Attention-GAN for Object Transfiguration in Wild Images

Xinyuan Chen[1,2,3], Chang Xu[3], Xiaokang Yang[1], Dacheng Tao[3]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{xychen91,xkyang}@sjtu.edu.cn
[2]Centre for Artificial Intelligence, SIT, FEIT, University of Technology Sydney
[3]UBTECH Sydney AI Centre, SIT, FEIT, University of Sydney
{c.xu,dacheng.tao}@sydney.edu.au

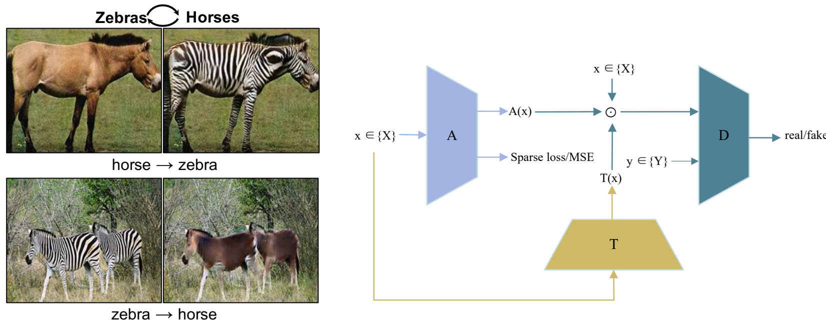
Abstract. This paper studies the object transfiguration problem in wild images. The generative network in classical GANs for object transfiguration often undertakes a dual responsibility: to detect the objects of interests and to convert the object from source domain to another domain. In contrast, we decompose the generative network into two separated networks, each of which is only dedicated to one particular sub-task. The attention network predicts spatial attention maps of images, and the transformation network focuses on translating objects. Attention maps produced by attention network are encouraged to be sparse, so that major attention can be paid on objects of interests. No matter before or after object transfiguration, attention maps should remain constant. In addition, learning attention network can receive more instructions, given the available segmentation annotations of images. Experimental results demonstrate the necessity of investigating attention in object transfiguration, and that the proposed algorithm can learn accurate attention to improve quality of generated images.

Keywords: Generative Adversarial Networks, Attention Mechanism

1 Introduction

The task of image-to-image translation aims to translate images from a source domain to another target domain, e.g., greyscale to color and image to semantic label. A lot of researches on image-to-image translation have been produced in the supervised setting, where ground truths in the target domain are available. [1] learns a parametric translation function using CNNs by minimizing the discrepancy between generated images and the corresponding target images. [2] uses conditional GANs to learn a mapping from input to output images. Similar ideas have been applied to various tasks such as generating photographs from sketch or from semantic layout [3, 4], and image super-resolution [5].

To achieve image-to-image translation in the absence of paired examples, a series of works has emerged by combining classical adversarial training [6] with different carefully designed constraints, e.g., circularity constraint [7–9], f-consistency constraint [10], and distance constraints [11]. Although there is no



(a) (b)

Fig. 1. (a): Object transfiguration of horse $\leftrightarrow$ zebra. (b): An illustration of Attention-GAN. A, T, D respectively represent the attention network, the transformation network and the discriminative network. Sparse loss denotes the sparse regularization for the predicted attention map. MSE denotes mean square error loss for supervised learning. $A(x)$ denotes the attention map predicted by the attention network. $T(x)$ denotes the transformed images. $\odot$ denotes the layered operation.

paired data, these constraints are able to establish the connections between two domains so that meaningful analogs are obtained. Circularity constraint [7–9] requires a sample from one domain to the other that can be mapped back to produce the original sample. f-consistency requires both input and output in each domain should be consistent with each other in intermediate space of a neural network. [11] learns the image translation mapping in a one-sided unsupervised way by enforcing high cross-domain correlation between matching pairwise distances computed in source and target domains.

Object transfiguration is a special task in the image-to-image translation problem. Instead of taking the image as a whole to accomplish the transformation, object transfiguration aims to transform a particular type of object in an image to another type of object without influencing the background regions. For example, in the top line of Figure 1(a), horses in the image are transformed into zebras, and zebras are transformed into horses, but the grassland and the trees are expected to be constant. Existing methods [7, 11] used to tackle object transfiguration as a general image-to-image task, without investigating unique insights of the problem. In such a one-shot generation, a generative network actually takes two distinct roles: detecting the region of interests and converting object from source domain to target domain. However, incorporating these two functionalities in a single network would confuse the aims of the generative network. In iterations, it could be unclear whether the generative network should improve its detection of the objects of interests or boost its transfiguration of the objects. The quality of generated images is often seriously influenced as a result, e.g. some background regions might be taken into transformation by mistake.

In this paper, we propose an attention-GAN algorithm for the object transfiguration problem. The generative network in classical GANs has been factorized as two separated networks: an attention network to predict where the attention should be paid, and a transformation network that actually carries out the transformation of objects. A sparse constraint is applied over the attention map, so that limited attention energy can be focused on regions of priority rather than spreaded on the whole image at random. A layered operation is adopted to finalize the generated images by combining the transformed objects and the original background regions with the help of the learned sparse attention mask. A discriminative network is employed to distinguish real images from these synthesized images, while attention network and transformation network cooperate to generate synthesized images that can fool the discriminative network. Cycle-consistent loss [7–9] was adopted to handle unpaired data. Moreover, if segmentation results of images are available, the attention network can be learned in a supervised manner and the performance of the proposed algorithm can be improved accordingly. Experimental results on three object transfiguration tasks, i.e. horse $\leftrightarrow$ zebra, tiger $\leftrightarrow$ leopard, and apple $\leftrightarrow$ orange [12], suggest the advantages of investigating attention in object transfiguration, and the quantitative and the qualitative performance improvement of the proposed algorithm over state-of-the-art methods.

