DA-GAN: Instance-level Image Translation by Deep Attention Generative Adversarial Networks

Shuang Ma*, Jianlong Fu†, Chang Wen Chen*♯ and Tao Mei †
*Computer Science and Engineering, SUNY at Buffalo †Microsoft Research
♯School of Science and Engineering, The Chinese University of Hong Kong
{shuangma,chencw}@buffalo.edu jianf@microsoft.com tmei@live.com

Abstract

Unsupervised image translation, which aims in translating two independent sets of images, is challenging in discovering the correct correspondences without paired data. Existing works build upon Generative Adversarial Networks (GANs) such that the distribution of the translated images are indistinguishable from the distribution of the target set. However, such set-level constraints cannot learn the instance-level correspondences (e.g. aligned semantic parts in object transfiguration task). This limitation often results in false positives (e.g. geometric or semantic artifacts), and further leads to mode collapse problem. To address the above issues, we propose a novel framework for instance-level image translation by Deep Attention GAN (DA-GAN). Such a design enables DA-GAN to decompose the task of translating samples from two sets into translating instances in a highly-structured latent space. Specifically, we jointly learn a deep attention encoder, and the instance-level correspondences could be consequently discovered through attending on the learned instances. Therefore, the constraints could be exploited on both set-level and instance-level. Comparisons against several state-of-the-arts demonstrate the superiority of our approach, and the broad application capability, e.g, pose morphing, data augmentation, etc., pushes the margin of domain translation problem. [1]

1. Introduction

Can machines possess human ability to relate different image domains and translate them? This question can be formulated as image translation problem. In other words, learning a mapping function, by finding some underlying correspondences (e.g. similar semantics), from one image domain to the other. Years of research have produced powerful translation systems in supervised setting, where example pairs are available, e.g. [14]. However, obtaining paired training data is difficult and expensive. Therefore, researchers turned to develop unsupervised learning approach which only relies on unpaired data.

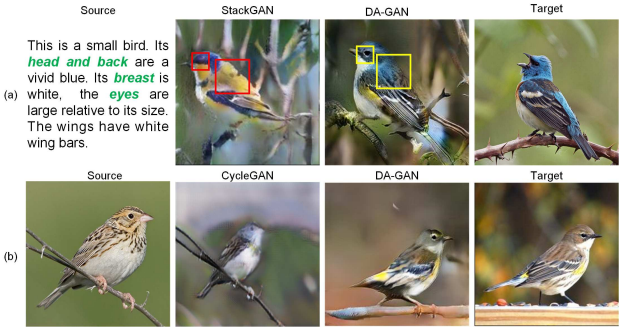

Figure 1: (a) text-to-image generation. (b) object configuration. We can observe that the absence of instance-level correspondences results in both semantic artifacts (labeled by red boxes) exist in StackGAN and geometry artifacts exist in CycleGAN. Our approach successfully produces the correct correspondences (labeled by yellow boxes) because of the proposed instance-level translating. Details can be found in Sec. 1

Following the principle of translation, i.e. remaining the expected identity from source domain (e.g., semantics in text-to-image, human-ID in face-to-animation), while generating samples that match the distribution of target domain, existing works typically build upon Generative Adversarial Networks (GANs). However, they are trained only on a holistic characterization of the data distribution, while lacks an inference mechanism to reason about data at abstract level. The implicit training process and weak controllability prevents them from finding meaningful instance-level correspondences, and the identity is thus hard to be governed. By 'instance-level correspondences', we refer to high-level content involving identifiable objects that shared by a set of samples. These identifiable objects could be adaptively

[1] This work was done while Shuang's internship in Microsoft Research

task driven. For example, in Fig 1 (a), the words in the description corresponds to according parts and attributes of the bird image. Therefore, false positives often occur because of the instance-level correspondences missing in existing works. For example, in object transfiguration, the results just showing changes of color and texture, while fail in geometry changes (Fig. 1). In text-to-image synthesis, fine-grained details are often missing (Fig. 1).

