Tag Disentangled Generative Adversarial Networks for Object Image Re-rendering

Chaoyue Wang, Chaohui Wang, Chang Xu, Dacheng Tao

Tag Disentangled Generative Adversarial Networks for Object Image Re-rendering

Chaoyue Wang[†], Chaohui Wang[‡], Chang Xu[⋆], Dacheng Tao[⋆]
[†]Centre for Artificial Intelligence, School of Software, University of Technology Sydney, Australia
[‡]Université Paris-Est, LIGM (UMR 8049), CNRS, ENPC, ESIEE Paris, UPEM, Marne-la-Vallée, France
[⋆]UBTech Sydney AI Institute, School of IT, FEIT, The University of Sydney, Australia
chaoyue.wang@student.uts.edu.au, chaohui.wang@u-pem.fr, {c.xu, dacheng.tao}@sydney.edu.au

Abstract

In this paper, we propose a principled Tag Disentangled Generative Adversarial Networks (TD-GAN) for re-rendering new images for the object of interest from a single image of it by specifying multiple scene properties (such as viewpoint, illumination, expression, *etc*.). The whole framework consists of a disentangling network, a generative network, a tag mapping net, and a discriminative network, which are trained jointly based on a given set of images that are completely/partially tagged (*i.e*., supervised/semi-supervised setting). Given an input image, the disentangling network extracts disentangled and interpretable representations, which are then used to generate images by the generative network. In order to boost the quality of disentangled representations, the tag mapping net is integrated to explore the consistency between the image and its tags. Furthermore, the discriminative network is introduced to implement the adversarial training strategy for generating more realistic images. Experiments on two challenging datasets demonstrate the state-of-the-art performance of the proposed framework in the problem of interest.

1 Introduction

The re-rendering of new images for the object of interest from a single image of it by specifying expected scene properties (such as viewpoint, illumination, expression, *etc*.) is of fundamental interest in computer vision and graphics. For example, re-rendering of faces for the continuous pose, illumination directions, and various expressions would be an essential component for virtual reality systems, where one tends to naturally "paste" persons into a virtual environment. More applications of image re-rendering can be found in architecture, simulators, video games, movies, visual effects, *etc*.

Conventional approaches to addressing the object image re-rendering problem are generally based on the following scheme: a 3D (static or dynamic) model of the object of interest is first reconstructed from the given image(s) and is then projected onto the 2D image plane corresponding to a specified configuration of scene properties. However, 3D model reconstruction from a single 2D object image is a highly ill-posed and challenging problem. Taking the 3D surface reconstruction of the static object for an example, it relies on the exploitation of class-specific statistical priors on the 3D representation of the object of interest [Kar *et al.*, 2015; Vicente *et al.*, 2014; Wang *et al.*, 2011], while the modeling of such 3D priors necessitates very high expenses in building the training data. Also considering the current state-of-the-art in the study of 3D reconstruction, it is desirable to directly integrate the 3D model reconstruction and 3D-2D projection process together so as to be able to focus on the 2D data only.

Recent works (*e.g*., [Cheung *et al.*, 2014; Kulkarni *et al.*, 2015]) have shown the promise of deep learning models in achieving the goal of interest. The basic idea is based on learning interpretable and disentangled representations [Bengio *et al.*, 2013] of images. Unlike most deep learning models which focus on the learning of hierarchical representations [Bengio *et al.*, 2015], these models aim to extract disentangled representations which correspond to different factors (*e.g*., identity, viewpoint, *etc*.) from the input image. Learning the disentangled representations aims to express the objective factors with different high-level representations. Using the human brain to make an analogy, the human understands the world by projecting real-world objects into abstract concepts of different factors and can generate new object images with these concepts via generalization.

Despite the great progress achieved in image re-rendering, existing methods share some important limitations. The first one is regarding the effectiveness and independence of the (disentangled) representations extracted from the input image. For most existing methods, (disentangled) representations are mostly extracted from images themselves, and the valuable tag information (*e.g*., photograph conditions and object characterizations) associated with images has not been finely explored. Besides, there was little attempt to make the re-rendering result more realistic by increasing the difficulty in distinguishing genuine and re-rendered images. Last but not least, previous works mostly focused on performing object image re-rendering w.r.t. a single scene property and the extension of the developed methods to the setting of multiple scene properties is not straightforward.

In order to boost the performance in re-rendering new images for the object of interest from a single image of it by specifying multiple scene properties, in this paper, we propose a principled Tag Disentangled Generative Ad-

versarial Networks (*TD-GAN*). The whole framework consists of four parts: a disentangling network, a generative network, a tag mapping net, and a discriminative network, which are trained jointly based on a given set of images that are completely/partially tagged (corresponding to the supervised/semi-supervised setting). Given an input image, the disentangling network extracts disentangled and interpretable representations, which are then used to generate images by the generative network. In order to boost the quality of the obtained disentangled representations, the tag mapping net is integrated to explore the consistency between the image and its tags. Considering that the image and its tags record the same object from two different perspectives [Xu *et al.*, 2015], they should share the same disentangled representations. Furthermore, the discriminative network is introduced to implement the adversarial training strategy for generating more realistic images. Experiments on two challenging datasets demonstrate the state-of-the-art performance of our framework in the problem of interest.