2 Related Work

Generative Adversarial Networks Generative adversarial networks (GANs) [6] have achieved impressive results in image generation [13–15] by way of a two-player minimax game: a discriminator aims to distinguish the generated images from real images while a generator aims to generate realistic images to fool the discriminator. A series of multi-stage generative models has been proposed to generate more realistic images [16–18]. [17] proposes composite generative adversarial network (CGAN) that disentangles complicated factors of images by employing multiple generators to generate different parts of the image. The layered recursive GANs [18] learns to generate image background and foregrounds separately and recursively. GANs have shown a great success on a variety of conditional image generation applications, e.g., image-to-image translation [7–9, 19], text-to-image generation [20, 21]. Different from the original GANs that generate images from noise variables, conditional GANs synthesize images based on the input information (e.g., category, image and text). [22] proposes a mask-conditional contrast-GAN architecture to disentangle image background with object semantic changes by exploiting the semantic annotations in both train and test phases. However, it is hard to collect segmentation mask for a large number of images, especially in test phase.

Attention Model in Networks Motivated by human attention mechanism theories [23], attention mechanism has been successfully introduced in computer vision and natural language processing tasks, e.g. image classification [24–26], image captioning [27], visual question answering [28], image segmentation [29]. Rather than compressing an entire image or a sequence into a static representation, attention allows the model to focus on the most relevant part of images or features as needed. Mnih et al. [24] propose a recurrent network model that is capable of only processing a sequence of regions or locations of an image or video. Bahdanau et al. [30] propose an attention model that softly weights the importance of input words in a source sentence when predicting a target word for machine translation. Following this, Xu et al. [27] and Yao et al. [31] use attention models for image captioning and video captioning respectively. The model automatically learns to fix its gaze on salient objects while generates the corresponding words in the output sequence. In visual question answering, [28] uses the question to choose relevant regions of the images for computing the answer. In image generation, Gregor et al. [32] proposes a generative network combined attention mechanism with a sequential variational auto-encoding framework. The generator attends a smaller region of an input image guided by the ground truth image, and generates a few pixels for an image at a time. Differently, our method combine the attention mechanism with GANs framework and produce region of interest in absence of ground truth images in target domain.

3 Preliminaries

In the task of image-to-image translation, we have two domains X and Y with training samples $\{x_i\}_i^N \in X$ and $\{y_i\}_i^N \in Y$. The goal is to learn mapping from one domain to the other $\mathcal{G} : X \rightarrow Y$, (e.g. horse→zebra). The discriminator D_Y aims to distinguish real image y from translated images $\mathcal{G}(x)$. On the contrary, the mapping function $\mathcal{G}$ tries to generate images $\mathcal{G}(x)$ that looks similar to images in Y domain to fool the discriminator. The objective of *adversarial loss* in LSGAN [33] is expressed as:

$$\mathcal{L}_{GAN}(\mathcal{G}, D_Y, X, Y) = \mathbb{E}_{y\in Y}\left[D_Y^2(y)\right] + \mathbb{E}_{x\in X}\left[(D_Y(\mathcal{G}(x)) - 1)^2\right], \tag{1}$$

The mapping function $\mathcal{F} : Y \rightarrow X$, in the same way, tries to fool the discriminator D_X:

$$\mathcal{L}_{GAN}(\mathcal{F}, D_X, X, Y) = \mathbb{E}_{x\in X}\left[D_X^2(x)\right] + \mathbb{E}_{y\in Y}\left[(D_X(\mathcal{F}(y)) - 1)^2\right]. \tag{2}$$

The discriminators D_X and D_Y try to maximize the loss while mapping functions $\mathcal{G}$ and $\mathcal{F}$ try to minimize the loss. However, a network of sufficient capacity can map the set of input images to any random permutation of images in the target domain. To guarantee that the learned function maps an individual input x to a desired output y, the *cycle consistency loss* is proposed to measure the discrepancy occurred when the translated image is brought back to the original image space:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x\in X}\left[\|\mathcal{F}(\mathcal{G}(x)) - x\|_1\right] + \mathbb{E}_{y\in Y}\left[\|\mathcal{G}(\mathcal{F}(y)) - y\|_1\right]. \tag{3}$$

Taking advantages of adversarial loss and cycle consistency loss, the model achieves a one-to-one correspondence mapping, and discovers the cross-domain relation [8]. The full objective is:

$$\begin{aligned}\mathcal{L}(\mathcal{G}, \mathcal{F}, D_X, D_Y) = &\mathcal{L}_{GAN}(\mathcal{G}, D_Y, X, Y) \\ &+ \mathcal{L}_{GAN}(\mathcal{F}, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(\mathcal{G}, \mathcal{F}),\end{aligned} \tag{4}$$

where λ controls the relative importance of the two objectives. However, the generative mapping functions $\mathcal{G}$ and $\mathcal{F}$ actually takes a dual responsibility for object transfiguration: to detect the objects of interest and to transfigure the object, which confuse the aims of the generative network.