Driven by this important issue, a question arises: *Can we seek an approach which is capable of finding meaningful correspondences from both set-level and instance-level under unsupervised setting?* To resolve this issue, in this paper, we introduce a dedicated unsupervised domain translation approach builds upon Generative Adversarial Networks - **DA-GAN**. It provides the first solution by decomposing the task of translating samples from two independent sets into translating instances in a highly-structured latent space. Specifically, we jointly learn a *Deep Attention Encoder* (DAE) to integrate the attention mechanism [7] into the learning of the mapping function F. The learning consists of two closely-related steps: 1) DAE first decomposes the original data into instances, which makes it possible to find correct semantic alignments (Fig.1(a)) and exploit geometry changes (Fig.1(b)); 2) once the aligned instances have been inferred, the instance-level constraints are employed (Eq.5,6), in which the identity can be exclusively governed but not interweaves with other factors from the target domain (e.g. different styles in face-to-animation). As a result, the instance-level constraints enable the mapping function to find the meaningful semantic corresponding, and therefore producing true positives and visually appealing results. Compared with existing works which uses foreground masks LRGAN [41] or normal maps SSGAN [39] to retain the identity in a specific task, the proposed DA-GAN can automatically and adaptively learn task-driven identity representations by attention mechanism without human-involvement. Our main contributions can be summarized into three-fold:

- We decompose the task to instance-level image translation such that the controllability is much enhanced, and the could be exploited on both instance-level and set-level.
- To the best of our knowledge, we are the first that integrate the attention mechanism into Generative Adversarial Networks.
- We introduce a novel framework DA-GAN, which produces visually appealing results and is applicable in a large variety of tasks.

2. Related Work

Generative Adversarial Networks

Since the Generative Adversarial Networks (GANs) was proposed by Goodfellow et al., [10] researchers have studied it vigorously. Several techniques have been proposed to stabilize the training techniques [25, 22, 28, 1, 44] and generate compelling results. Built upon these generative models, several methods were developed to generate images based on GAN. Most methods utilized conditioning variables such as attributes or class labels [40, 37, 5, 24, 11]. There are also works conditioned on images to generate images, e.g. photo editing [4, 13], and super-resolution [19, 33]. Other approaches used conditional features from a completely different domain for image generation. Reed et al. [26] used encoded text description of images as the conditional information to generating 64×64 images that match the description. Their follow-up work [26] can produce 128×128 images by utilizing additional annotations on object part locations. In StackGAN [43], two GANs in different stages are adopted to generate high resolution images. Comparing with StackGAN, the proposed DA-GAN can generated 256×256 images directly. More importantly, we trained the network by unpaired data, and achieve visually appealing results.

Image-to-Image Translation

'pix2pix'[14] of Isola et al. uses a conditional GAN [10] to learn a mapping from input to output images. Similar ideas have been applied to various tasks such as generating photographs from sketches [30] or from attribute and semantic layouts [16]. Recently, [8] proposed the domain transformation network (DTN) and achieved promising results on translating small resolution face and digit images. CoGAN [21] and cross modal scene networks [2] use a weight-sharing strategy to learn a common representation across domains. Another line of concurrent work [3, 31, 34] encourages the input and output to share certain content features even though they may differ in style. They also use adversarial networks, with additional terms to enforce the output to be close to the input in a predefined metric space, such as class label space, image pixel space, and image feature space. In CycleGAN [45], a cycle consistency loss is proposed to enforce one-to-one mapping. We note that several contemporary works [42, 17] are all introduced the cycle-consistency constraint for the unsupervised image translation. Neural Style Transfer [9, 15, 36] is another way to perform image-to-image translation, which synthesizes an image by combining the content of one image with the style of another image based on pre-trained deep features. Different with style transfer, domain translation aims in learning the mapping between two image collections, rather than between two specific images.

