Learning to Discover Cross-Domain Relations
with Generative Adversarial Networks

Taeksoo Kim [1] **Moonsu Cha** [1] **Hyunsoo Kim** [1] **Jung Kwon Lee** [1] **Jiwon Kim** [1]

Abstract

While humans easily recognize relations between data from different domains without any supervision, learning to automatically discover them is in general very challenging and needs many ground-truth pairs that illustrate the relations. To avoid costly pairing, we address the task of discovering cross-domain relations when given unpaired data. We propose a method based on generative adversarial networks that learns to discover relations between different domains (DiscoGAN). Using the discovered relations, our proposed network successfully transfers style from one domain to another while preserving key attributes such as orientation and face identity.

1. Introduction

Relations between two different domains, the way in which concepts, objects, or people are connected, arise ubiquitously. Cross-domain relations are often natural to humans. For example, we recognize the relationship between an English sentence and its translated sentence in French. We also choose a suit jacket with pants or shoes in the same style to wear.

Can machines also achieve a similar ability to relate two different image domains? This question can be reformulated as a conditional image generation problem. In other words, finding a mapping function from one domain to the other can be thought as generating an image in one domain given another image in the other domain. While this problem tackled by generative adversarial networks (GAN) (Isola et al., 2016) has gained a huge attention recently, most of today's training approaches use explicitly paired data, provided by human or another algorithm.

This problem also brings an interesting challenge from a

[1]SK T-Brain, Seoul, South Korea. Correspondence to: Taeksoo Kim <jazzsaxmafia@sktbrain.com>.
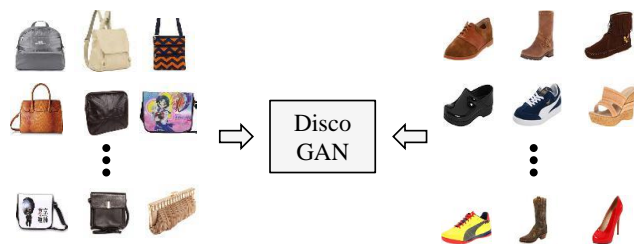
(a) Learning cross-domain relations **without any extra label**

(b) Handbag images (input) & **Generated** shoe images (output)

(c) Shoe images (input) & **Generated** handbag images (output)

Figure 1. Our GAN-based model trains with two independently collected sets of images and learns how to map two domains *without any extra label*. In this paper, we reduce this problem into generating a new image of one domain given an image from the other domain. (a) shows a high-level overview of the training procedure of our model with two independent sets (e.g. handbag images and shoe images). (b) and (c) show results of our method. Our method takes a handbag (or shoe) image as an input, and generates its corresponding shoe (or handbag) image. Again, it is worth noting that our method does not take any extra annotated supervision and can self-discover relations between domains.

learning point of view. Explicitly supervised data is seldom available and labeling can be labor intensive. Moreover, pairing images can become tricky if corresponding images are missing in one domain or there are multiple best candidates. Hence, we push one step further by discovering relations between two visual domains without any explic-
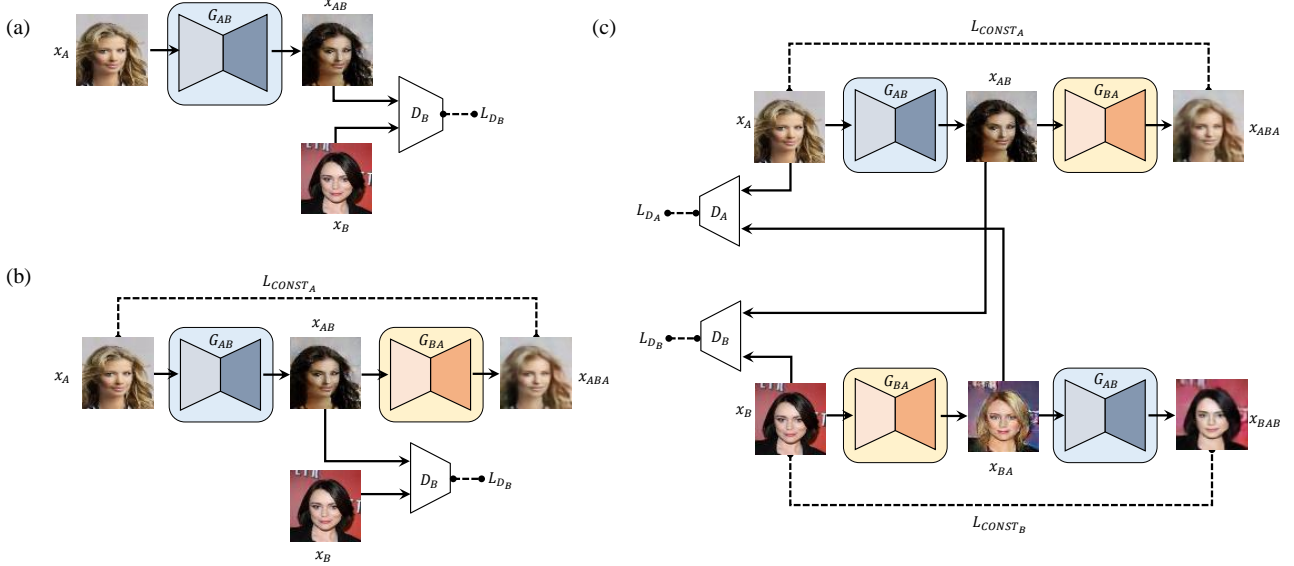
Figure 2. Three investigated models. (a) standard GAN (Goodfellow et al., 2014), (b) GAN with a reconstruction loss, (c) our **proposed model (DiscoGAN)** designed to discover relations between two unpaired, unlabeled datasets. Details are described in Section 3.