2 Related Works

The last decade has witnessed the emergence of algorithms for image modeling and rendering (*e.g.*, [Kingma and Welling, 2013; Dosovitskiy *et al.*, 2015; Liu *et al.*, 2017]). In particular, a large effort has been devoted to the new view synthesis problem and quality results can be obtained for real-world objects, even provided with only one single object image. To name a few, by integrating auto-encoder with recurrent neural networks, [Yang *et al.*, 2015] proposed the recurrent convolutional encoder-decoder network (*RCEDN*) to synthesize novel views of 3D objects through rotating the input image with a fixed angle. [Tatarchenko *et al.*, 2016] also adopted convolutional network to generate novel object images using a single object image as input.

Besides the viewpoint, other factors, including illumination and scale, have also been studied in re-rendering object images. *Transforming auto-encoders* [Hinton *et al.*, 2011] is among the earliest works that attempted to learn a whole vector of instantiation parameters, which is then used to generate a transformed version of the input image, through training auto-encoder capsules. [Cheung *et al.*, 2014] introduced an unsupervised cross-covariance penalty (*XCov*) for learning disentangled representations of input images through the hidden layers of deep networks. Novel images can then be generated by resetting these disentangled representations according to specified configurations. [Kulkarni *et al.*, 2015] proposed the deep convolutional inverse graphics network (*DC-IGN*), which employs convolutional layers to de-render the input images and then re-render the output images using deconvolutional layers. Neurons of the *graphics codes* layer between convolutional and de-convolutional layers are encouraged to represent disentangled factors (identity, illumination, *etc*.) of the input image. Although the *XCov* and *DC-IGN* methods utilize label (tag) information to guide their models to learn disentangled representations from the input image, they do not consider the consistency between the image and its tags, *i.e*., the fact that the image and its tags share the same latent (disentangled) representation. In our framework, besides learning disentangled representations from images, we dig up the correlation between tags and disentangled representations, which enables our proposed *TD-GAN* to achieve better performance in learning disentangled representations.

Moreover, image re-rendering is related to image generation, which aims to generate images using scene-level features. [Dosovitskiy *et al.*, 2015] proposed an 'up-convolutional' network to generate images of chairs given the label of chair style and sine (cosine) value of azimuth angle. Experimentally, the 'up-convolutional' network has exhibited strong ability to generate images with specified features. However, it can only generate images for the objects existing in the training set and simply interpolate between them. Similarly, generative adversarial nets (*GANs*) [Goodfellow *et al.*, 2014] and *GANs*-based models [Gauthier, 2014; Chen *et al.*, 2016] are trained to generate images from different variables, such as, unstructured noise vector [Radford *et al.*, 2015], text [Reed *et al.*, 2016], latent code [Chen *et al.*, 2016], *etc*. Based on the adversarial training strategy, the generative network in *GANs* maps input variables into image space through playing a minimax optimization with the discriminative network.

3 The Proposed Approach

A set $\mathcal{X}^L$ of tagged images is considered in the supervised setting. Let $\{(\mathbf{x}_1, \mathbf{C}_1), \cdots, (\mathbf{x}_{|\mathcal{X}^L|}, \mathbf{C}_{|\mathcal{X}^L|})\}$ denote the whole training dataset, where $\mathbf{x}_i \in \mathcal{X}^L$ denotes the i^{th} image and $\mathbf{C}_i \in \mathcal{C}$ its corresponding *tag codes* which can be represented as: $\mathbf{C}_i = (\mathbf{c}_i^{\text{ide}}, \mathbf{c}_i^{\text{view}}, \mathbf{c}_i^{\text{exp}}, \dots)$. One-hot encoding vectors are employed to describe tag information (*e.g*., in the case of expression tag with three candidates [neutral, smile, disgust], $\mathbf{c}_i^{\text{exp}} = [0, 1, 0]$ for an image with "smile" tag). An additional untagged training image set $\mathcal{X}^U$ is considered in the semi-supervised setting.

3.1 Main Framework

As shown in Fig. 1, *TD-GAN* is composed of four parts: a tag mapping net g, a generative network G, a disentangling network R and a discriminative network D. The tag mapping net g and the disentangling network R aim to map *tag codes* and real-world images into latent disentangled representations, which can then be decoded by the generative network G. Finally, the discriminative network D is introduced to perform the adversarial training strategy which plays a minimax game with the networks R and G. There are three interrelated objectives in the framework of *TD-GAN* for realizing the optimal image re-rendering task.

