IMAGE-TO-IMAGE TRANSLATION USING GENERATIVE ADVERSARIAL NETWORKS

Zhijian Li

October 28, 2020

ABSTRACT

Image-to-image translation is a kind of problem about image processing, where the goal is to learn a mapping between two types of images. Recently, great progress has been made in this field by using Generative Adversarial Networks (GAN) and Variational Auto-Encoders (VAE), which provide outstanding capability of image generation and information disentanglement. Herein, we survey several approaches for the image-to-image translation task, divide them into four different categories by scope of application, including one-to-one method, multi-domain method, multi-modal method and multi-mapping method. We firstly introduce the background of image-to-image translation task and illustrate basic idea of GAN, VAE and their variants. Then, we analyze the principles and existing limits of these approaches in detail, and discuss the difficulties and solutions of image-to-image translation task.

Keywords Image-to-image translation · Generative Adversarial Networks · Variational Auto-Encoders

1 Introduction

For a given image, image-to-image translation task aims at change a particular aspect of it to another. Moreover, many problems in computer vision can be formulated as image-to-image translation task, including super-resolution, colorization, domain adaptation, and style transfer. This task has experienced significant improvments due to the fantastic progress in generative model, such as generative adversarial networks (GAN)[5] and variational auto-encoders (VAE)[11]. Specifically, given training data from two domains, the models need to translate images from one domain to the other. We assume that image has some attributes which represent inherent meaningful features, such as hair color, age, mustache. We further denote that domain is a set of images which share the same attributes, e.g., images of woman could be regarded as one domian while those of man represent another.

Generative Adversarial Networks (GAN) is a kind of widely used generative model which have gained considerable attention. GAN aim at translate random noise to images whose content are similar to those of the given dataset in an unsupervised way. More specifically, GAN train two different kinds of networks simutaneously, one is generator (G), and the other is discriminator (D). The generator takes in a random noise and translate it into a realistic image, while the discriminator is a binary classifier learns to classify the real and generated images properly. Though GAN has made great success, it still suffers from some troubling problems, such as unstable training and mode collapse. Based on the plain GAN model, Wasserstein GAN (WGAN) [1] replaces the binary classifier (discriminator) with a substitute loss function which represents the wasserstein distance between two distributions. WGAN cures the unstable training problem arise in GAN, and the mode collapse phenomenon is also dramatically reduced. Therefore, GAN variants in recent years are mostly based on WGAN rather than the plain GAN. On the other hand, GAN models have introduced many applications, and image-to-image translation task is one of them. Replace the input random noise with images of domain A, and regrad images of domain B as "true" images, we could train a GAN model to translate images from domain A (noise) to domain B (realistic image). Approaches based on GAN model for image-to-image translation task have shown impressive performence, e.g., CycleGAN [18], DiscoGAN [10], StarGAN [3], StarGANv2 [4].

Another widely used generative model is variational auto-encoders (VAE), which can also be divided into two parts like GAN model, a probabilistic encoder network and a probabilistic decoder network. Different from the adversarial training in GAN, encoder network aims at translate an image into a latent code while decoder network learns to reconstruct the input image according to the latent code. Encoded by the same encoder network, latent codes from similar images tend to aggregate which acts like classification algorithms in unsupervised learning. This phenomenon in VAE gives model the ability to extract attribute information from training data, which makes the encode-decode structure widely used in image-to-image translation model. Some researches [14, 7, 13, 2, 12, 17] try to disentangle the information from different domains using this encode-decode structure, and achieve image-to-image translation task by recombining these information.

There are three main challenges in image-to-image translation task: (1) the lack of aligned training pairs, in other words, we need to train models in an unsupervised way. (2) a single input image should be translated into multiple outputs, note that naively adding noise in the input image brings little effect. (3) multi-domain translation using an unified model, since one model for each pair of domain costs a lot and unable to utilize cross-domain information. We investigate several algorithms in the field of image-to-image translation, analyse how they solve these problems and point out their disadvantages.