itly paired data.

In order to tackle this challenge, we introduce a model that discovers cross-domain relations with GANs (DiscoGAN). Unlike previous methods, our model can be trained with two sets of images without any explicit pair labels (Figure 1a) and does not require any pre-training. Our proposed model can then take an image in one domain and generate its corresponding image in another domain (see Figure 1b). The core of our model is based on two different GANs coupled together – each of them ensures our generative functions can map each domain to its counterpart domain. A key intuition we rely on is to constrain *all images* in one domain to be *representable* by images in the other domain. For example, when learning to generate shoe image from handbag image, we force this generated image to be an image-based representation of the handbag image (and hence reconstruct the handbag image) through a reconstruction loss, and to be as close to images in the shoe domain as possible through a GAN loss. We use these two properties to encourage the mapping between two domains to be well covered on both directions (i.e. encouraging one-to-one rather than many-to-one or one-to-many). In the experiments section, we show that this simple intuition discovers common properties and styles of two domains very well.

Both experiments on toy domain and real world image datasets support the claim that our proposed model is well-suited for discovering cross-domain relations. When translating data points between simple 2-dimensional domains and between face image domains, our DiscoGAN model was more robust to the mode collapse problem compared

to two other baseline models of Figures 2a and 2b. It also learns the bidirectional mapping between two image domains, such as faces, cars, chairs, edges and photos, and successfully apply the mapping in image translation. Translated images consistently change specific attributes such as hair color, gender and orientation while maintaining all other components.

2. Model

We now formally define cross-domain relations and present the problem of learning to discover such relations in two different domains. Standard GAN model and a similar variant model with additional components are investigated for their applicability for this task. Limitations of these models are then explained, and we propose a new architecture based on GANs that can be used to discover cross-domain relations.

2.1. Formulation

Relation is mathematically defined as a function $\mathbf{G}_{AB}$ that maps elements from domain A to elements in its codomain B and vice versa for $\mathbf{G}_{BA}$. In fully unsupervised setting, $\mathbf{G}_{AB}$ and $\mathbf{G}_{BA}$ can be arbitrarily defined. To find a meaningful relation, we need to impose a condition on the relation of interest. Here, we constrain relation to be a one-to-one correspondence (bijective mapping). That means $\mathbf{G}_{AB}$ is the inverse mapping of $\mathbf{G}_{BA}$.

The range of function $\mathbf{G}_{AB}$ – the complete set of all possible resulting values $\mathbf{G}_{AB}(x_A)$ for all x_A's in domain A – should be contained in domain B and similarly for
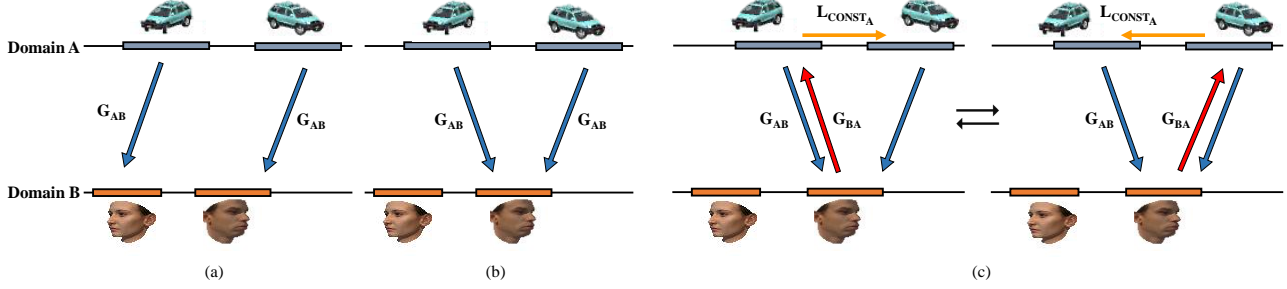
Figure 3. Illustration of our models on simplified one dimensional domains. (a) ideal mapping from domain A to domain B in which the two domain A modes map to two different domain B modes, (b) GAN model failure case, (c) GAN with reconstruction model failure case.