Exploring consistency between the image and its tags.
In practice, the image provides a visual approach to recording the real-world object. Meanwhile, tags (identity, viewpoint, illumination, *etc*.) accompanied with the image can describe object in a textual or parametric way. The image and its tags consistently represent the same object, despite the difference in physical properties. It is therefore meaningful to explore the consistency between the image and its tags in the image re-rendering task. In the training procedure, the disentangling network R aims to extract the disentangled representations

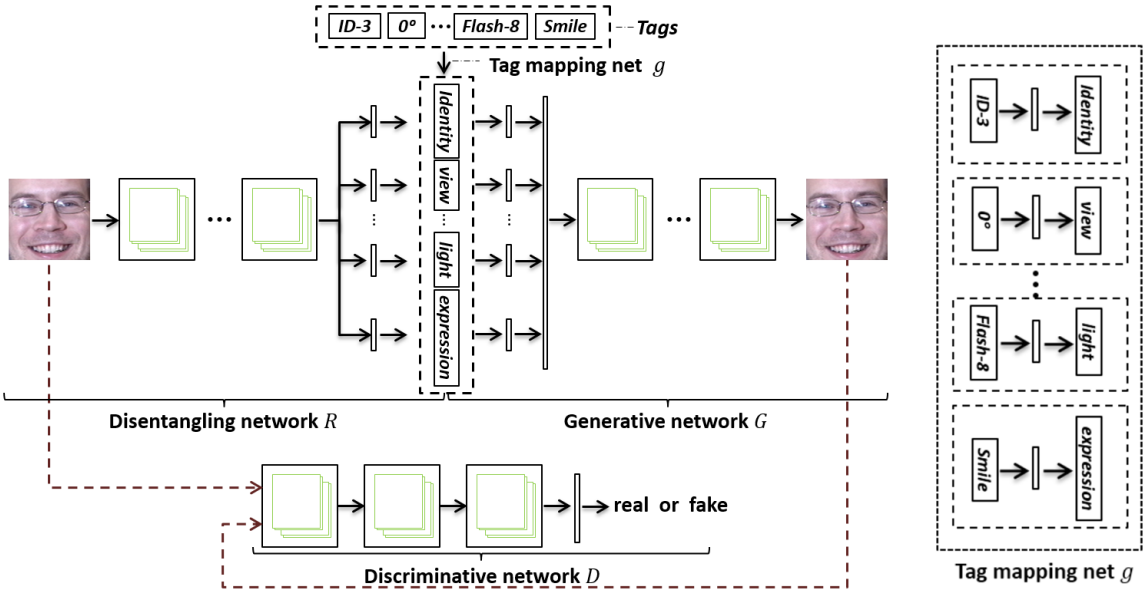

Figure 1: Model architecture. *TD-GAN* is composed of four parts: a tag mapping net g, a disentangling network R, a generative network G and a discriminative network D. (a) During training, the tag mapping net g and the generative network G are trained to render images with their tags. The disentangling network R aims to extract disentangled representations, which can be decoded by the network G. The discriminative network D plays a minimax game with networks G and R based on the adversarial training strategy. (b) During test, the disentangling network R extracts disentangled representations from the input image. After replacing one or multiple disentangled representations with the specified representations generated by the tag mapping net g, the image can be re-rendered through the generative network G.

$R(\mathbf{x})$ of the input image $\mathbf{x}$[1]. Besides, *tag codes* $\mathbf{C}$ are fed through the tag mapping net g to obtain the disentangled representations[2] $g(\mathbf{C})$ of the tags. The first objective of *TD-GAN* is to penalize the discrepancy between the disentangled representations $R(\mathbf{x})$ and $g(\mathbf{C})$ generated from the image and its tags. Using the *L2* norm to penalize such a discrepancy, the energy function is defined as follows:

$$f_1(R, g) = \frac{1}{|\mathcal{X}^L|} \sum_{\mathbf{x}_i \in \mathcal{X}^L} \| R(\mathbf{x}_i) - g(\mathbf{C}_i) \|_2^2 \qquad (1)$$

Compared with valuable tag information, lots of untagged images can be easily harvested in practice. In order to explore useful information contained in these untagged images for network learning in the semi-supervised setting, by applying the generative network G on their disentangled representations $R(\mathbf{x})$, we utilize the following objective function so as to encourage the reconstructed image from the disentangled representations $R(\mathbf{x}_i)$ to be close to the genuine image $\mathbf{x}_i$:

$$\tilde{f}_1(G, R) = \frac{1}{|\mathcal{X}^U|} \sum_{\mathbf{x}_i \in \mathcal{X}^U} \| G(R(\mathbf{x}_i)) - \mathbf{x}_i \|_2^2 \qquad (2)$$