On the other hand, we notice that the model can be viewed as two 'autoencoders': $\mathcal{F} \circ \mathcal{G} : X \to X$ and $\mathcal{G} \circ \mathcal{F} : Y \to Y$, where the translated image $\mathcal{G}(x)$ and $\mathcal{F}(y)$ can be viewed as intermediate representations trained by adversarial loss. In object transfiguration task, the generative mappings $\mathcal{G}$ and $\mathcal{F}$ are trained to generate objects to fool the discriminator. Therefore, the image background can be coded as any representation so long as it can be decoded back to the original, which does not guarantee background consistency before and after transformation. As a result, the proposed Attention-GAN that decomposes the generative network into two separated network: an attention network to predict the object of interests and a transformation network focuses on transforming object.

4 Model

The proposed model consists of three players: an attention network, a transformation network, and a discriminative network. The attention network predicts the region of interest from the original image x. The transformation network focuses on transforming the object from one domain to the other. The resulting image is therefore a combination of the transformed object and the background of original image with a layered operator. Finally, the discriminator aims to distinguish the real image $y \in Y$ and the generated image. The overview of the proposed model is illustrated in Figure 1(b). For notation simplicity, we only show the forward process that transforms images from domain X to domain Y, and the backward process from domain Y back to the domain X can be easily obtained in the similar approach.

4.1 Formulations

The architecture of the proposed model is shown in Figure 2. Given an input image x in domain X, the attention network A_X outputs a spatial score map $A_X(x)$, whose size is the same as the original image x. The element value of score map is from 0 to 1. The attention network assigns higher scores of visual attention to the region of interest while suppressing background. In another branch, the transformation network T outputs the transformed image $T(x)$ that looks similar to those in the target domain Y. Then we adopt a layered operation to construct

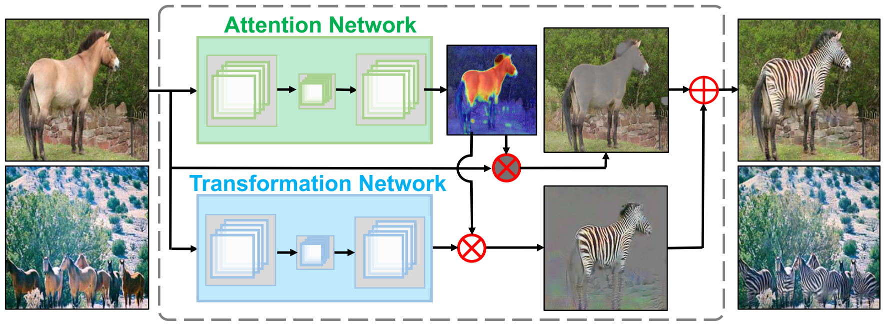


Fig. 2. The proposed Attention-GAN for object transfiguration from one class to another. The attention network predicts the attention maps. The transformation network synthesizes the target object. A layered operation is applied on the background and transformed images to output the resulting image.

the final image. Given transformed region $A_X(x)$, a transformed image $T_X(x)$ and image background from original image x are combined as:

$$\mathcal{G}(x) \equiv A_X(x) \odot T_X(x) + (1 - A_X(x)) \odot x, \tag{5}$$

where $\odot$ denotes the element-wise multiplication operator. Another mapping function $\mathcal{F}$ is introduced to bring transformed images $\mathcal{G}(x)$ back to the original space $\mathcal{F}(\mathcal{G}(x)) \approx x$. The mapping from an image y in target domain Y to the source domain follows:

$$\mathcal{F}(y) \equiv A_Y(y) \odot T_Y(y) + (1 - A_Y(y)) \odot y. \tag{6}$$

Followed by Section 3, the *adversarial loss* (Equations (1) and (2)) and the *cycle consistency loss* (Equation (3)) are introduced to learn the overall mappings $\mathcal{G}$ and $\mathcal{F}$. In classical GANs [7–9], the generative mapping $\mathcal{G}$ transforms the whole image to target domain and then the generative mapping $\mathcal{F}$ is required to bring the transformed image back to original image $\mathcal{F}(\mathcal{G}(x)) \approx x$. However, in practice, the background of the generated image appears to be unreal and significantly different from the original image background, so that the cycle consistency loss can hardly reach 0. In our method, the attention network outputs a mask that separates the image into region of interest and background. The background part will not be transformed, so that the cycle consistency loss in the background reaches 0.

4.2 Attention Losses

Similar to cycle consistency, the attention map of object x in domain X predicted by attention network A_X should be consistent with the attention map of the transformed object by attention network A_Y. For example, if a horse is transformed into a zebra, the region of the zebra should be brought back to the horse as a cycle. That is to say, the regions of interest in the original image and the transformed image should be the same: $A_X(x) \approx A_Y(\mathcal{G}(x))$. Similarly, for each image y from domain Y, attention network A_Y and A_X should satisfy consistency: $A_Y(y) \approx A_X(\mathcal{F}(y))$. To that end, we propose an attention cycle-consistent loss:

$$\mathcal{L}_{A_{cyc}}(A_X, A_Y) = \mathbb{E}_{x \in X}\left[\|A_X(x) - A_Y(\mathcal{G}(x))\|_1\right] + \mathbb{E}_{y \in Y}\left[\|A_Y(y) - A_X(\mathcal{F}(y))\|_1\right] \tag{7}$$

In addition, we introduce a sparse loss to encourage the attention network to pay attention to a small region related to the object instead of the whole image:

$$\mathcal{L}_{sparse}(A_X, A_Y) = \mathbb{E}_{x \in X}\left[\|A_X(x)\|_1\right] + \mathbb{E}_{y \in Y}\left[||A_Y(y)||_1\right]. \tag{8}$$

Considering Equation (7), the attention maps of $A_X(\mathcal{F}(y))$ and $A_Y(\mathcal{G}(x))$ should be consistent to $A_Y(y)$ and $A_X(x)$, so they do not include additional sparse loss on $A_X(\mathcal{F}(y))$ and $A_Y(\mathcal{G}(x))$.