$\mathbf{G}_{BA}(x_B)$.

We now relate these constraints to objective functions. Ideally, the equality $\mathbf{G}_{BA} \circ \mathbf{G}_{AB}(x_A) = x_A$ should be satisfied, but this hard constraint is difficult to optimize and a relaxed soft constraint is more desirable. For this reason, we minimize the distance $d(\mathbf{G}_{BA} \circ \mathbf{G}_{AB}(x_A), x_A)$, where any form of metric function such as L_1, L_2 or Huber loss can be used. Similarly, we also need to minimize $d(\mathbf{G}_{AB} \circ \mathbf{G}_{BA}(x_B), x_B)$.

Guaranteeing that $\mathbf{G}_{AB}$ maps to domain B is also very difficult to optimize. We relax this constraint as follows: we instead minimize generative adversarial loss $-\mathbb{E}_{x_A \sim P_A}\left[\log \mathbf{D}_B(\mathbf{G}_{AB}(x_A))\right]$. Similarly, we minimize $-\mathbb{E}_{x_B \sim P_B}\left[\log \mathbf{D}_A(\mathbf{G}_{BA}(x_B))\right]$.

Now, we explore several GAN architectures to learn with these loss functions.

2.2. Notation and Architecture

We use the following notations in sections below. A generator network is denoted as $\mathbf{G}_{AB} : \mathbb{R}_A^{h \times w \times c} \to \mathbb{R}_B^{h \times w \times c}$, subscripts denote the input and output domains, and superscripts denote the input and output image size. The discriminator network is denoted as $\mathbf{D}_B : \mathbb{R}_B^{h \times w \times c} \to [0, 1]$, and the subscript B denotes that it discriminates images in domain B. Notations $\mathbf{G}_{BA}$ and $\mathbf{D}_A$ are used similarly.

Each generator takes an image of size $h \times w \times c$ and feeds it through an encoder-decoder pair. The encoder part of each generator is composed of convolution layers, each followed by leaky ReLU (Maas et al., 2013; Xu et al., 2015). The decoder part is composed of deconvolution layers, followed by a ReLU, and outputs a target domain image of size $h \times w \times c$. The number of convolution and deconvolution layers ranges from four to five depending on the domain.

The discriminator is similar to the encoder part of the generator. In addition to the convolution layers and leaky Re-

LUs, the discriminator has an additional convolution layer, and a final sigmoid to output a scalar output between $[0, 1]$.

2.3. GAN with a Reconstruction Loss

We first consider a standard GAN model (Goodfellow et al., 2014) for the relation discovery task (see Figure 2a). Originally, a standard GAN takes random Gaussian noise z, encodes it into hidden features h and generates images in domains such as MNIST. We make a slight modification to this model to fit our task: the model we use takes in image as input instead of noise.

In addition, since this architecture only learns one mapping from domain A to domain B, we add a second generator that maps domain B back into domain A (see Figure 2b). We also add a reconstruction loss term that compares the input image with the reconstructed image. With these additional changes, each generator in the model can learn mapping from its input domain to output domain and discover relations between them.

A generator $\mathbf{G}_{AB}$ translates input image x_A from domain A into x_{AB} in domain B. The generated image is then translated into a domain A image x_{ABA} to match the original input image (Equation 1, 2). Various forms of distance functions, such as MSE, cosine distance, and hinge-loss, can be used as the reconstruction loss d (Equation 3). The translated output x_{AB} is then scored by the discriminator which compares it to a real domain B sample x_B.

$$x_{AB} = \mathbf{G}_{AB}(x_A) \tag{1}$$

$$x_{ABA} = \mathbf{G}_{BA}(x_{AB}) = \mathbf{G}_{BA} \circ \mathbf{G}_{AB}(x_A) \tag{2}$$

$$L_{CONST_A} = d(\mathbf{G}_{BA} \circ \mathbf{G}_{AB}(x_A), x_A) \tag{3}$$

$$L_{GAN_B} = -\mathbb{E}_{x_A \sim P_A}\left[\log \mathbf{D}_B(\mathbf{G}_{AB}(x_A))\right] \tag{4}$$