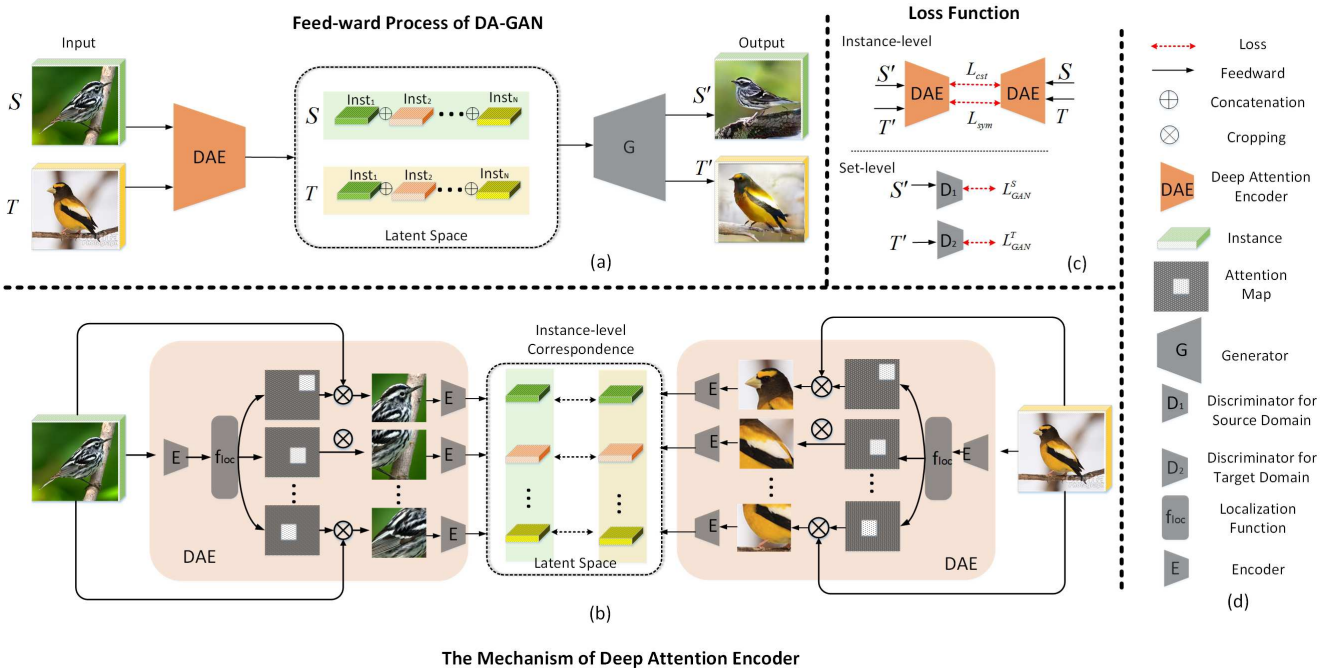

Figure 2: A pose morphing example for illustration the pipeline of DA-GAN. Given two images of birds from source domain S and target domain T, the goal of pose morphing is to translate the pose of source bird s into the pose of target one t, while still remain the identity of s. The feed-ward process is shown in (a), where two input images are fed into DAE which projects them into a latent space (labeled by dashed box). Then G takes these structured representations ($DAE(s)$ and $DAE(t)$) from the latent space to generated the translated samples, i.e.$s' = G(DAE(s))$, $t' = G(DAE(t))$. The details of the proposed DAE (labeled by orange block) is shown in (b). Given an image X, a localization function f_{loc} first predicts N attention regions' coordinates from the feature map of X, (i.e. $E(X)$, where E is an encoder, which can be utilized in any form). Then N attention masks are generated and activated on X to produce N attention regions $\{R_i\}_{i=1}^{N}$. Finally, each region's feature consists the instance-level representations $\{Inst_i\}_{i=1}^{N}$. By operating the same way on both S and T, the instance-level correspondences can consequently be found in the latent space. We exploit constraints on both instance-level and set-level for optimization, it is illustrated in (c). All of the notations are listed in (d). [Best viewed in color.]

3. Approach

Our aim is to learn a mapping function F that maps samples from source domain $S : \{s_i\}_{i=1}^{N}$ to target domain $T : \{t_i\}_{i=1}^{M}$, denoted as $F : S \rightarrow T$. As illustrated in Fig. 2, the proposed DA-GAN consists of four modules: a Deep Attention Encoder (DAE), a Generator(G) and two discriminators (D_1, D_2). The mapping is conducted from both source domain and target domain. The translated samples sets from source domain and target domain are denoted as S' and T', respectively. We introduce the DAE in Sec. 3.1. The translation on instance-level and set-level are introduced in Sec. 3.2 and in Sec. 3.3, respectively.