Maximizing image rendering capability.
Given image *tag codes* $\mathbf{C}$, the generative network G should have the capability to render the image with its disentangled representations $g(\mathbf{C})$. Note that the disentangled representations extracted from the image and its tags have already been encouraged to be the same (see Eq. (1)). To maximize the rendering capability of network G, we tend to minimize the discrepancy between genuine images and rendered images via the following objective function:

$$f_2(G, g) = \frac{1}{|\mathcal{X}^L|} \sum_{\mathbf{x}_i \in \mathcal{X}^L} \| G(g(\mathbf{C}_i)) - \mathbf{x}_i \|_2^2 \qquad (3)$$

We can re-render images based on the networks R, G and g discussed above. However, such a way lacks strong driving force for continuously improving the re-rendering performance. Therefore, we introduce the discriminative network D into *TD-GAN* so as to minimize genuine image recognition loss, the detail of which is shown below.

Minimizing genuine image recognition loss.
Ideally, image re-rendering should be able to mix re-rendered images with genuine images so that they cannot be distinguished. To this end, we adopt the adversarial training strategy, based on the *GANs* model [Goodfellow *et al.*, 2014]. Adversarial training suggests the use of the discriminative network D as an adversary of the networks R and G. The discriminative network D outputs the probability that the input image is genuine, and tries its best to detect all those re-rendered images. Competition in this game encourages R, G and D to improve their solutions until the re-rendered images are indistinguishable from those genuine ones. Formally, the objective function can be formulated as:

$$f_3(R, G, D) = \mathbb{E}[\log D(\mathbf{x})] + \mathbb{E}[\log(1 - D(G(R(\mathbf{x}))))] \qquad (4)$$

where $\mathbb{E}$ is computed over those genuine images in the training set. Higher values of f_3 indicates better discriminative abilities, and vice versa.

3.2 Training Process

TD-GAN is composed of R, D, G and g, which can be optimized and learned using the alternating optimization strategy.

[1]Formally, $R(\mathbf{x}) = (r_{\text{ide}}, r_{\text{view}}, \dots)$, where those r_* denote disentangled representations extracted from the input image.

[2]The function g actually consists of a set of independent sub-functions designed to translate the tags into the corresponding representations, *i.e.*, $g(\mathbf{C}) = (g_{\text{ide}}(\mathbf{c}^{\text{ide}}), g_{\text{view}}(\mathbf{c}^{\text{view}}), \dots)$.

Disentangling Net R	**Discriminative Net D**
Input: Image $\mathbf{x}$ – $(128 \times 128 \times 3)$	
[Layer 1] Conv. (3, 3, 64) Stride=2. *LReLU* BatchNorm	
[Layer 2] Conv. (3, 3, 128) Stride=2. *LReLU* BatchNorm	
[Layer 3] Conv. (3, 3, 256) Stride=2. *LReLU* BatchNorm	
[Layer 4] Conv. (3, 3, 256) Stride=1. *LReLU* BatchNorm	
[Layer 5] Conv. (3, 3, 256) Stride=1. *LReLU* BatchNorm	
[Layer 6] Conv. (3, 3, 512) Stride=2. *LReLU* BatchNorm	
[Layer 7] Conv. (3, 3, 512) Stride=1. *ReLU* BatchNorm	
[Layer 8] FC. 5120 *LReLU*	FC. 2560 *LReLU*
[Layer 9] FC. separate *Tanh*	FC. 1 *Sigmoid*
Output: Disentagnled Representations $R(\mathbf{x})$	Probability of being genuine $D(\mathbf{x})$

Generative Network G
Input: Disentangled representations $R(\mathbf{x})$
[Layer 1] Concatenate Layer
[Layer 2] FC. (4, 4, 1024) *ReLU* BatchNorm
[Layer 3] Deconv. (4, 4, 512) Stride=2 *ReLU* BatchNorm
[Layer 4] Deconv. (4, 4, 256) Stride=2 *ReLU* BatchNorm
[Layer 5] Deconv. (4, 4, 128) Stride=2 *ReLU* BatchNorm
[Layer 6] Deconv. (4, 4, 64) Stride=2 *ReLU* BatchNorm
[Layer 7] Deconv. (4, 4, 3) Stride=2 *Tanh*.
Output: Generated Image $G(R(\mathbf{x}))$ – $(128 \times 128 \times 3)$

Table 1: Details of the disentangling, discriminative and generative networks used for all the experiments.

In the following, we first present the optimization with respect to each component of *TD-GAN* at each iteration in the context of the supervised setting, and then clarify the difference in the optimization for the semi-supervised setting.

