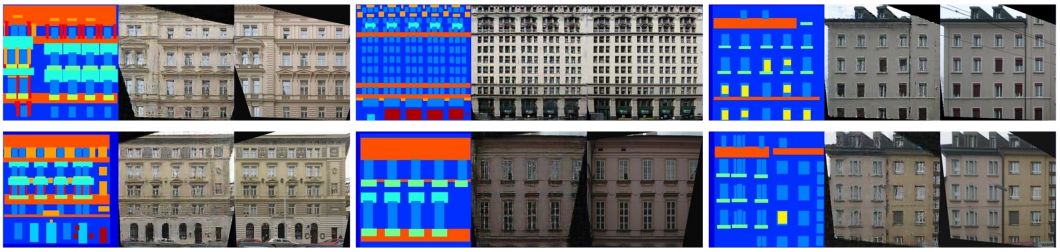

Figure 6: Results of architectural labels-to-photos translation. In each group from left to right are the input of labels, the translated architecture photos, and the ground truth.

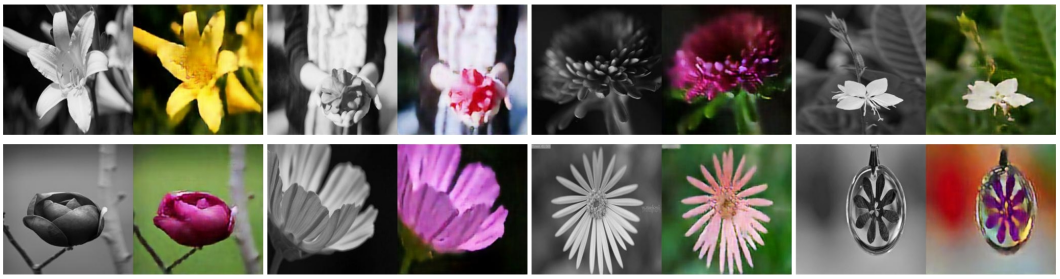

Figure 7: Results of image colorization. In each group, the input is gray images, and the results are translated color images.