Hence, by combining Equations (1-3), (7) and (8), our full objective is:

$$\begin{aligned}\mathcal{L}(T_X, T_Y, D_X, D_Y, A_X, A_Y) &= \mathcal{L}_{GAN}(\mathcal{G}, D_Y, X, Y) + \mathcal{L}_{GAN}(\mathcal{F}, D_X, X, Y) \\ &+ \lambda_{cyc}\mathcal{L}_{cyc}(\mathcal{G}, \mathcal{F}) + \lambda_{A_{cyc}}\mathcal{L}_{A_{cyc}}(A_X, A_Y) + \lambda_{sparse}\mathcal{L}_{sparse}(A_X, A_Y),\end{aligned} \tag{9}$$

where λ_{sparse} and λ_{cyc} balance the relative importance of different terms. Attention network, transformation network and discriminative network in both X domain and Y domain can be solved in the following min-max game:

$$\arg \min_{T_X, T_Y, A_X, A_Y} \max_{D_X, D_Y} \mathcal{L}(T_X, T_Y, D_X, D_Y, A_X, A_Y), \tag{10}$$

the optimization algorithm is described in the supplementary material.

4.3 Extra Supervision

In some cases, segmentation annotations can be collected and used as attention map. For example, our horse $\rightarrow$ zebra image segmentation of horse is exactly the region of interest. We therefore supervise the training of the attention network by segmentation label. Given a training set $\{(x_1, m_1), \cdots, (x_N, m_N)\}$ of N examples, where m_i indicates the binary labels of segmentation, we minimize the discrepancy between predicted attention maps $A(x_i)$ and segmentation label m_i. To learn the attention maps for both X domain and Y domain, the total attention loss can be written as:

$$\mathcal{L}_{A_{sup}}(A_X, A_Y) = \sum_{i=1}^{N_X} \|m_i - A_X(x_i)\|_1 + \sum_{j=1}^{N_Y} \|m_j - A_Y(y_j)\|_1. \tag{11}$$

The full objective thus becomes:

$$\begin{aligned}\mathcal{L}(T_X, T_Y, D_X, D_Y, A_X, A_Y) &= \mathcal{L}_{GAN}(\mathcal{G}, D_Y, X, Y) + \mathcal{L}_{GAN}(\mathcal{F}, D_X, X, Y) \\ &+ \lambda_{cyc}\mathcal{L}_{cyc}(\mathcal{G}, \mathcal{F}) + \lambda_{A_{sup}}\mathcal{L}_{A_{sup}}(A_X, A_Y),\end{aligned} \tag{12}$$

where λ_{cyc} and $\lambda_{A_{sup}}$ control the relative importance of the objectives. As the attention maps are supervised by semantic annotations, we do not incorporate the constraints of Equations (7) and (8).

5 Experiments

In this section, we first introduce two metrics to evaluate the quality of generated images. We then compare unsupervised Attention-GAN against CycleGAN [7]. Next, we study the importance of the *attention sparse loss*, and compare our method against some variants. Lastly, we demonstrate empirical results of supervised Attention-GAN.

We first evaluated the proposed Attention-GAN on three tasks: horse $\leftrightarrow$ zebra, tiger $\leftrightarrow$ leopard and apple $\leftrightarrow$ orange. The images for horse, zebra, apple and orange were provided by CycleGAN [7]. The images for tiger and leopards are from ImageNet [12], which consists of 1,444 images for tiger, 1,396 for leopard. We randomly selected 60 images for test, and the rest for training set. In supervised experiment, we performed horse $\leftrightarrow$ zebra task where images and annotations can be obtained from MSCOCO dataset [34]. For each object category, images in MSCOCO training set were used for training and those in MSCOCO val set were for testing. For all experiments, the training samples were first scaled as 286×286, and then randomly flipped and cropped as 256×256. In test phase, we scaled input images to the size of 256×256.

For all experiments, the networks were trained with an initial learning rate of 0.0002 for the first 100 epoch and a linearly decaying rate that goes to zero over the next 100 epochs. We used the Adam solver [35] with batch size of 1. We updated the discriminative networks using a randomly selected sample from a buffer of previously generated images followed by [36]. The training process is shown in supplementary material. The architectures of transformation networks and attention networks are based on Johnson *et al.* [37]. The discriminators are adapted from the Markovian Patch-GAN [38, 2, 7, 9]. Details are listed in the supplementary material.

5.1 Assessment of Image Quality

Since object transiguration is required to predict the region of interest and transform the object while preserve the background, we introduce metrics to estimate quality of transformed image.