We optimize the tag mapping net g, by fixing G^*, R^* and D^*. Since g maps tags to disentangled representations, optimizing g involves the first and second objectives (f_1 and f_2):

$$\mathcal{L}_g = \min_g \ \lambda_1 f_1(R^*, g) + \lambda_2 f_2(G^*, g) \tag{5}$$

Here, those λ's are hyper-parameters balancing the influence of the corresponding terms, and R^* and G^* are fixed as the configurations obtained from the previous iteration (similarly for the other optimization steps presented below).

By fixing g^*, D^* and R^*, the search of the generative network G can be formulated as:

$$\mathcal{L}_G = \min_G \ \lambda_2 f_2(G, g^*) + \lambda_3 f_3(R^*, G, D^*) \tag{6}$$

where G determines the re-rendering procedure with disentangled representations as input.

The disentangling network R is trained by fixing g^*, G^* and D^*. Similarly, the main target of the network R is to infer disentangled representations from the input image (*i.e.*, f_1), and it is also a part of adversarial training (*i.e.*, f_3). Hence, the loss function with respect to R should be:

$$\mathcal{L}_R = \min_R \ \lambda_1 f_1(R, g^*) + \lambda_3 f_3(R, G^*, D^*) \tag{7}$$

The discriminative network D is introduced to cooperate with the adversarial training strategy, via the following optimization:

$$\mathcal{L}_D = \max_D \ \lambda_3 f_3(R^*, G^*, D) \tag{8}$$

We can observe that the above processing forms a minimax game, which aims to maximize the image re-rendering capability and to minimize the error in distinguishing between genuine and re-rendered images. In the semi-supervised setting, since G and R are used to re-render the untagged images as well, the objective functions in Eqs. (6) and (7) also include a weighted loss $\lambda_4 \tilde{f}_1$ for untagged image re-rendering.

	Tag	Hidden layer	Disentangled
$\text{Chair}^{\text{identity}}_{\text{fully}}$	500	FC.1024 *ReLU*	FC.1024 *Tanh*
$\text{Chair}^{\text{identity}}_{\text{semi}}$	100	FC.1024 *ReLU*	FC.1024 *Tanh*
$\text{Chair}^{\text{viewpoint}}$	31	FC. 512 *ReLU*	FC. 512 *Tanh*
$\text{Face}^{\text{identity}}$	200	FC.1024 *ReLU*	FC.1024 *Tanh*
$\text{Face}^{\text{illumination}}$	19	FC. 512 *ReLU*	FC. 512 *Tanh*
$\text{Face}^{\text{viewpoint}}$	13	FC. 256 *ReLU*	FC. 256 *Tanh*
$\text{Face}^{\text{expression}}$	3	FC. 256 *ReLU*	FC. 256 *Tanh*

Table 2: Experimental settings of the tag mapping net g.

3.3 Image Re-rendering

In our framework, the trained disentangling network R is able to transform the input image into disentangled representations corresponding to those scene properties (*e.g.*, identity, viewpoint, *etc.*). Hence, in the test stage, given an unseen image $\mathbf{x}_{\text{test}}$ of the object of interest as input, the disentangling network R will output its disentangled representations. The image re-rendering task is performed simply by replacing one or multiple disentangled representations with the specified representation(s) generated by the tag mapping net g. The obtained disentangled representations are then fed to the generative network G so as to output the re-rendering result.

4 Experiments

We evaluated the performance of our *TD-GAN* method based on two challenging datasets: *3D-chairs* dataset [Aubry *et al.*, 2014] and *Multi-PIE* database [Gross *et al.*, 2010]. The obtained results demonstrate the advantage of our method in learning interpretable disentangled representations and re-rendering images of unseen objects with tag configurations.

4.1 Implementation Details

In our experiments, all images were resized to $128 \times 128 \times 3$. Meanwhile, all layers, except those fully-connected with disentangled representations were fixed and the details of these layers, are shown in Table 1. Distinct *tag codes* (*i.e.*, different architecture configurations of the tag mapping net g) were chosen for different tasks and we summarize in Table 2 all the settings used in our experiments.

Moreover, our proposed *TD-GAN* was implemented based on *Theano* [Bergstra *et al.*, 2010]. All the parameters were initialized using a zero-mean Gaussian with variance 0.05. The same weights ($\lambda_1 = 10, \lambda_2 = 10, \lambda_3 = 1$, and $\lambda_4 = 50$) were used in all the experiments. Based on the training process described above, we alternatively updated the disentangling network, the discriminative network and the generative network (with the tag mapping net). The optimization was done based on the *ADAM* solver [Kingma and Ba, 2014] with momentum 0.5, where the learning rate was set as 0.0005 for the disentangling network and 0.0002 for all the other networks. We utilized a minibatch size of 50 and trained for around 900 epochs.