3.1. Deep Attention Encoder

To project samples into the latent space, we integrate attention mechanism to jointly learn an Deep Attention Encoder DAE. Given a feature map $E(X)$ of an input image X (where E is an encoder that could be utilized in any form), we first adopt a localization function $f_{loc}(\cdot)$ to predict a set of attention regions' location, which is given by:

$$f_{loc}(E(X)) = [x_i, y_i]_{i=1}^{N'}, \tag{1}$$

where $[x_i, y_i]$ denotes a region's center coordinates, N' denotes the number of regions predicted. Once the location of an attended region is hypothesized, we generate an attention mask $\mathcal{M}_i$. Specifically, we denote w and h as half of the width and half of the height of X. Then we can adopt the parameterizations of attend region by:

$$\begin{aligned} x_i^{left} &= x_i - w, \quad x_i^{right} = x_i + w, \\ y_i^{top} &= y_i - h, \quad y_i^{bottom} = y_i + h. \end{aligned} \tag{2}$$

The cropping operation can therefore be achieved by an element-wise multiplication applied on X, i.e. $R_i = X \circ \mathcal{M}_i$, which produces the attended regions $\{R_i\}_{i=1}^{N'}$. Then instance-level representations of X in the latent space are defined by:

$$\{E(R_i)\}_{i=1}^{N'} = \{Inst\}_{i=1}^{N'}, \tag{3}$$

To allow back-propagation, here we adopt the attention mask as:

$$\begin{aligned}\mathcal{M}_i = [\sigma(x - x_i^{left}) - \sigma(x - x_i^{right})]\cdot \\ [\sigma(y - y_i^{top}) - \sigma(y - y_i^{bottom})],\end{aligned} \tag{4}$$

where $\sigma(\cdot) = 1/(1+\exp^{-kx})$ is a sigmoid function. In theory, when k is large enough, $\sigma(\cdot)$ is approximated as a step function and $\mathcal{M}_i$ will become a two dimensional rectangular function, then the derivation could be approximated. For learning these attention regions, we add a geometric regularization $\mathbb{E}_{X\sim P_{data}(X)}[d(Y, DAE(X))]$. Y is the label of image X, and d is some similarity metrics in the data space, In practice, there are many options for the distance measure d. For instance, a VGG classifier.

3.2. Instance-Level Image Translation

As the DAE projects s and t into a shared latent space, we can constrain them to be matched with each other in this latent space. Therefore, we adopt a consistency loss on the samples from source domain $\{s_i\}_{i=1}^N$ and the according translated samples $\{s'_i\}_{i=1}^N$:

$$\mathcal{L}_{cst} = \mathbb{E}_{s\sim P_{data}(s)}\, d(DAE(s),\ DAE(F(s))), \tag{5}$$

On the other hand, we also consider the samples from the target domain to further enforce the mapping to be deterministic. In theory, if a mapping is bijective (one-to-one corresponding), the operation from a set to itself form a symmetric group. The mapping can then be considered as a permutation operation on itself. We therefore exploit a symmetry loss to enforce F can map samples from T to themselves, i.e. $t_i \approx F(t_i)$. The loss function is defined as:

$$\mathcal{L}_{sym} = \mathbb{E}_{t\sim P_{data}(t)}\, d(DAE(t),\ DAE(F(t))), \tag{6}$$

this can also be considered as an auto-encoder type of loss applied on samples from T, where d is a distance measure. In theory, there are many options for d. For instance, the L^n distance, or the distance of learned features by the discriminator or by other networks, such as a VGG classifier.

3.3. Set-Level Image Translation

It is straight-forward to use a discriminator D_1 to distinguish the translated samples $\{s'_i\}_{i=1}^N$ from the real samples in the target domain $\{t\}_{i=1}^M$, and generator is forced to translate samples that is indistinguishable from real samples in target domain, which is given by:

$$\begin{aligned}\mathcal{L}_{GAN}^s = \mathbb{E}_{t\sim P_{data}(t)}[\log D_1(t)] \\ + \mathbb{E}_{t\sim P_{data}(s)}[\log(1 - D_1(F(s)))].\end{aligned} \tag{7}$$

While there still exists another issue - mode collapse. In theory, large modes usually have a much higher chance of attracting the gradient of discriminator, and the generator is not penalized for missing modes. In practice, all input samples map to the same output, and the optimization fails to make progress. This issue asks for adding penalty on generator for missing modes.