To assess the background consistency of transformation, we compute PSNR and SSIM between generated image background and original image background. PSNR is an approximation to human perception of reconstruction quality, which is defined via mean squared error (MSE). Given testing samples $\{(x_1, m_1), \cdots, (x_N, m_N)\}$, we use pixels-wise multiplication $\odot$ by the segmentation mask to compute image background PSNR:

$$\frac{1}{N}\sum_{i=1}^{N} PSNR\left(x_i \odot (1 - m_i), \mathcal{G}(x_i) \odot (1 - m_i)\right), \tag{13}$$

where x_i indicates original image, $\mathcal{G}(x_i)$ indicates the resulting image, $(1 - m_i)$ indicates the image background, the pixels-wise multiplication $x_i \odot (1 - m_i)$ indicates the background of original image, and $\mathcal{G}(x_i) \odot (1 - m_i)$ indicates the background of generated image. Similarly, we use SSIM to assess the structural

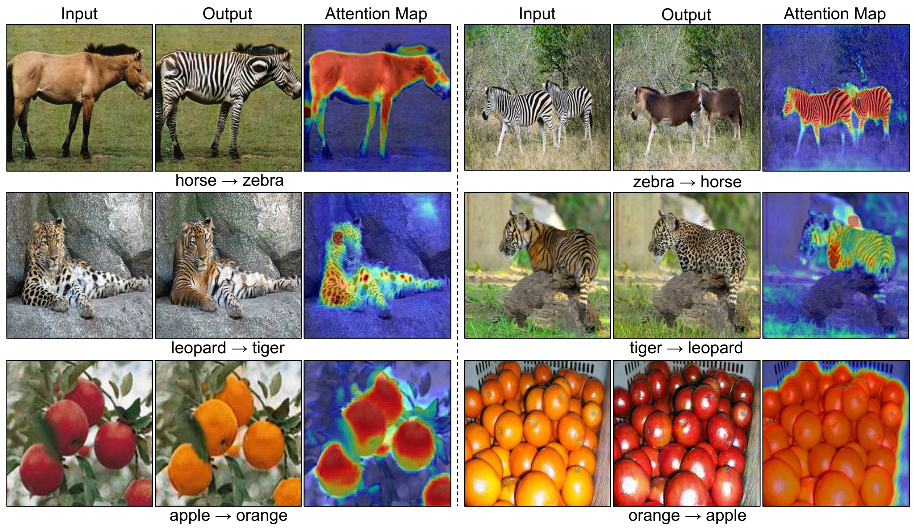


Fig. 3. Results of object transfiguration on different tasks: horse ↔ zebra, leopard ↔ tiger and apple ↔ orange. In each case, the first image is the original images, the second image is the synthesized image, and the third image is the predicted attention map. Our proposed model only manipulates the attention parts of image and preserves the background consistency.

Table 1. Background consistency performance of different object transfiguration tasks for background PSNR and SSIM.

	Task	CycleGAN	DistanceGAN	Ours (Unsupervised)	Ours (Supervised)
PSNR	horse → zebra	18.1875	11.1896	22.2629	24.589
	zebra → horse	18.1021	10.1153	21.5360	23.9330
SSIM	horse → zebra	0.6725	0.2630	0.9003	0.9482
	zebra → horse	0.7155	0.3627	0.8988	0.9534

similarity between background of original image and composited output by using pixels-wise multiplication:

$$\frac{1}{N}\sum_{i=1}^{N} SSIM\left(x_i \odot (1-m_i)\,, y_i \odot (1-m_i)\right). \tag{14}$$

In experiment, we use MSCOCO [34] dataset's test images and segmentation mask to evaluate background quality of generated image.

5.2 Unsupervised Results Comparisons to State-of-the-Art

Quantitative Comparison. We compare our method with CycleGAN [7] and DistanceGAN [11] by computing the image background PSNR and SSIM (Equations (13) and (14)). The test dataset is from MSCOCO dataset [34]. As MSCOCO

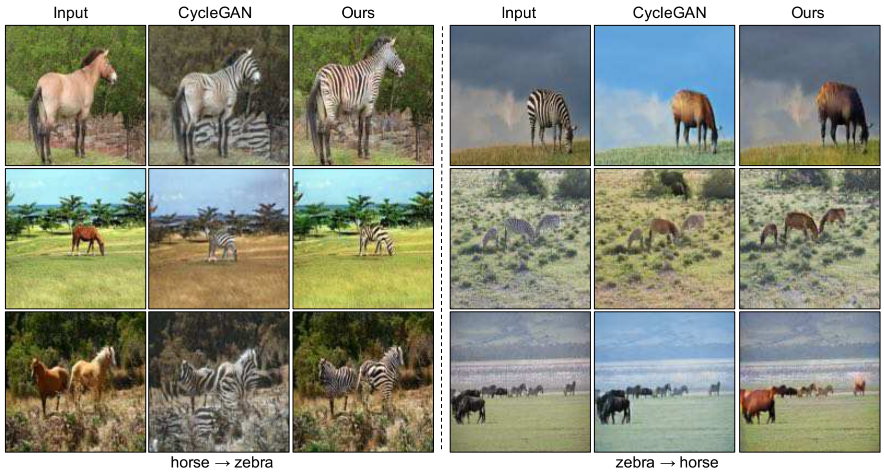


Fig. 4. Comparison with CycleGAN on horse ↔ zebra. In each case, the first image is the input image, the second is the result of CycleGAN [7], and the third is the result of our Attention-GAN.