4.2 Performance Criteria

We qualitatively and quantitatively evaluated the performance of our method. On the one hand, following existing

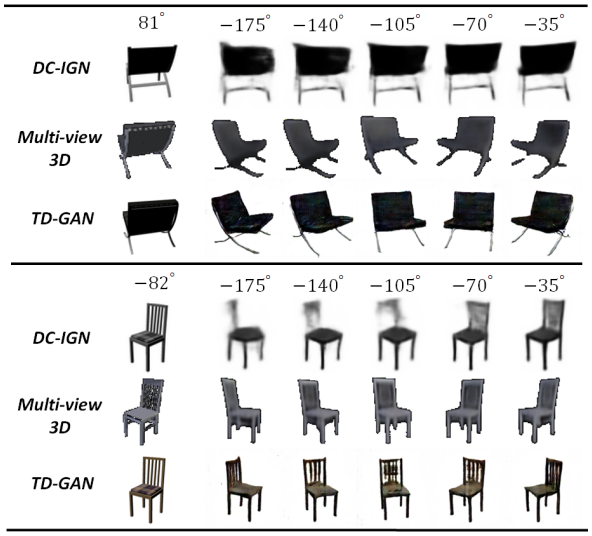

Figure 2: Novel view synthesis results of two previous method and ours. For each method, the leftmost image is the input image, and the images on its right side were re-rendered under different viewpoints.

works [Tatarchenko *et al.*, 2016; Cheung *et al.*, 2014], we choose to use the qualitative criteria to evaluate the performance of re-rendered images, *i.e.*, evaluating through human observers. To this end, we directly report the input image and its re-rendered images for a set of representative test examples. On the other hand, as for quantitative criteria, same as previous work [Kulkarni *et al.*, 2015], we measured the mean squared error (*MSE*) between the re-rendered images and the corresponding ground truths over the test set.

4.3 Experimental Results

3D-chairs dataset.

The *3D-chairs* dataset [Aubry *et al.*, 2014] contains 86,366 images rendered from 1,393 3D CAD models of different chairs. For each chair, 62 viewpoints are taken from 31 azimuth angles (with a step of 11 or 12 degrees) and 2 elevation angles (20 and 30 degrees). Since the *3D-chairs* dataset records chair images from different viewpoints, we firstly performed the novel view synthesis task, which takes a single image of an unseen object as input, and attempts to re-render images under novel viewpoints. Following the experimental setting of existing works [Tatarchenko *et al.*, 2016; Yang *et al.*, 2015; Dosovitskiy *et al.*, 2015], we selected 809 chair models from all 1,393 chair models by removing near-duplicate (*e.g.*, those differing only in color) and low-quality models. The first 500 models were used for training purpose, while the remaining 309 models for test.

In Fig. 2, we show the novel view synthesis results of *TD-GAN* and two previous methods on the *3D-chairs* dataset. Following the comparison in [Tatarchenko *et al.*, 2016], we adopt the results reported by existing works [Kulkarni *et al.*, 2015; Tatarchenko *et al.*, 2016] as references, and exhibit our results over the same (or very similar) test images. Among them, *Multi-View 3D* [Tatarchenko *et al.*, 2016] was designed to re-render unseen viewpoints of 3D objects. Compared with *Multi-View 3D*, our method not only achieves comparable performance in viewpoint transformation, but also be able to

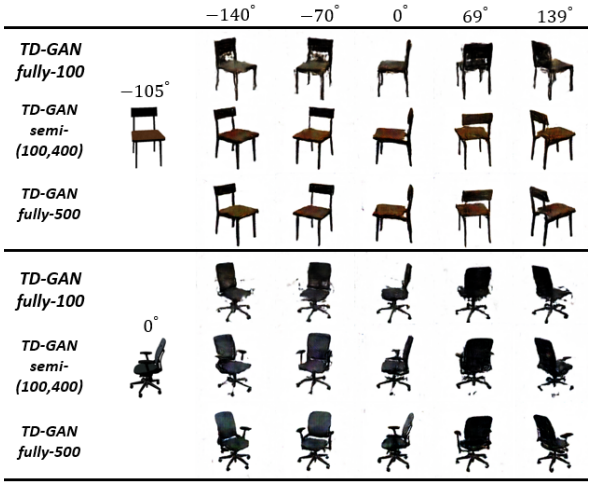

Figure 3: Novel view synthesis results of *TD-GAN* trained in three settings. The images are arranged similarly to Fig. 2.