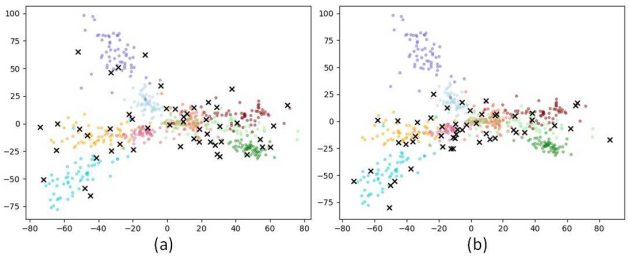

Figure 3: Visualized distribution of 10 classes of birds. Each color represents a birds class. Black crosses represents the distribution of the generated samples. (a): generated data distribution of DA-GAN. (b) generated data distribution of StackGAN [43].

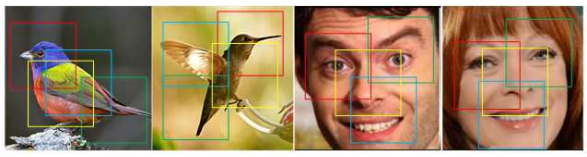

Figure 4: The attention locations predicted by DAE on birds images and face images.

As we mentioned before, $DAE \circ G$ can be considered as an auto-encoder for $\{t_i\}_{i=1}^M$. Then for every modes in T, $F(t)$ is expected to generate very closely located modes. We therefore add another discriminator D_2 for samples from the target domain to enforce the reconstructed t' is indistinguishable from t. An additional optimization objective for the generator is hence added $\mathbb{E}_{t\sim p_{data}(t)}[\log D_2(F(t))]$. The objective function is given by:

$$\begin{aligned}\mathcal{L}_{GAN}^t = \mathbb{E}_{t\sim P_{data}(t)}[\log D_2(t)] \\ + \mathbb{E}_{t\sim P_{data}(t)}[\log(1 - D_2(F(t)))].\end{aligned} \tag{8}$$

This multi-adversarial training procedure is critical for penalizing the missing modes, it encourage $F(t)$ to move towards a nearby mode of the data generating distribution. In this way, we can achieve fair probability mass distribution across different modes.

3.4. Full Objective and Implementation Details

Our full objective is given by:

$$\begin{aligned}\mathcal{L}(DAE, G, D_1, D_2) = \mathcal{L}_{GAN}^s(DAE, G, D_1, S, T) \\ + \mathcal{L}_{GAN}^t(DAE, G, D_2, T) \\ + \alpha\mathcal{L}_{cst}(DAE, G, S) \\ + \beta\mathcal{L}_{sym}(DAE, G, T),\end{aligned} \tag{9}$$

where α and β are weights for the consistency loss and symmetry loss, respectively. We aim to solve:

$$F^* = \arg\min_F \max_{D_1, D_2} \mathcal{L}(F, D_1, D_2) \tag{10}$$

where $F = DAE \circ G$.

Figure 5: Examples results of text to image synthesis.

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 [35]	None	4	parts	L2
Colorization [45]	Binary	1	object	ResBlock

Table 1: Implementation details for learning the DAE.

During the training stage, the outputs of DAE are connected to two ways. One is connected to the generator, the other one is connected to the geometric regularization term (Sec. 3.1). We utilize different regularization terms for different tasks, which can be found in Table 1. The regularization term enables DAE learns to attend on meaningful instances. In the meantime, these instances are fed into G to generate the translated samples. Thus, the DAE and GAN are playing a collaborative game during training. At the testing stage, the outputs of DAE are just sent into G to produce the translated results.