2 Background

During the research process of image-to-image translation, pix2pix [8] firstly use conditional generative adversarial networks to develop a common framework for this task. However, due to the lack of paired training data, pix2pix is actually not very pratical and CycleGAN was proposed to solve this problem. In the meantime, UINT [14] makes a shared-latent space assumption and propose an unsupervised image-to-image translation framework by combining VAE and GAN. Both CycleGAN and UINT present impressive results in the task with deterministic one-to-one mapping, which means only a pair of images and a pair of domains be translated at a time. Subsequent researches focus on one-to-many translation and explore this field from two perspectives: multi-modal translation and multi-domain translation. MUNIT [7] and DRIT [12] are two multi-modal translation methods, which learn one-to-many mapping between two domains in an unsupervised way. UFDN [13] and StarGAN [3] learn multi-domain translation by an unified model, however, they are not able to generate diverse outputs under only one input image. At last, StarGANv2 [4] and DMIT [17] achieve multi-domain translation and multi-modal translation simutaneously in an unsupervised way.

Note that the progress has been made on the research of image-to-image translation is accompanied by the advance in generative adversarial networks and normalization techniques. More specifically, the usage of Adaptive Instance Normalization (AdaIN) [6] layer in StyleGAN [9] provide an effective way to insert attribute information into the synthetic images, which benefit model design in image-to-image translation task. On the other hand, WGAN helps to stabilize training model and Auxiliary Classifier GAN (ACGAN) [16] provides a novel structure of discriminator which helps class-conditional image synthesis. Moreover, Mode seeking generative adversarial networks (MSGAN) [15] propose a simple and effective regularization term to address the mode collapse issue for GAN, which benefits the diversity of output images.

3 Supervised method: pix2pix

Pix2pix investigate conditional adversarial networks as a general-purpose solution to image-to-image translation problems [8]. More specifically, given paired images from two different domains, pix2pix trains a cGAN model in a supervised way. while GAN is a generative model which learns a mapping from random noise z to a realistic output image y, the conditional GAN used in pix2pix learns a mapping from an input image x combined with a random noise z to the corresponding output image. The generator G is trained to translate images from domain A to domain B that can not be distinguished by an adversarially trained discriminator D, which is trained to detect fake domain B images. The loss function of a conditional GAN is formulated as follow:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log (1 - D(x, G(x, z)))]. \tag{1}$$

Additionally, [8] introduce another supplementary loss function which is beneficial to the training process:

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1].$$

So, the final objective is

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L_1}(G). \tag{2}$$

However, to be clear, although the input data include random noise that used to generate diverse outputs even with a single input image, the generator just simply learns to ignore it. And that's the reason to design specialised networks rather than naively combine image with random noise to achieve multi-modal translation.

4 Unsupervised methods

4.1 one to one methods

4.1.1 CycleGAN

CycleGAN [18] is an approach for learning to translate images between two domains in the absence of paired training data, where pix2pix is incapable. The objective in CycleGAN consists of two different parts, one is traditional GAN loss function for indistinguishable mapping and the other is a cycle consistent term for unpaired training. More specifically, given domain X and domain Y with their corresponding images denoted as x and y, Cyclegan trains two translators $G : X \to Y$ and $F : Y \to X$ to perform image translation, with two discriminators D_X and D_Y to distinguish realistic and generated images for domain X and domain Y respectively. For the mapping function $G : X \to Y$ and its discriminator D_Y, the objective is formulated as follow:

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))]. \tag{3}$$

Where the generator G tries to generator images $G(x)$ which are similar to the images from domain Y, while the discriminator D_Y learns to distinguish between generated images $G(x)$ and realistic images from domain Y. As for generator F and its discriminator D_X, the objective is formulated in the same form:

$$\mathcal{L}_{GAN}(F, D_X, X, Y) = \mathbb{E}_{x \sim p_{data}(x)}[\log D_X(x)] + \mathbb{E}_{y \sim p_{data}(y)}[\log(1 - D_X(F(y)))].$$