The generator $\mathbf{G}_{AB}$ receives two types of losses – a reconstruction loss L_{CONST_A} (Equation 3) that measures how well the original input is reconstructed after a sequence of

two generations, and a standard GAN generator loss L_{GAN_B} (Equation 4) that measures how realistic the generated image is in domain B. The discriminator receives the standard GAN discriminator loss of Equation 6.

$$L_{G_{AB}} = L_{GAN_B} + L_{CONST_A} \qquad (5)$$

$$L_{D_B} = -\mathbb{E}_{x_B \sim P_B}[\log \mathbf{D}_B(x_B)] \\ -\mathbb{E}_{x_A \sim P_A}[\log(1 - \mathbf{D}_B(\mathbf{G}_{AB}(x_A)))] \quad (6)$$

During training, the generator $\mathbf{G}_{AB}$ learns the mapping from domain A to domain B under two relaxed constraints: that domain A maps to domain B, and that the mapping on domain B is reconstructed to domain A. However, this model lacks a constraint on mapping from B to A, and these two conditions alone does not guarantee a cross-domain relation (as defined in section 2.1) because the mapping satisfying these constraints is one-directional. In other words, the mapping is an injection, not bijection, and one-to-one correspondence is not guaranteed.

Consider two possibly multi-modal image domains A and B. Figure 3 illustrates the two multi-modal data domains on a simplified one-dimensional representation. Figure 3a shows the ideal mapping from input domain A to domain B, where each mode of data is mapped to a separate mode in the target domain. Figure 3b, in contrast, shows the mode collapse problem, a prevalent phenomenon in GANs, where data from multiple modes of a domain map to a single mode of a different domain. For instance, this case is where the mapping $\mathbf{G}_{AB}$ maps images of cars in two different orientations into the same mode of face images.

In some sense, the addition of a reconstruction loss to a standard GAN is an attempt to remedy the mode collapse problem. In Figure 3c, two domain A modes are matched with the same domain B mode, but the domain B mode can only direct to one of the two domain A modes. Although the additional reconstruction loss L_{CONST_A} forces the reconstructed sample to match the original (Figure 3c), this change only leads to a similar symmetric problem. The reconstruction loss leads to an oscillation between the two states and does not resolve mode-collapsing.

2.4. Our Proposed Model: Discovery GAN

Our proposed GAN model for relation discovery – DiscoGAN – couples the previously proposed model (Figure 2c). Each of the two coupled models learns the mapping from one domain to another, and also the reverse mapping for reconstruction. The two models are trained together simultaneously. The two generators $\mathbf{G}_{AB}$'s and the two generators $\mathbf{G}_{BA}$'s share parameters, and the generated images x_{BA} and x_{AB} are each fed into separate discriminators L_{D_A} and L_{D_B}, respectively.

One key difference from the previous model is that input images from both domains are reconstructed and that there are two reconstruction losses: L_{CONST_A} and L_{CONST_B}.

$$L_G = L_{G_{AB}} + L_{G_{BA}} \qquad (7)$$
$$= L_{GAN_B} + L_{CONST_A} + L_{GAN_A} + L_{CONST_B}$$

$$L_D = L_{D_A} + L_{D_B} \qquad (8)$$

As a result of coupling two models, the total generator loss is the sum of GAN loss and reconstruction loss for each partial model (Equation 7). Similarly, the total discriminator loss L_D is a sum of discriminator loss for the two discriminators $\mathbf{D}_A$ and $\mathbf{D}_B$, which discriminate real and fake images of domain A and domain B (Equation 8).

Now, this model is constrained by two L_{GAN} losses and two L_{CONST} losses. Therefore a bijective mapping is achieved, and a one-to-one correspondence, which we defined as cross-domain relation, can be discovered.

3. Experiments

3.1. Toy Experiment

To empirically demonstrate our explanations on the differences between a standard GAN, a GAN with reconstruction loss and our proposed model (DiscoGAN), we designed an illustrative experiment based on synthetic data in 2-dimensional domains A and B. Both source (domain A) and target (domain B) data samples are drawn from Gaussian mixture models.

In Figure 4, the left-most figure shows the initial state of the toy experiment where all the domain A modes map to almost a single point because of initialization of the generator. For all plots the target domain 2D plane is shown with target domain modes marked with black x's. Colored points represent samples from domain A that are mapped to domain B, and each color denotes samples from each domain A mode. In this case, the task is to discover cross-domain relations between domains A and B, and translate samples from five domain A modes into domain B, which has ten modes spread around the arc of a circle.