We adopt the generator consists of several residual blocks [13]. For the generator, the instance-level representations are concatenated along the channel dimension and fed into several residual blocks. Finally, a series of up-sampling layers are used to generate a the translated image. For the discriminator, the generated image is fed through a series of down-sampling blocks. Finally, a fully-connected layer with one node is used to produce the decision score. The up-sampling blocks consist of the nearest-neighbor up-sampling followed by a 3×3 stride 1 convolution. Batch normalization and ReLU activation are applied after every convolution except the last one. The residual blocks consist of 3×3 stride 1 convolutions, Batch normalization and ReLU. All networks are trained using Adam solver with batch size 64 and an initial learning rate of 0.0002.

4. Experiments

In this section, we validate the effectiveness of the proposed DA-GAN in a large variety of tasks, including domain adaption, text-to-image synthesis, object configuration, pose morphing for data augmentation, face-to-animation synthesis and skeleton to cartoon figure synthesis. We conduct these experiments on several datasets, including MNIST [18], CUB-200-2011 [38], SVHN [12], FaceScrub [23] and AnimePlanet [2].

Experiments (Sec.4.2 Table 1(a)) further validate the

[2]It is retrieved from http://www.anime-planet.com/, which has about 60k images.

Figure 6: Example results of object configuration. Each row from top to bottom are the real samples, results generated by VAT[20], CycleGAN [45] and DA-GAN, respectively.

Figure 7: Example results of pose morphing. In each group, the first column are a source bird s, the second column are the target bird t, the third column are birds that generated by DA-GAN

contribution of DAE. Especially, it is important for more complected and structured 074 data. Without DAE, the accuracy for CUB-200-2011 drops 075 from 71.2

4.1. Baselines

- **GAN-INT-CLS** [26] succeeds in synthesizing 64 $\times$ 64 birds and flowers images based on text descriptions.
- **GAWWN** is Reed's follow-up work [27] that was able to generate 128 $\times$ 128 images.
- **StackGAN** is the latest work that can synthesize high-quality images in 256 $\times$ 256, from text descriptions.
- **SA** is an early work that explored ideas from subspace learning for domain adaption [6].
- **DANN** It is another domain adaption work that conducted by [8] deep feature learning.
- **UNIT** is a recent unsupervised image-to-image translation work [21] which based on the shared-latent space assumption and cycle loss.
- **DTN** [34] employs a compound loss function for unsupervised domain translation.
- **CycleGAN** is an image-to-image translation work that adopt GAN with cycle-loss [45, 17, 21, 42].
- **VAT** [20] is a new technique derives from style transfer, while it is different in finding dense correspondences.

4.2. Component Analysis of DA-GAN

We trained a classifier on MNIST dataset and employ it on the translated samples for quantitative evaluation. The results are shown in Table 2. As we can see, the DA-GAN approaches very high accuracy on the translated sample set. While the results is impaired without the DAE. We also fine-tune a VGG [32] classifier on on the CUB-200-2011, and use it to test our generated images from text, The accuracy drops a lot to 60.6 %. We also show some results produced by DAE in Fig. 4. It can be seen that f is capable of attending on semantic regions. For example, birds head, wings, and etc. human's eyes, mouth, and etc.

To validate that the proposed DA-GAN is effective in mitigating the mode collapse problem. We conduct a toy experiment on a subset of samples from CUB-200-2011. We select 10 classes of birds. To mimic the large mode, we picked some similar classes (e.g. some of them are from the same category). The dense region in Fig. 3 shows the birds that have similar looking. We generate about 600 images by input the according text descriptions and the distribution of the generated data is shown in Fig. 3(a). The same setting is conducted on StackGAN, the results is shown in 3(b). As we can see that, comparing with StackGAN, the samples generated by DA-GAN are more divers and have a larger coverage.

4.3. Domain Adaptation

We applied the proposed framework to the problem of domain adaptation, i.e. adapting a classifier trained using

(a) (b)

Figure 8: (a) Comparisons of face translation with VAT. Each row from left to right is human face, results produced by DA-GAN and VAT, respectively. (b) Human face to animation face synthesis. Human faces are placed in the first and third rows, the according translated animation faces are placed in the second row and fourth row.