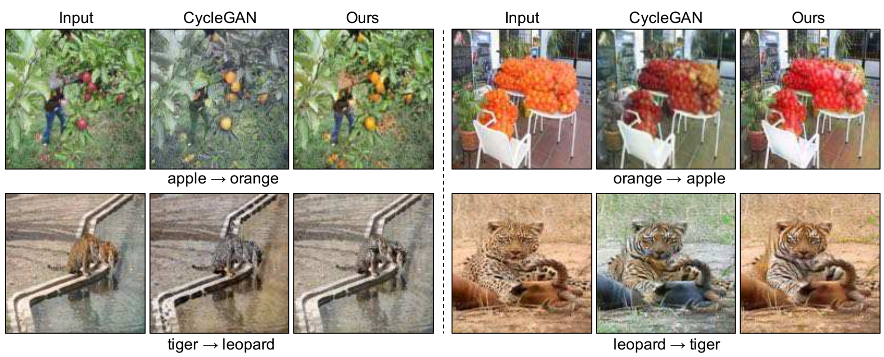


Fig. 5. Comparison with CycleGAN on apple ↔ orange and tiger ↔ leopard. In each case: input image (left), result of CycleGAN [7] (middle), and result of our Attention-GAN (right).

dataset does not have the classes of tiger or leopard, and apples and oranges in images are too small, we only compare the results of horse ↔ zebra. Results are shown in Table 1. As can be seen, for both PSNR and SSIM, our method in unsupervised fashion outperforms CycleGAN and DistanceGAN, which indicates that the proposed model predicts accurate attention map and achieves a better performance of transformation quality. Since our method outperforms DistanceGAN by a large margin, we only explore qualitative quality and human perceptual study with CycleGAN.

Qualitative comparison Results of horse ↔ zebra are shown in Fig. 4. We observed that our method provides translation results of higher visual quality on test data compared to those of CycleGAN. For example, in the horse → zebra task, CycleGAN mistakes some parts of background as target and trans-

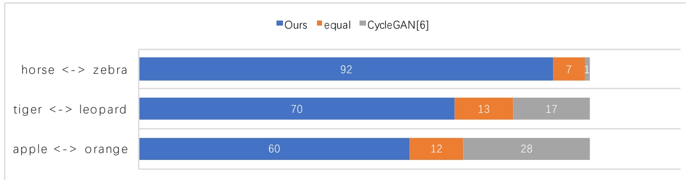


Fig. 6. The stacked bar chart of participants preferences for our methods compared to CycleGAN [21]. The blue bar indicates the number of images that more participants prefer our results. The gray bar indicates the number of images that more participants prefer CycleGAN's results. The orange bar indicates the number of images where two methods get a equal number of votes from 10 participants.

forms them into black and white stripes. In second column of Fig 4, CycleGAN translates the green grass and trees into brown in the zebra $\rightarrow$ horse task. In contrast, our method generates zebra in the correct location and preserves background consistency. Comparison results on tiger $\leftrightarrow$ leopard and apple $\leftrightarrow$ orange are shown in Figure 5. The results of Attention-GAN are more visually pleasing than those of CycleGAN. In most cases, CycleGAN can not preserve background consistency, e.g., the blue jeans in the first image are transformed to yellow, the blue water in third image is transformed to yellow and the yellow weeds in the last image is transformed to green. One possible reason is that our Attention-GAN disentangles the background and object of interests by the attention network and only transforms the object, while the compared method only uses one generative network that manipulate the whole image.

Human Perceptual Study We further evaluate our algorithm via a human study. We perform pairwise A/B tests deployed on the Amazon Mechanical Turk platform. We follow the same experiment procedure in [39, 40]. The participants are asked to select the more realistic image from each pair. Each pair contains two images translated from the same source image by two approaches. We test the tasks of horse $\leftrightarrow$ zebra, tiger $\leftrightarrow$ leopard and apple $\leftrightarrow$ orange. In each task, we randomly select 100 images from test set. Each image are compared by 10 participants. Figure 6 shows the participants preference among 100 examples. We observe that 92 results of our methods outperforms results of CycleGAN in horse $\leftrightarrow$ zebra task. In tiger $\leftrightarrow$ leopard, still only 17% results of compared method beat ours, which indicates that qualitative assessments obtained by our proposed approaches are better than those obtained by existing methods. We also notice that in apple $\leftrightarrow$ orange task, only 60 results of our methods outperform the compared method. We consider the reason is that a large portion of images in apple and orange dataset are close-up images whose background is simple so that CycleGAN could reach a competitive result.

5.3 Model Analysis

We perform model analysis on the horse $\rightarrow$ zebra task. Figure 7 shows the generated images, along with the intermediate generation results of model. In the

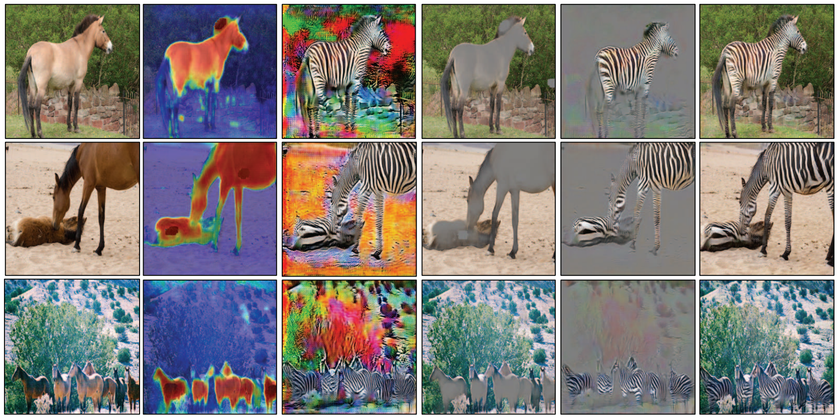


Fig. 7. Generation results of our model on horse → zebra. From left to right: Inputs, attention maps, outputs of transformation network, background images factorized by attention maps, object of images factorized by attention maps, final composite images.