We use a neural network with three linear layers each followed by a ReLU nonlinearity as the generator. For the discriminator we use five linear layers each followed by a ReLU, except for the last layer which is switched out with a sigmoid that outputs a scalar $\in [0, 1]$. The colored background shows the output value of the discriminator $\mathbf{D}_B$, which discriminates real target domain samples from synthetic translated samples from domain A. The contour lines show regions of same discriminator value.

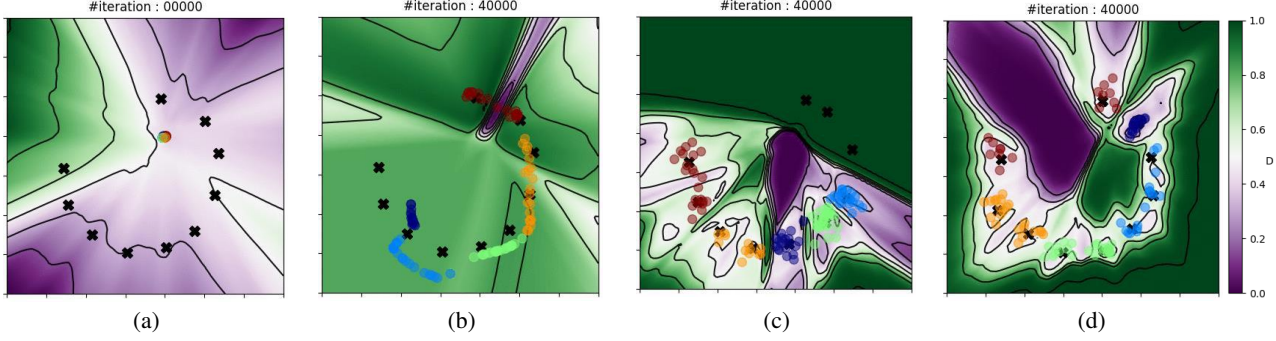50,000 iterations of training were performed, and due to the

Figure 4. Toy domain experiment results. The colored background shows the output value of the discriminator. 'x' marks denote different modes in domain B, and colored circles indicate samples of domain A mapped to domain B, where each color corresponds to a different mode in domain A. (a) ten target (domain B) modes and initial translations, (b) results from standard GAN model, (c) GAN with reconstruction loss, (d) our proposed DiscoGAN model.

domain simplicity our model often converged much earlier. Illustrations of translated samples in Figure 4 show very different behavior depending on the model used.

In the baseline (standard GAN) case, many translated points of different colors are located around the same domain B mode. For example, navy and light-blue colored points are located together, as well as green and orange colored points. This result illustrates the mode-collapse problem of GANs since points of multiple colors (multiple A domain modes) are mapped to the same domain B mode. The baseline model still oscillate around domain B modes throughout the iterations.

In the case of GAN with a reconstruction loss, the collapsing problem is less prevalent, but navy, green and light-blue points still overlap at a few modes. The contour plot also demonstrates the difference from baseline: regions around all domain B modes are leveled in a green colored plateau in the baseline, allowing translated samples to freely move between modes, whereas in the case with reconstruction loss the regions between domain B modes are clearly separated.

In addition, both this model and the standard GAN model fail to cover all modes in domain B since the mapping from domain A to domain B is injective. Our proposed Disco-GAN model, on the other hand, is able to not only prevent mode-collapse by translating into distinct well-bounded regions that do not overlap, but also generate domain B samples in all ten modes as the mappings in our model is bijective. It is noticeable that the discriminator for domain B is perfectly fooled by translated samples from domain A (white regions near B domain modes).

Although this experiment is limited due to its simplicity, the results clearly support the superiority of our proposed model over other variants of GANs.

3.2. Real Domain Experiment

To evaluate whether our DiscoGAN successfully learns underlying relationships between domains, we trained and tested our model using several image-to-image translation tasks that require the use of discovered cross-domain relations between source and target domains.

In each real domain experiment, all input images and translated images were size of $64 \times 64 \times 3$. For training, we employed learning rate of 0.0002 and used the Adam optimizer (Kingma & Ba, 2015) with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We applied Batch Normalization (Ioffe & Szegedy, 2015) to all convolution and deconvolution layers except the first and the last layers, and applied weight decay regularization coefficient of 10^{-4} and minibatch of size 200. All computations were conducted on a single machine with an Nvidia Titan X Pascal GPU and an Intel(R) Xeon(R) E5-1620 CPU.