Method	Accuracy
DA-GAN	94.3 %
DA-GAN w/o DAE	90.2 %
DA-GAN w/o const	79.8%
DA-GAN w/o Sym	90.6%
DA-GAN w/o D_2	88.2%

(a)

Method	Accuracy
SA[6]	59.32 %
DANN[8]	73.85 %
DTN [34]	84.44%
UNIT [21]	90.53 %
DA-GAN	**93.60** %

(b)

Method	Inception	♯ miss
GAN-INT-CLS	2.9 ± 0.4	89.0
GAWWN	3.6 ± 0.4	61.0
StackGAN	3.7 ± 0.4	36.0
DA-GAN	**5.6± 0.4**	**19.0**

(c)

Table 2: (a): Component evaluation of DA-GAN. (b): Comparisons with state-of-the-arts on domain adaption. (c): Comparisons with state-of-the-arts on text-to-image synthesis.

Figure 9: Example results of skeleton-cartoon synthesis.

labeled samples in one domain (source domain) to classify samples in a new domain (target domain) where labeled samples in the new domain are unavailable during training. For this purposes, we transform images from SHVN to the MNIST domain. The results of this experiment are reported in Table 2. We found that our method achieved a 94.6 % accuracy for the SVHN to MNIST translation task, which was much better than 90.53 % achieved by the previous state-of-the-art method.

4.4. Text to Image Synthesis

We conduct qualitative and quantitative evaluation on the text-to-image synthesis task. Comparisons with several state-of-the-arts [43, 27, 26] on CUB-200-2011 dataset are shown in Fig. 6. The quantitative evaluation are measured by two metrics: inception score [29] and the number of missing modes (denote as ♯ miss). In our experiments, we fine-tune a VGG19 model which is introduced in Sec. 4.2. While the inception score is considered as a good assessment for sample quality. However, the inception score is sometimes not a good metric for missing modes evaluation. For stronger validation, we adopt another evaluation metric - missing mode (♯miss) It represents the classifier reported number of missing modes, i.e. the size of the numbers that the model never generates. As shown in Table 2(c), DA-GAN achieves much improvements in terms of inception score, and the missing modes drop dramatically, which again proves the effectiveness of our proposed framework. Some examples results are shown in Fig. 5 for a visualized comparison.

4.5. Object Transfiguration

We use images of seven classes form the CUB-200-2011 dataset to perform object transfiguration, i.e. translate a source bird into a target breed. Some example results are show in Fig. 6. The first row is real samples form each breed, and we aims in translating bird (a) into the following six breeds. Among these selected target birds, (b) is selected as the most similar one with (a) in both spatial and geometry attributes. (c) is selected sharing similar spatial attribute while different in geometry attribute. (d-e) are all selected that have different spatial and geometry attributes

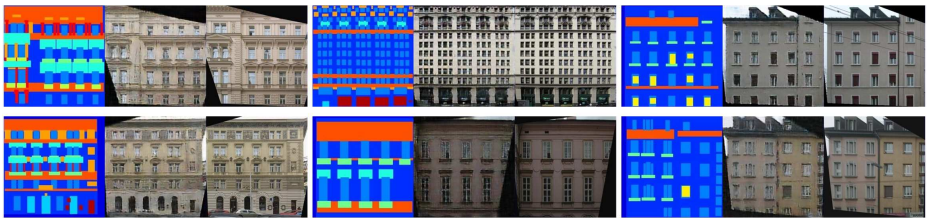

Figure 10: In each group from left to right are labels, translated architecture photos, and the ground truth.

Figure 11: Results of image colorization. In each group, the input is gray images, and the results are translated color images.

	Real	VAT	CycleGAN	DA-GAN
Top-1 acc	98.6 %	42.1 %	62.1 %	**88.9** %
Realism	20.9	10.3	11.4	**18.9**

Table 3: Qualitative evaluations for object configuration.

Method	Training Data	Accuracy
no data augmentation	8K	**79.0** %
DA-GAN	8K + 10K	**81.6** %

Table 4: Data augmentation results.

with (a). We can see that, without similar semantic structure, VAT [20] fails in translating birds, due to their limited corresponding matching method. CycleGAN [45] is robust to spatial changes while fail in changing the birds geometries. Comparing with the results that produced by DA-GAN, both shows blurred images that missing fine-grained details. We can see that, DA-GAN succeeds in translating images that even have large variance of spatial and geometry attributes. It strongly validates our claim that the instance-level corresponding is critical in translation task. We further conducted quantitative evaluation, which can be found in Table 3. The images produced by DA-GAN out performs in both classification accuracy and realism.