Adversarial losses alone can not guarantee a mapping from an input x_i to a desirable output, generally speaking this mapping will not keep the content in image x_i unchanged because of unpaired trainging data. To solve this problem, [18] argue that the mapping function G and F should be cycle-consistent, the image translation cycle should be able to bring x back to the original image: $F(G(x)) \approx x$, $G(F(y)) \approx y$. CycleGAN introduces an extra cycle consistent loss to accomplish this behavior:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1]. \tag{4}$$

The full objective is:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) \tag{5}$$

4.1.2 DiscoGAN

DiscoGAN [10] has the same objective loss funtion as CycleGAN, but it further discuss why bidirectional translation is needed. For translating images between two domains, it seems like only half of the loss functions mentioned above are necessary, e.g., $\mathcal{L}_{GAN}(G, D_Y, X, Y)$ and $\mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1]$. However, this will lead to severe mode collapse, translator F will not work properly if only $\mathcal{L}_{GAN}(G, D_Y, X, Y)$ and $\mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1]$ being used. For more details, referring to the experiment part of DiscoGAN.

4.1.3 UINT

Different from GAN based methods mentioned above, UINT [14] propose an image-to-image translation framework based on a shared-latent space assumption, which assumes a pair of corresponding images in different domains can be mapped to a same latent representation in a shared-latent space. There are three parts in UINT method, two encoders E_1 and E_2 encode images from two domains into the same latent space separately, two decoders G_1 and G_2 translate latent codes to images that belongs to their domain accordingly, two discriminators D_1 and D_2 distinguish between realistic and generated images for the two domains. Given image x_i from domain i, z_i is a latent code generated by encoder E_i and $x_i^{i \to j}$ is a generated image from domain i to domain j through E_i and G_j in sequence. The full objective in UINT can be devided into three kinds of information streams, $\mathcal{L}_{VAE_i}$, $\mathcal{L}_{GAN_i}$ and $\mathcal{L}_{CC_i}$:

$$\begin{aligned}\min_{E_1,E_2,G_1,G_2} \max_{D_1,D_2} & \mathcal{L}_{VAE_1}(E_1, G_1) + \mathcal{L}_{GAN_1}(E_2, G_1, D_1) + \mathcal{L}_{CC_1}(E_1, G_1, E_2, G_2) + \\ & \mathcal{L}_{VAE_2}(E_2, G_2) + \mathcal{L}_{GAN_2}(E_1, G_2, D_2) + \mathcal{L}_{CC_2}(E_2, G_2, E_1, G_1).\end{aligned} \tag{6}$$

$\mathcal{L}_{VAE_i}$ item aims at minimizing a variational upper bound, the objects are:

$$\mathcal{L}_{VAE_1}(E_1, G_1) = \lambda_1 KL(q_1(z_1|x_1)\|p_\eta(z)) - \lambda_2 \mathbb{E}_{z_1 \sim q_1(z_1|x_1)}[\log p_{G_1}(x_1|z_1)]$$

$$\mathcal{L}_{VAE_2}(E_2, G_2) = \lambda_1 KL(q_2(z_2|x_2)\|p_\eta(z)) - \lambda_2 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log p_{G_2}(x_2|z_2)],$$

where the prior distribution is a zero mean Gaussian $p_\eta(z) = \mathcal{N}(z|0, I)$. The GAN objective functions $\mathcal{L}_{GAN_i}$ are given by

$$\mathcal{L}_{GAN_1}(E_2, G_1, D_1) = \lambda_0 \mathbb{E}_{x_1 \sim P_{\mathcal{X}_1}}[\log D_1(x_1)] + \lambda_0 \mathbb{E}_{z_2 \sim q_2(z_2|x_2)}[\log(1 - D_1(G_1(z_2)))]$$

$$\mathcal{L}_{GAN_2}(E_1, G_2, D_2) = \lambda_0 \mathbb{E}_{x_2 \sim P_{\mathcal{X}_2}}[\log D_2(x_2)] + \lambda_0 \mathbb{E}_{z_1 \sim q_1(z_1|x_1)}[\log(1 - D_2(G_2(z_1)))].$$