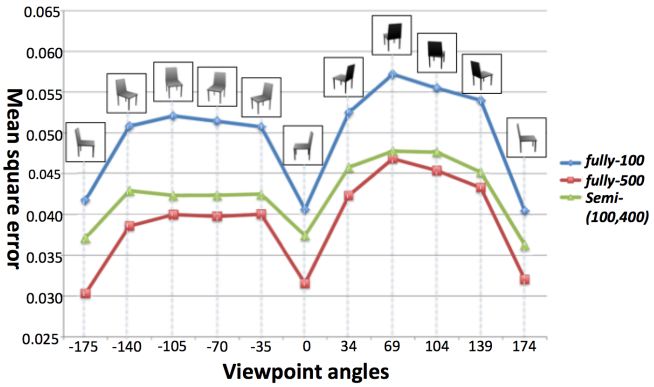

Figure 4: *MSE* of *TD-GAN* trained in three settings. *MSE* was calculated over all the chair images under 0° in the test set as inputs. Chair images are used to indicate target viewpoints.

carry out other image re-rendering tasks, such as illumination transformation, expression transformation, *etc*. Despite the fact that both *DC-IGN* [Kulkarni *et al.*, 2015] and *TD-GAN* were developed to re-render images with multiple different configurations, the obtained experimental results demonstrate that *TD-GAN* achieves much better results. Furthermore, we evaluated the semi-supervised extension of our *TD-GAN* method. To this end, within the aforementioned experimental setting, we tested the following three representative training settings: using all the 500 training models and their tags, only the first 100 models and their tags, and all the 500 training models but only the tags of the first 100 models (referred to as *fully-500*, *fully-100*, and *semi-(100,400)*, respectively). We show in Fig. 3 some representative qualitative results on the same test images, from which we can observe that: (i) the fully-supervised setting with all the 500 models (*i.e.*, *fully-500*) achieves the best performance; (ii) the semi-supervised setting (*i.e.*, *semi-(100,400)*) shows slightly degraded performance compared to *fully-500*; and (iii) both of *fully-500* and *semi-(100,400)* perform much better than the fully-supervised setting with only the first 100 models (*i.e.*, *fully-100*). For quantitative results, we selected all the images of chair objects under 0° in the test set as inputs, and re-rendered the images under different viewpoints. We report the *MSE* val-

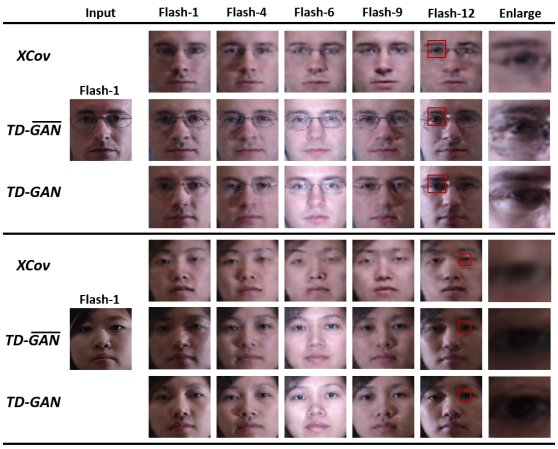

Figure 5: Illumination transformation of human face (best view in color). Given an image (leftmost) of human face, we reported a set of re-rendered images on its right side.

	XCov	*TD-GAN*	*TD-GAN*
Flash-1	0.5776	0.5623	0.5280
Flash-4	5.0692	0.8972	0.8818
Flash-6	4.3991	1.2509	0.1079
Flash-9	3.4639	0.6145	0.5870
Flash-12	2.4624	0.7142	0.6973
All Flash (mean)	3.8675	0.6966	0.6667

Table 3: *MSE* ($\times 10^{-2}$) of illumination transformation.

ues for a set of individual target viewpoints in Fig. 4, and the global *MSE* over all the target viewpoints is 0.03875, 0.04379 and 0.05018 in the *fully-500*, *semi-(100,400)* and *fully-100* settings, respectively. Finally, from both the qualitative and quantitative results, we can conclude that the introduction of untagged data (semi-supervised learning) is indeed beneficial for improving the performance of the *TD-GAN* method.

Multi-PIE database.

Multi-PIE [Gross *et al.*, 2010] is a face dataset containing images of 337 people under 15 camera-poses (13 camera-poses with 15° intervals at head height and 2 additional ones located above the subject) and 19 illumination conditions in up to four sessions. Following previous works [Cheung *et al.*, 2014; Ding *et al.*, 2015], we cropped all face images based on manually annotated landmarks on eyes, nose and mouth. Since *Multi-PIE* records several factors (tags) of human faces, it is suitable for training *TD-GAN* to perform various image re-rendering tasks.