3.2.1. CAR TO CAR, FACE TO FACE

We used a Car dataset (Fidler et al., 2012) which consists of rendered images of 3D car models with varying azimuth angles at 15° intervals. We split the dataset into train set and test set, and again split the train set into two groups, each of which is used as A domain and B domain samples. In addition to training a standard GAN model, a GAN with a reconstruction model and a proposed DiscoGAN model, we also trained a regressor that predicts the azimuth angle of a car image using the train set. To evaluate, we translated images in the test set using each of the three trained models, and azimuth angles were predicted using the regressor for both input and translated images. Figure 5 shows the predicted azimuth angles of input and translated images for each model. In standard GAN and GAN with reconstruction (5a and 5b), most of the red dots are grouped in a few clusters, indicating that most of the input images are translated into images with same azimuth
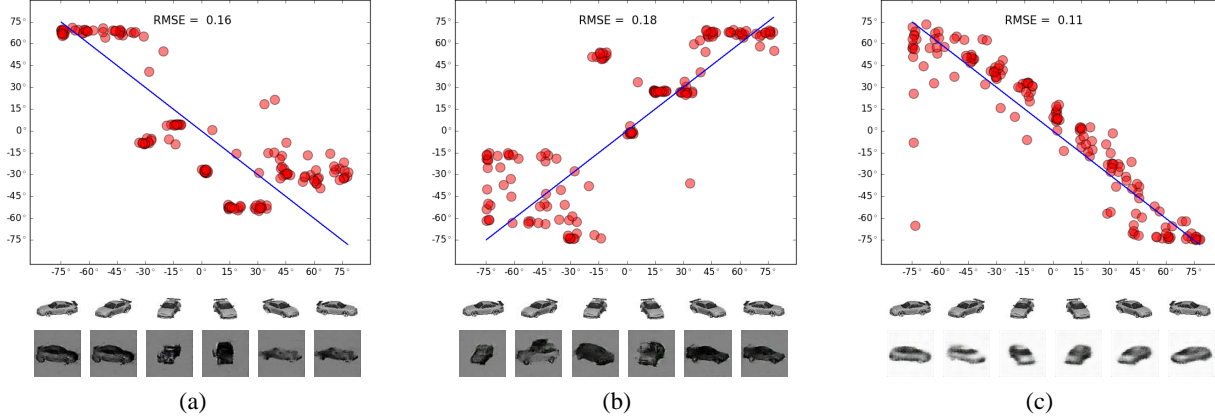
Figure 5. Car to Car translation experiment. Horizontal and vertical axes in the plots indicate predicted azimuth angles of input and translated images, where the angle of input image ranges from -75° to 75°. RMSE with respect to ground truth (blue lines) are shown in each plot. Images in the second row are examples of input car images ranging from -75° to 75° at 15° intervals. Images in the third row are corresponding translated images. (a) plot of standard GAN (b) GAN with reconstruction (c) DiscoGAN. The angles of input and output images are highly correlated when our proposed DiscoGAN model is used. Note the angles of input and translated car images are reversed with respect to 0° (i.e. mirror images) for (a) and (c).

angles, and that these models suffer from mode collapsing problem. Our proposed DiscoGAN (5c), on the other hand, shows strong correlation between predicted angles of input and translated images, indicating that our model successfully discovers azimuth relation between the two domains. In this experiment, the input and translated images either have positive correlation where they have the same range of azimuth angles (5b), or negative correlation where they have opposite range of angles (5a and 5c).

Next, we use a Face dataset (Paysan et al., 2009) shown in Figure 6a, in which the face images vary in azimuth rotation from -90° to +90°. Similar to previous Car to Car experiment, input images in the -90° to +90° rotation range generated output images either in the same range, from -90° to +90°, or the opposite range, from +90° to -90° when our proposed model was used (6d). We also trained a standard GAN and a GAN with reconstruction loss for comparison. When a standard GAN and GAN with reconstruction loss were used, the generated images do not vary as much as the input images in terms of rotation. In this sense, similar to what has been shown in previous Car to Car experiment, the two models suffered from mode collapse.

3.2.2. FACE CONVERSION

We also applied the face attribute conversion task on CelebA and Facescrub dataset (Liu et al., 2015; Ng & Winkler, 2014), where a specific feature, such as gender or hair color, varies between two domains and other facial features are preserved. The results are listed in Figure 7.