Table 2. Performance of horse → zebra for different losses.

	$\lambda_{attn} = 0$	$\lambda_{attn} = 1$	$\lambda_{attn} = 5$
PSNR	19.8621	22.2629	24.2173
SSIM	0.8291	0.9003	0.9367

second column, the attention maps with are shown. As can be seen, while being completely unsupervised, the attention network of model is able to successfully disentangle the objects of our interests and the background from input image. The third column is the output of the transformation network, where the transformed zebra are visually pleasing while the background parts of images are meaningless. It demonstrates that the transformation network only focuses on transforming the object of interests. Moreover, Figure 7 shows that the final output images in the last column are combined by the background parts in the forth column and the objects of interests in the fifth column.

Figure 8 shows the qualitative results of variants of our model on horse → zebra. It can be seen that without the sparse loss ($\lambda_{sparse} = 0$ in Equation (8)), the attention network would predict some parts of image background as regions of interests. When λ_{sparse} was set to 5, the attention mask shrinked too much to cover the whole object of interests. It is because if we emphasize too much on the relative importance of sparse loss, the attention network can not comprehensively predict the object location. We find $\lambda_{sparse} = 1$ is an appropriate choice, which makes a good balance to pay enough attention on the objects of interests. In Table 2, we observe that with the value of λ_{sparse} becoming larger, the performance of background consistency is better. However, the qualities of transformed object decrease if λ_{sparse} is set too large. This indicates that the λ_{sparse} can be viewed as a parameter that balance the performance of background consistency and transformation quality.

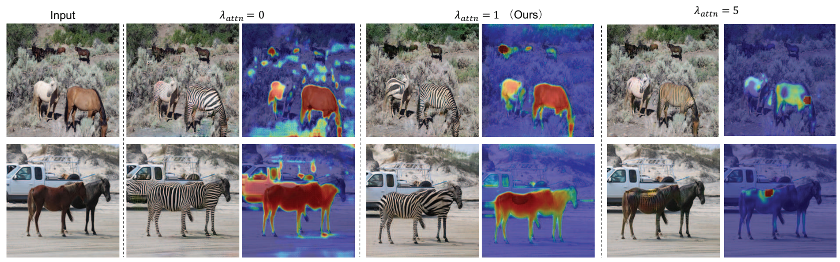


Fig. 8. The effect of sparse loss with different parameters λ_{sparse} for mapping horse $\rightarrow$ zebra. From left to right: input, output and attention map without sparse loss, input and attention map when $\lambda_{sparse} = 1$, input and attention map when λ_{sparse}=5.

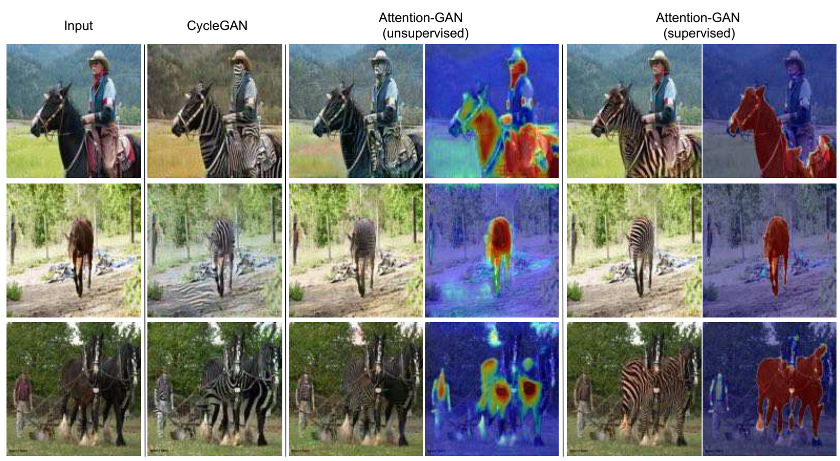


Fig. 9. Comparison of horse $\leftrightarrow$ zebra between CycleGAN [7], unsupervised Attention-GAN and supervised Attention-GAN.

5.4 Comparisons of Supervised Results

We compute PSRN, SSIM of background region between generated and original images in horse $\leftrightarrow$ zebra task. In Table 1, the Attention-GAN with supervision outperforms unsupervised Attention-GAN and CycleGAN from the perspective of background consistency. This demonstrates that the attention network predicts the object of interests more accurately with the segmentation mask. In Figure 9, CycleGAN and unsupervised Attention-GAN predict some parts of the person as region of interests and transform them into texture of zebra (see the first row of Figure 9). We also notice that the attention maps with supervision tend to be dark red or dark blue, which indicates the supervised attention network predicts with higher confidence, and disentangles the background and object of interests more clearly.