4.6. More Applications

We further conduct pose morphing, which considered harder in changing the geometries, by DA-GAN. The results are shown in Fig. 7. It can be seen that, we succeed in morphing the birds' pose even when there exists very large gap of geometry variance. For practical usage, we also make used of these morphed samples for data augmentation. For each image, we randomly picked 10 references as the pose targets. Top-1 result is picked for each image and is used for augmented data, which produced about 10K images of birds. We then applied a pre-trained VGG on the augmented data, which shows improvement on fine-grained classification task. The results is shown in Table 4.

We adopt the DA-GAN to translate a human face into a animation face while still preserve the human identity, the results are shown in Fig. 8. We also compare our results with the ones produces by VAT [20] in Fig. 8(a). We can see that, VAT cannot solve the task we are tackling. The produced images does not belong to the target domain, i.e. an animation face. More severely, when two query face shows different shooting angle, VAT produces artifacts due to the incorrect semantic correspondences. We also conducted experiments on skeleton to cartoon figure translation, and paired datasets (e.g. image colorization, labels to photos.) The results are shown in Fig. 9, Fig. 11 and Fig. 10. More experimental details can be found in supplementary materials.

5. Conclusion

In this paper, we propose a novel framework for unsupervised image translation. Our intuition is to decompose the task of translating samples from two sets into translating instances in a highly-structured latent space. The instance-level corresponding could then be found by integrating attention mechanism into GAN. Extensive quantitative and qualitative results validate that, the proposed DA-GAN can significantly improve the state-of-the-arts for image translation. It is superiority in scalable for broader application, and succeeds in generating visually appealing images. We find that, some failure cases are caused by the incorrect attention results. It is because the instances are learned by a weak supervised attention mechanism, which sometimes showing a large gap with that learned under fully supervision. To tackle this challenge we may seek for more robust and effective algorithm in the future.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 214–223, 2017.

[2] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. Cross-modal scene networks. *CoRR*, abs/1610.09003, 2016.

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *CoRR*, abs/1612.05424, 2016.

[4] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. *CoRR*, abs/1609.07093, 2016.

[5] X. Chen, X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. pages 2172–2180, 2016.

[6] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, pages 2960–2967, Washington, DC, USA, 2013. IEEE Computer Society.

[7] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4476–4484, 2017.

[8] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, Jan. 2016.

[9] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, June 2016.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. pages 2672–2680, 2014.

[12] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. D. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. June 2016.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks.

[15] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.

[16] L. Karacan, Z. Akata, A. Erdem, and E. Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *CoRR*, abs/1612.00215, 2016.

[17] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. 2017.

[18] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.

[19] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. 2017.

[20] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.*, 36(4):120:1–120:15, 2017.

[21] M. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. 2017.

[22] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. 2016.

[23] H. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. pages 343–347, 2014.

[24] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[25] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[26] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1060–1069. JMLR.org, 2016.

[27] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 217–225. Curran Associates, Inc., 2016.

[28] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. Improved techniques for training gans. pages 2234–2242, 2016.

[29] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016.

[30] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays. Scribbler: Controlling deep image synthesis with sketch and color. *CoRR*, abs/1612.00835, 2016.

[31] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. 2017.

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *international conference on learning representations*, 2015.

[33] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár. Amortised MAP inference for image super-resolution. 2017.

[34] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. 2017.

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

[36] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *CoRR*, abs/1603.03417, 2016.

[37] A. van den Oord, N. Kalchbrenner, L. Espeholt, k. kavukcuoglu, O. Vinyals, and A. Graves. Conditional image generation with pixelcnn decoders. pages 4790–4798, 2016.

[38] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[39] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016.

[40] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. 2016.

[41] J. Yang, A. Kannan, D. Batra, and D. Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. *CoRR*, abs/1703.01560, 2017.

[42] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image anslation. 2017.

[43] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. 2017.

[44] J. J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *CoRR*, abs/1609.03126, 2016.

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