In Figure 7a, we can see that various facial features are

well-preserved while a single desired attribute (gender) is changed. Similarly, in Figures 7b and 7c hair color conversion and conversion between with and without eyeglasses are successfully performed.
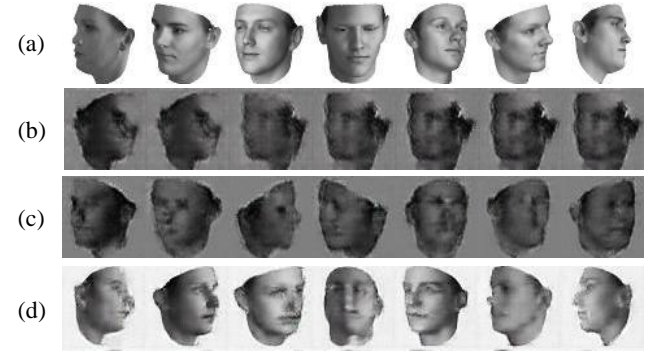

Figure 6. Face to Face translation experiment. (a) input face images from -90° to +90° (b) results from a standard GAN (c) results from GAN with a reconstruction loss (d) results from our DiscoGAN. Here our model generated images in the opposite range, from +90° to -90°.

3.2.3. CHAIR TO CAR, CAR TO FACE

We also investigated the case where the discrepancy between two domains is large, and there is a single shared feature between two domains. 3D rendered images of chair (Aubry et al., 2014) and the previously used car and face datasets (Fidler et al., 2012; Paysan et al., 2009) were used

Figure 7. (a) Translation of gender in Facescrub dataset. (b) Blond to black and black to blond hair color conversion in CelebA dataset. (c) Eyeglasses to no eyeglasses, no eyeglasses to eyeglasses conversion in CelebA dataset.
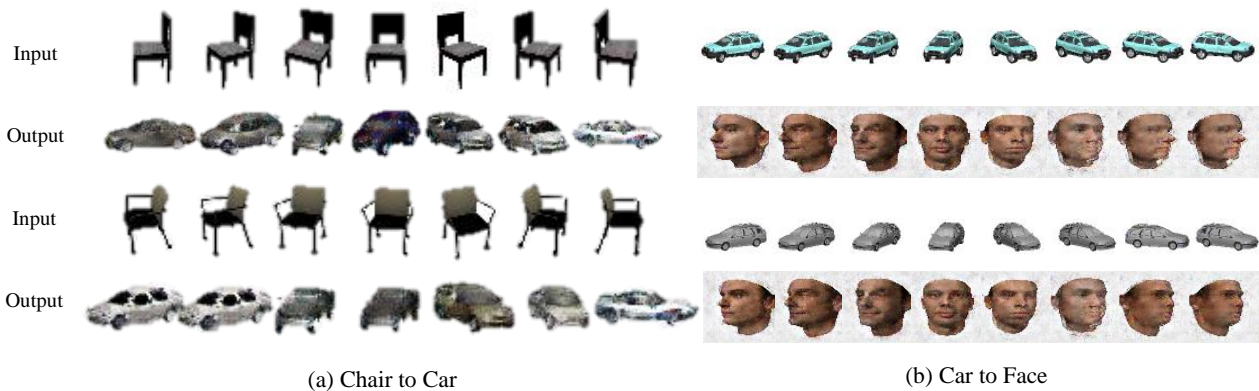

(a) Chair to Car

(b) Car to Face

Figure 8. Discovering relations of images from visually very different object classes. (a) chair to car translation. DiscoGAN is trained on chair and car images (b) car to face translation. DiscoGAN is trained on car and face images. Our model successfully pairs images with similar orientation.

in this task. All three datasets vary along the azimuth rotation. Figure 8 shows the results of image-to-image translation from chair to car and from car to face datasets. The translated images clearly match the rotation features of the input images while preserving visual features of car and face domain respectively.

3.2.4. EDGES TO PHOTOS

Edges-to-photos is an interesting task as it is a 1-to-N problem, where a single edge image of items such as shoes and handbags can generate multiple colored images of such items. In fact, an edge image can be colored in infinitely many ways. We validated that our DiscoGAN performs very well on this type of image-to-image translation task and generate realistic photos of handbags (Zhu et al., 2016) and shoes (Yu & Grauman, 2014). The generated images

Figure 9. Edges to photos experiment. Our model is trained on a set of object sketches and colored images and learns to generate corresponding photos or sketches. (a) colored handbag images are generated from handbag sketches, (b) colored shoe images are generated from shoe sketches, (c) handbag sketches are generated from colored handbag images.

are presented in Figure 9.

3.2.5. HANDBAG TO SHOES, SHOES TO HANDBAG

Finally, we investigated the case with two domains that are visually very different, where shared features are not explicit even to humans. We trained a DiscoGAN using previously used handbags and shoes datasets, not assuming any specific relation between those two. In the translation results shown in Figure 1, our proposed model discovers fashion style as a related feature between the two domains. Note that translated results not only have similar colors and patterns, but also have similar level of fashion formality as the input fashion item.