These are two conditional GAN objective functions with hyperparameter λ_0 controls the impact of these funcitons. Moreover, they use a VAE-like objective funciton to model the cycle-consistency constraint, which is given by

$$\mathcal{L}_{CC_1}(E_1, G_1, E_2, G_2) = \lambda_3 KL(q_1(z_1|x_1)\|p_\eta(z)) + \lambda_3 KL(q_2(z_2|x_1^{1\to2})\|p_\eta(z)) - \lambda_4 \mathbb{E}_{z_2 \sim q_2(z_2|x_1^{1\to2})}[\log p_{G_1}(x_1|z_2)]$$

$$\mathcal{L}_{CC_2}(E_2, G_2, E_1, G_1) = \lambda_3 KL(q_2(z_2|x_2)\|p_\eta(z)) + \lambda_3 KL(q_1(z_1|x_2^{2\to1})\|p_\eta(z)) - \lambda_4 \mathbb{E}_{z_1 \sim q_1(z_1|x_2^{2\to1})}[\log p_{G_2}(x_2|z_1)].$$

Clearly, the negative log-likelihood term accomplish cycle-consistency that a twice translaed image resembles the input one, similar to (4).

4.2 multi-domain method

4.2.1 StarGAN

CycleGAN and UINT achieve translation from one domain to another, but they have limited scalability in handling more than two domains, since different models should be built independently for every pair of image domains. StarGAN was proposed in [3] as a scalable method which can perform image-to-image translation for multiple domains using only one networks. The idea of StarGAN is simple and easy to implement, combining image and target domain information (binary or one-hot vector) as the input data, the generator translate images from any domain to the target domain according to this information. Moreover, training joingly between domains from different datasets is able to work in StarGAN by adding a mask to the target domain label. The objective function consists of three parts, adversarial loss $\mathcal{L}_{adv}$, domain classification loss $\mathcal{L}_{cls}$, and reconstruction loss (cycle consistency loss) $\mathcal{L}_{rec}$ Adversarial loss acts as the traditional GAN loss and aims at generating indistinguishable images, which is formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log D_{src}(x)] + \mathbb{E}_{x,c}[\log(1 - D_{src}(G(x, c)))].$$

Where x is the input image and c is the target domain label, D_{src} represents traditional discriminator outputs like in the plain GAN or WGAN. Under the situation of multi-domain translation in an unified model, discriminator needs to handl images from different domains separately. To achieve such goal, StarGAN adds an auxiliary classifier [16] on top of discriminator and impose the domain classification loss, which is formulated as:

$$\mathcal{L}^r_{cls} = \mathbb{E}_{x,c'}[-\log D_{cls}(c'|x)],$$

$$\mathcal{L}^f_{cls} = \mathbb{E}_{x,c}[-\log D_{cls}(c|G(x, c))].$$

Here c' represents the corresponding domain label of x, and D_{cls} is a classification output which is different from traditional GAN output. With only adversarial loss and domain classification loss does not guarantee the translated images preserve the content of its input images. Cycle consistency constrain reflects in the reconstruction loss of StarGAN, which is defined as:

$$\mathcal{L}_{rec} = \mathbb{E}_{x,c,c'}[\|x - G(G(x, c), c')\|_1].$$

Given an image x with a target domain label c, G translates it to a fake image $G(x, c)$ that belongs to domain c, and learns to reconstruct image x through $G(x, c)$ with the source domain label c'. Finally, the full objective to optimize G and D are

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^r_{cls}, \tag{7}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{rec}\mathcal{L}_{rec}. \tag{8}$$

4.2.2 UFDN

UFDN [13] is another multi-domain image-to-image tranlsation method, it learns disentangled cross-domain information explicitly and achieves tranlsation by combining this information with domain-specific vector. Different from StarGAN, UFDN utilize an encode-decode structure (E and G) to disentangle and recombine attribute information of images, a discriminator D_x to distinguish generated images from realistic images, and another discriminator D_v to eliminate domain-specific information. Importantly, without discriminator D_v, simple encode-decode technique will not guarantee disentangled information, since latent code z still contains image-specific information. Information stream during training of UFDN consists of three parts, one VAE loss and two adversarial losses. To be more specific, UFDN uses an encoder E to derive a latent code z from an input image x_c, which belongs to domain c. Combining the latent code z with its domain vector v_c, G generates a output image $\hat{x}_c$ to reconstruct x_c. This objective function is defined as:

$$\mathcal{L}_{vae} = \|\hat{x}_c - x_c\|_F^2 + KL(q(z|x_c)\|p(z)).$$

To eliminate domain-specific information of latent code z, UFDN introduces domain discriminator D_v, results in the objective function:

$$\mathcal{L}_{D_v}^{adv} = \mathbb{E}[\log P(l_v = v_c|E(x_c))]$$

$$\mathcal{L}_{E}^{adv} = -\mathcal{L}_{D_v}^{adv}$$

Finally, UFDN uses an ACGAN discriminator to not only detect fake images but also classify these images to their corresponding domains, which is the same as StarGAN, commonly used in multi-domain situation. The objective functions are formulated as:

$$\mathcal{L}_{D_x}^{adv} = \mathbb{E}[\log(D_x(\hat{x}_{\overline{c}}))] + \mathbb{E}[\log(1 - D_x(x_c))]$$

$$\mathcal{L}_{G}^{adv} = -\mathbb{E}[\log(D_x(\hat{x}_{\overline{c}}))]$$

$$\mathcal{L}_{cls} = \mathbb{E}[\log P(l_x = v_{\overline{c}}|\hat{x}_{\overline{c}})] + \mathbb{E}[\log P(l_x = v_c|x_c)].$$

4.3 multi-modal method

4.3.1 DRIT and MUINT

DRIT [12] and MUINT [7] are able to generate diverse outputs using a single input image, while have limited scalability in handling more than two domains, which is so-called multi-modal translation. Briefly speaking, DRIT and MUINT inherit the encode-decode structure used in UFDN but with an extra networks to encode domain-specific information. To be more specific, there are two kinds of encoders in these methods. Two domain-specific encoders aim at extracting attribute information for images from two domains separately, and force the distribution of latent code to be a zero mean gaussian, which is an important process in multi-modal translating. Two domain-invariant encoders aim at extracting content information for images cross different domains, it's the same part used in UFDN. Differences between DRIT and MUINT lie at the process in generating images with content vector and attribute vector. DRIT concatenates these two vectors into one just like the way in UFDN, but it inevitably contains an extra discriminator to eliminate domain-specific information as we have discussed in 4.2.2. On the other hand, MUINT imitate the generator architecture used in StyleGAN, which consists of multiple convolution and AdaIN layers, and therefore no need for the discriminator. This kind of architecture automatically eliminate domain-specific information and insert new attribute information layer-by-layer, from coarse features to fine features. Both DRIT and MUINT generate diverse and high quality images, which is a huge progress in the field of image-to-image translation.

4.4 multi-mapping method

4.4.1 StarGAN v2

Previous work solve multi-domain and multi-modal problems separately, but in the most cases, a unified model that address both issues simultaneously is highly demanded. StarGAN v2 was proposed in [4], a single framework that tackles both the problems and shows significantly improved results over the baselines. In summary, StarGAN v2 enriches StarGAN with the networks used in StyleGAN and a style encoder structure used in other researches. As for the proposed framework, it consists of four parts, a generator G, a mapping network F, a style encoder E, and a discriminator D. Note that the layers in all of them, except G, consist of two modules, some shared layers and one domain specific output for each domain. On the other hand, StarGAN v2 trains the framework using the following objectives, adversarial objective $\mathcal{L}_{adv}$, style reconstruction $\mathcal{L}_{sty}$, diversity sensitive $\mathcal{L}_{ds}$, and cycle consistency loss $\mathcal{L}_{cyc}$. During the training, StarGAN v2 sample a latend code z and a target domain $\tilde{y}$ randomly, generates a target style code $\tilde{s} = F_{\tilde{y}}(z)$ using mapping network F. Adversarial objective is formulated as follow:

$$\mathcal{L}_{adv} = \mathbb{E}_{x,y}[\log D_y(x)] + \mathbb{E}_{x,\tilde{y},z}[\log(1 - D_{\tilde{y}}(G(x,\tilde{s})))].$$