We evaluate the foreground mask of horse in terms of UoI and mAPr@0.5. The unsupervised Attention-GAN got 28.1% of UoI and 20.3% of mAPr@0.5. On the other hand, the supervised Attention-GAN got 37.8% of UoI score and

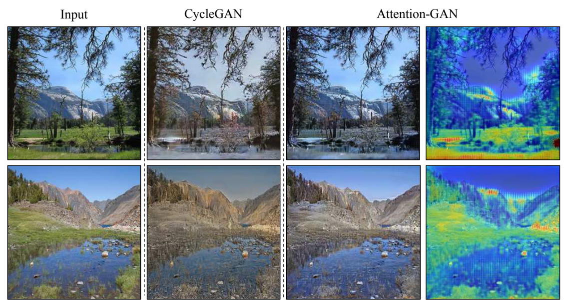


Fig. 10. Results of Summer → Winter comparing with CycleGAN

30.5% of mAPr@0.5. Although our algorithm is not particularly designed for semantic segmentation, the proposed attention network is able to learn the object of interests in an unsupervised way and achieve a reasonable performance.

5.5 Global Image Transformation

Both local and global image transformation are important. We study object transfiguration, and evaluate it on horse ↔ zebra, apple ↔ orange and tiger ↔ leopard. More applications include virtual try-on [41] with regard to a desired clothing item of a person, and face attributes (e.g. mustache and glass) changing [42]. The proposed attention GAN is effective to identify important regions in object transfiguration problems, and it can also lead to some interesting observations in global image transformation. In summer ↔ winter, there is no explicit object of interests, but the algorithm does recognize some regions with more attention, e.g. grass and trees in Fig. 10, which are usually green in summer and brown in winter. Meanwhile, regions without distinctive characteristics, e.g., blue sky would not be attended.

6 Conclusion

This paper introduces attention mechanism into the generative adversarial nets considering image context and structure information on object transfiguration task. We develop a three-player model that consists of an attention network, a transformation network and a discriminative network. The attention network predicts the regions of interest whilst the transformation network transforms the object from one class to another. We show that our model has advantages over the one-shot generation method [7] in preserving background consistency and transformation quality.

Acknowledgment This work was supported in part by State Key Research and Development Program (2016YFB1001003), NSFC (61527804, 61521062), in part by Australian Research Council Projects FL-170100117, DE-180101438, DP-180103424 and LP-150100671, and in part by USyd-SJTU Partnership Collaboration Award.

References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3431–3440
2. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 (2016)
3. Karacan, L., Akata, Z., Erdem, A., Erdem, E.: Learning to generate images of outdoor scenes from attributes and semantic layouts. arXiv preprint arXiv:1612.00215 (2016)
4. Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: Controlling deep image synthesis with sketch and color. arXiv preprint arXiv:1612.00835 (2016)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2) (2016) 295–307
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
7. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
8. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. (2017) 1857–1865
9. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
10. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. CoRR **abs/1611.02200** (2016)
11. Benaim, S., Wolf, L.: One-sided unsupervised domain mapping. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. (2017) 752–762
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 248–255
13. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in neural information processing systems. (2015) 1486–1494
14. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
15. Wang, C., Wang, C., Xu, C., Tao, D.: Tag disentangled generative adversarial network for object image re-rendering. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. (2017) 2901–2907
16. Im, D.J., Kim, C.D., Jiang, H., Memisevic, R.: Generating images with recurrent adversarial networks. arXiv preprint arXiv:1602.05110 (2016)
17. Kwak, H., Zhang, B.T.: Generating images part by part with composite generative adversarial networks. arXiv preprint arXiv:1607.05387 (2016)

18. Yang, J., Kannan, A., Batra, D., Parikh, D.: Lr-gan: Layered recursive generative adversarial networks for image generation. In: 5th International Conference on Learning Representations (ICLR). (2017)
19. Wang, C., Xu, C., Wang, C., Tao, D.: Perceptual adversarial networks for image-to-image transformation. IEEE Transactions on Image Processing **27**(8) (2018) 4066–4079
20. Balcan, M., Weinberger, K.Q., eds.: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. Volume 48 of JMLR Workshop and Conference Proceedings., JMLR.org (2016)
21. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
22. Liang, X., Zhang, H., Xing, E.P.: Generative semantic manipulation with contrasting gan. arXiv preprint arXiv:1708.00315 (2017)
23. Rensink, R.A.: The dynamic representation of scenes. Visual cognition **7**(1-3) (2000) 17–42
24. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: Advances in neural information processing systems. (2014) 2204–2212
25. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2921–2929
26. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 842–850
27. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. (2015) 2048–2057
28. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision, Springer (2016) 451–466
29. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3640–3649
30. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3th International Conference on Learning Representations (ICLR). (April 2015)
31. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE international conference on computer vision. (2015) 4507–4515
32. Gregor, K., Danihelka, I., Graves, A., Rezende, D., Wierstra, D.: Draw: A recurrent neural network for image generation. In Bach, F., Blei, D., eds.: Proceedings of the 32nd International Conference on Machine Learning. Volume 37 of Proceedings of Machine Learning Research., Lille, France, PMLR (07–09 Jul 2015) 1462–1471
33. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)

34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755
35. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
36. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
37. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, Springer (2016) 694–711
38. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European Conference on Computer Vision, Springer (2016) 702–716
39. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: The IEEE International Conference on Computer Vision (ICCV). Volume 1. (2017)
40. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. arXiv preprint arXiv:1711.11585 (2017)
41. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2018)
42. Upchurch, P., Gardner, J., Pleiss, G., Pless, R., Snavely, N., Bala, K., Weinberger, K.: Deep feature interpolation for image content changes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)