4. Related Work

Recently, a novel method to train generative models named Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) has been developed. A GAN is composed of two modules – a generator G and a discriminator D. The generator's objective is to generate (synthesize) data samples whose distribution closely matches that of real data samples while the discriminator's objective is to distinguish between real and generated samples. The two models G and D, formulated as a two-player minimax game, are trained simultaneously.

Researchers have studied GANs vigorously in the past few years: network models such as LAPGAN (Denton et al., 2015) and DCGAN (Radford et al., 2016) and improved training techniques (Salimans et al., 2016; Arjovsky et al., 2017) has been explored. More recent GAN works are described in (Goodfellow, 2017).

Several methods were developed to generate images based on GANs. Conditional Generative Adversarial Nets (cGANs) (Mirza & Osindero, 2014) used MNIST digit class label as an additional information to both generator and discriminator and can generate digit images of the specified class. Similarly, Dosovitskiy et al. (2015) showed that GAN can generate images of objects based on specified characteristic codes such as color and viewpoint. Other approaches instead used conditional features from a completely different domain for image generation. For example, Reed et al. (2016) used encoded text description of images as the conditional information to generating images that match the description.

In addition to using conditional information such as class labels and text encodings, several works in the field of image-to-image translation used images of one domain to generate images in another domain. (Isola et al., 2016) translated black-and-white images to colored images by training on paired black-and-white and colored image data. Similarly, Taigman et al. (2016) translated face images to emojis by providing image features from pre-trained face recognition module as conditional input to a GAN.

Several recent papers (Tong et al., 2017; Chen et al., 2016; Makhzani, 2015) used similar architectures which reconstruct the original input by successively applying a mapping and its inverse mapping, and encourage the reconstruction to match the original input.

The most similar work to ours (Mathieu et al., 2016) proposed an architecture that combined variational autoencoder (Kingma et al., 2014) with generative adversarial network for factor decomposition and domain transfer.

5. Conclusion

This paper presents a learning method to discover cross-domain relations with a generative adversarial network called DiscoGAN. Our approach works without any explicit pair labels and learns the relations between datasets from very different domains. One possible future directions is to extend DiscoGAN to handle mixed modalities such as text and image.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. In *arXiv preprint arXiv:1701.07875*, 2017.

Aubry, M., Maturana, D., Efros, A. A., Russell, B., and Sivic, J. 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 (CVPR)*, 2014.

Chen, Xi, Chen, Xi, Duan, Yan, Houthooft, Rein, Schulman, John, Sutskever, Ilya, and Abbeel, Pieter. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016.

Denton, E. L., Chintala, S., Szlam, A., and Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Dosovitskiy, A., Springenberg, J. T., and Brox, T. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Fidler, S., Dickinson, S., and Urtasun, R. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Goodfellow, Ian J. NIPS 2016 tutorial: Generative adversarial networks. In *arXiv preprint arXiv:1701.00160*, 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accerlerating deep network training by reducing internal covariate shift. In *arXiv preprint arXiv:1502.03167*, 2015.

Isola, P., Zhu, J., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *arXiv preprint arXiv:1611.07004*, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

Kingma, D. P., , and Welling, M. Auto-encoding variational bayes. In *Proceedings of the 2rd International Conference on Learning Representations (ICLR)*, 2014.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Makhzani, A., Shlens J. Jaitly N. Goodfellow I. Frey B. Adversarial autoencoders, 2015.

Mathieu, Michael F, Zhao, Junbo Jake, Zhao, Junbo, Ramesh, Aditya, Sprechmann, Pablo, and LeCun, Yann. Disentangling factors of variation in deep representation using adversarial training, 2016.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. In *arXiv preprint arXiv:1411.1784*, 2014.

Ng, H. W. and Winkler, S. A data-driven approach to cleaning large face datasets. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2014.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. A 3d face model for pose and illumination invariant face recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, 2009.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Taigman, Y., Polyak, A., and Wolf, L. Unsupervised cross-domain image generation. In *arXiv preprint arXiv:1611.02200*, 2016.

Tong, C., Li, Y., Jacob, A. P., Bengio, Y., and Li, W. Mode regularized generative adversarial networks. In *Proceedings of the 5rd International Conference on Learning Representations (ICLR)*, 2017.

Xu, B., Wang, N., T., Chen, and Li, M. Empirical evaluation of rectified activations in convolutional network. In *arXiv preprint arXiv:1505:00853*, 2015.

Yu, A. and Grauman, K. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Zhu, Jun-Yan, Krähenbühl, Philipp, Shechtman, Eli, and Efros, Alexei A. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.