We evaluated the performance of *TD-GAN* in re-rendering face images under various illuminations. By fixing viewpoint (0°) and expression (neutral), we selected all 337 identities in all four sessions under different illuminations as the data setting. Among the 337 identities, the first 200 ones were used for training, and the remaining 137 ones for test. Besides our *TD-GAN*, we also evaluated *XCov* [Cheung *et al.*, 2014] and *TD-GAN* without the use of generative adversarial loss (*i.e.*, f_3 in Eq. (4)) in the same task (refer to as *TD-GAN*). In Fig. 5, we show some images re-rendered by those methods using the same images as inputs. Table 3 reports *MSE* for the images re-rendered with five individual target illuminations and with all the target illuminations, using the images of all human faces under a fixed illumination (*Flash-1*) in the test

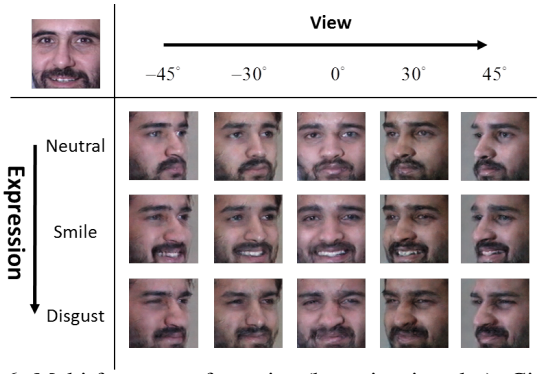

Figure 6: Multi-factor transformation (best view in color). Given a single image as input (the up-left one), its viewpoint and expression were jointly transformed.

set as inputs. According to Fig. 5 and Table 3, we can draw the following two conclusions. Firstly, *XCov* performs much worse in re-rendering images when the specified illumination condition is different from that of the input image, while our method performs well in all those re-rendering with illumination transformation experiments. Indeed, ideally, the re-rendering performance should be independent with the scene properties (*e.g.*, illumination) of the input image, unless the extracted identity representation is correlated to those properties. Secondly, the introduction of generative adversarial loss really helps our method to generate more realistic images with fine details.

Last but not least, in order to validate the capability of re-rendering images with multi-factor transformation, we performed an experiment where the viewpoint and expression were jointly configured. We used the data of session 3 as the dataset, which contains 230 identities with three expressions (neutral, smile and disgust). Similarly, the first 200 identities were used as the training set, and the remaining 30 ones served as the test set. As shown in Fig. 6, given a test image, our method is able to effectively re-render images of the same identity with different expressions and viewpoints. The re-rendered images look quite natural and make us believe that they are indeed the genuine images of the same face exhibiting in the input image.

5 Conclusion

In this paper, we have investigated the image re-rendering problem by developing a principled Tag Disentangled Generative Adversarial Networks (*TD-GAN*). Images and their associated tags have been fully and finely explored to discover the disentangled representations of real-world objects. The whole framework is established with three interrelated objectives, and can be effectively optimized with respect to the involved four essential parts. Experimental results on real-world datasets demonstrate that the proposed *TD-GAN* framework is effective and promising for practical image re-rendering applications.

Acknowledgments

This work is supported by the Australian Research Council Projects FT-130101457, DP-140102164, LP-150100671, and CNRS INS2I-JCJC-INVISANA.

References

[Aubry *et al.*, 2014] Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3769, 2014.

[Bengio *et al.*, 2013] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[Bengio *et al.*, 2015] Yoshua Bengio, Ian J Goodfellow, and Aaron Courville. Deep learning. *An MIT Press book in preparation. Draft chapters available at http://www. iro. umontreal. ca/ bengioy/dlbook*, 2015.

[Bergstra *et al.*, 2010] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: A cpu and gpu math compiler in python. In *Proc. 9th Python in Science Conf*, pages 1–7, 2010.

[Chen *et al.*, 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

[Cheung *et al.*, 2014] Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

[Ding *et al.*, 2015] Changxing Ding, Chang Xu, and Dacheng Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 24(3):980–993, 2015.

[Dosovitskiy *et al.*, 2015] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1546, 2015.

[Gauthier, 2014] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014, 2014.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[Gross *et al.*, 2010] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[Hinton *et al.*, 2011] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer, 2011.

[Kar *et al.*, 2015] Abhishek Kar, Shubham Tulsiani, Joo Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1966–1974. IEEE, 2015.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kulkarni *et al.*, 2015] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547, 2015.

[Liu *et al.*, 2017] Tongliang Liu, Dacheng Tao, Mingli Song, and Stephen J. Maybank. Algorithm-dependent generalization bounds for multi-task learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):227–241, February 2017.

[Radford *et al.*, 2015] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 3, 2016.

[Tatarchenko *et al.*, 2016] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision*, pages 322–337. Springer, 2016.

[Vicente *et al.*, 2014] Sara Vicente, João Carreira, Lourdes Agapito, and Jorge Batista. Reconstructing pascal voc. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE, 2014.

[Wang *et al.*, 2011] Chaohui Wang, Yun Zeng, Loic Simon, Ioannis Kakadiaris, Dimitris Samaras, and Nikos Paragios. Viewpoint invariant 3d landmark model inference from monocular 2d images using higher-order priors. In *2011 International Conference on Computer Vision*, pages 319–326. IEEE, 2011.

[Xu *et al.*, 2015] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view intact space learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2531–2544, 2015.

[Yang *et al.*, 2015] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pages 1099–1107, 2015.