Where $D_y(\cdot)$ denotes the output of D corresponding to domain y. The mapping network F learns to provide the style code $\tilde{s}$ that is likely in the target domain $\tilde{y}$. The Style reconstruction loss is defined as:

$$\mathcal{L}_{sty} = \mathbb{E}_{x,\tilde{y},z}[\|\tilde{s} - E_{\tilde{y}}(G(x,\tilde{s}))\|_1].$$

This objective trains a mapping from an input image to its style code. To further enable the generator G to produce diverse images, StarGAN v2 regularize G with a diversity sensitive loss:

$$\mathcal{L}_{ds} = \mathbb{E}_{x,\tilde{y},z_1,z_2}[\|G(x,\tilde{s}_1) - G(x,\tilde{s}_2)\|_1].$$

Where the target style codes $\tilde{s}_1$ and $\tilde{s}_2$ are produced by F conditioned on two random latent codes z_1 and z_2. To preserve the content of input images, StarGAN v2 employs the cycle consistency loss:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x,y,\tilde{y},z}[\|x - G(G(x,\tilde{s},\hat{s}))\|_1],$$

where $\hat{s} = E_y(x)$ is the estimated style code of the input image x, and y is the original domain of x. Finally, the full objective functions can be summarized as follows:

$$\min_{G,F,E} \max_{D} \mathcal{L}_{adv} + \lambda_{sty}\mathcal{L}_{sty} - \lambda_{ds}\mathcal{L}_{ds} + \lambda_{cyc}\mathcal{L}_{cyc}, \tag{9}$$

StarGAN v2 further trains the model in the same manner as the above objective, using reference images instead of latent vectors when generating style codes.

5 Conclusion and future research directions

In this study, we have presented a survey of the image-to-image translation task, its targets and main challenges, and various methods to sovle these problems. For unpaired training data issue, cycle consistency loss was introduced in CycleGAN to handle this situation, which became a commenly used term in other methods from then on. To break the limitaion of multi-domain translation, StarGAN concatenate the input image with a target domain label to generate output images of certain domains. After that, encode-decode architecture which is derived from VAE was introduced to realize multi-modal translation by disentangling domain-invariant and domain-specific information. As for achieving multi-domain and multi-modal simutaneously, StarGAN v2 absorbs advance techniques from previous researches and obtain impressive results. The above discussion shows that GAN and VAE are commenly used approaches in this task, and we have seen a tendency to mix GAN models with encode-decode architecture, which benefits a lot for the disentanglement between attribute information and content information.

However, irrespective of the significant progress of this task in recent years, several problems remains to be solved for future research. First, though we could disentangle attribute information and content information of a given image, we still unable to manipulate certain attributes of images, especially the scene of human face translation. For example, when translating one human face to another, many attributes are changed simutaneously, such as hair color, eye color, eyeglasses, we can not decide which attribute to be kept.

Second, the essence of GAN model aims at translating one distribution to another, however, WGAN does not perform well in the case of high dimension distribution with few training samples. Images from the same domain could be regarded as sample vetors from an unknown distribution, therefore, high resolution images represent vectors of high dimension. Actually, image-to-image translation task can be interpreted as translating one distribution to another, but always with few training images in one dataset, which cause poor translation performance of high resolution images when using the discriminator in WGAN.

We wish this survey will give a comprehensive overview of image-to-image translation task and help the researchers to propose more advance and novel methods.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] JieZhang Cao, Langyuan Mo, Qing Du, Yong Guo, Peilin Zhao, Junzhou Huang, and Mingkui Tan. Joint wasserstein distribution matching. *arXiv preprint arXiv:2003.00389*, 2020.

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[6] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[7] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[10] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

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

[12] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[13] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in neural information processing systems*, pages 2590–2599, 2018.

[14] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[15] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019.

[16] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651, 2017.

[17] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *Advances in Neural Information Processing Systems*, pages 2994–3004, 2019.

[18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.