Comparative Review of Cross-Domain Generative Adversarial Networks

B Zeno¹, I Kalinovskiy² and Yu Matveev¹

¹ITMO University, St. Petersburg, Russia
basilzeno@gmail.com; matveev@mail.ifmo.ru

²STC-innovation Ltd s, St. Petersburg, Russia
kalinovskiy@speechpro.com

Abstract. This paper provides the comparative analysis between two recent image-to-image translation models that based on Generative Adversarial Networks. The first one is UNIT which consists of coupled GANs and variational autoencoders (VAEs) with shared-latent space, and the second one is Star-GAN which contains a single GAN model. Given training data from two different domains from dataset CelebA, these two models learn translation task in two directions. The term domain denotes as a set of images sharing the same attribute value. So, the attributes that are prepared: eye glasses, blond hair, beard, smiling and age. Five UNIT models are trained separately, while only one Star-GAN model is trained. For evaluation, we conduct some experiments and provide a quantitative comparison using direct metric GAM (Generative Adversarial Metric) to quantify the ability of generalization and the ability of generating photorealistic photos. The experimental results show the superiority of cross-model UNIT over multi-model StarGAN on generating age and eye glasses attributes, and the equivalent performance to synthesize other attributes.

1. Introduction

Many problems in computer vision can be formulated as a translation problem, mapping an input image in one domain to a corresponding output image in another domain by changing a particular aspect or more in the given image. For example, changing the seasons of scenery images; changing the resolution of images from low to high; translating cat or dog images between different species. And more, this translation can handle details of the face such as adding expressions like a smile and adding attributes like beard or sunglass. For many years of research, this problem is studied in supervised settings, where paired input-output examples in different domains are available [1,2]. Due to the difficulty and cost of collecting this data, little a number of unsupervised solutions are proposed [3,4,5,6], where there are two independent sets of images, the first one consists of images in one domain, while the other consists of images in another domain.

Traditionally, many of image-to-image translation tasks have been tackled for special purpose. For example, image colorization [7]; removing the effects of camera shake from blurred images [8]; composing realistic picture from a simple freehand sketch is presented [9,10]; synthesizing a plausible image at a different time of day from an input image [11]. But Lately with the remarkable progress in generative machine learning models, a quite a number of research dedicated to the use of Generative Adversarial Networks (GANs) to address translation problem. Some of them are general to solve all these tasks. For instance, pix2pix [2] for synthesizing photos from label maps, reconstructing objects from edge maps; and colorizing images; CycleGAN [5] for style transfer, season transfer, photo enhancement and generating photos from paintings; icGAN [6], StarGAN [4] and UNIT [3] for
adding facial expressions and attributes; UNIT also tested for other purposes such as changing street scene image and cat species conversion.

The two-recent image-to-image translation models, StarGAN and UNIT have achieved impressive results, but there no experiments to determine which one of them is better. In UNIT, the domain adaption experiment is conducted between the Street View House Number (SVHN) dataset, MNIST and USPS datasets. Then the performance of adapted classifiers is evaluated using classification accuracy and compared to performance of other domain adaptation algorithms. In StarGAN, facial attributes transfer task is evaluated using Amazon Mechanical Turk (AMT), while facial expression synthesis task is evaluated using external facial expression classifier and the results are compared to [5,6].

In this paper, we conduct some experiments and provide a comparative analysis between StarGAN and UNIT to measure the ability of generalization and ability of generating photorealistic photos. The main contribution is demonstrating that the cross-model UNIT is better than the multi-model StarGAN. Despite, the superiority of StarGAN over its baseline cross-domain models.

2. Background
We begin by reviewing the relevant background material on Variational Autoencoders [22] and Generative Adversarial Networks [23].

2.1 Variational Autoencoder
Variational Autoencoder (VAE) [22] as a deep generative model which is based on a regularized version of the standard autoencoder. It modifies the autoencoder architecture by learning the representation as a latent variable \( z \) in a probabilistic generative model (e.g. Gaussian distribution) instead of the deterministic function. VAE is comprised of two multilayer perceptrons: encoder and decoder. The encoder maps an observation \( x_i \) into its latent representation \( z_i \) (or latent code) and a decoder (the generative network) decodes an observation from a latent code. The generative process can be written mathematically:

\[
    z_i \sim p(z), x_i \sim p(x|z_i)
\]

where \( p(z) \) is the prior distribution over the latent variables and \( p_{\theta}(x|z_i) \) is probabilistic decoder network with parameters \( \theta \). The approximate posterior distribution of this generative process, call it, \( q_{\phi}(z|x_i) \) (with parameters \( \phi \)).

To admit efficient inference, the variational Bayes approach simultaneously learns both the parameters of \( p_{\theta}(x|z_i) \) as well as those of a posterior approximation \( q_{\phi}(z|x_i) \). This is achieved by maximizing the evidence lower bound (ELBO):

\[
    L(\phi, \theta, x) = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x, z) - \log q_{\phi}(z|x)],
\]

with \( L(\phi, \theta, x) \leq \log p_{\theta}(x) \).

2.2 Generative Adversarial Networks
As one of the most significant improvements on the research of deep generative models, GAN framework was introduced by Goodfellow et al. [23] for generative modeling of data through learning a transformation from points belonging to a simple prior distribution \( z \sim p_z \) to those from the data distribution \( x \sim p_{data} \). A typical GAN model consists of two modules that play an adversarial game: a discriminator and a generator. while the generator learns to generate fake samples \( G(z) \) that are indistinguishable from real samples, the discriminator learns to distinguish between these fake samples \( G(z) \sim p_G \) and real ones \( x \sim p_{data} \), thus giving a scalar output \( y \in \{0, 1\} \). The goal of the generator is to fool the discriminator by generating photorealistic samples that resemble those from the real data while that of the discriminator is to accurately distinguish between real and generated data. The two models, typically designed as neural networks, play a min max game with the objective function as shown in Equation (1):

\[
    \min_{G} \max_{D} V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]
\]
Image to Image Translation Models

In this section, we present overview for both image-to-image translation models, StarGAN and UNIT:

2.3 UNIT: Unsupervised Image-to-Image Translation Networks

Liu et al. [3] proposed unsupervised image-to-image translation framework based on Coupled VAE-GANs. They analysed the problem from a probabilistic modeling perspective, where the goal is to learn a joint distribution of images in different domains. But in the unsupervised settings, where the two sets consist of images from two marginal distributions in two different domains, make inferring the joint distribution from these marginal distributions is a highly ill-posed problem, due to existence an infinite set of joint distributions. To address this ill-posed problem, they make a shared-latent space assumption [13], which assumes a pair of corresponding images in different domains can be mapped to a same latent representation in a shared-latent space. Based in this assumption, as shown figure 1, the system consists of 6 subnetworks: including two encoders E1 and E2 for mapping images to shared-latent codes z, two generators G1 and G2 for mapping latent codes z to images, and two adversarial discriminators D1 and D2 for the respective domains, in charge of evaluating whether the translated images are fake or realistic.

In UNIT, the input image can be reconstructed from translating back the translated input image. Since the assumption mentioned above implies the cycle consistency constraint [14,15], which means an image in the source domain can be mapped to an image in the target domain and this translated image in the target domain can be mapped back to the original image in the source domain.

To learn UNIT, learning problems of coupled VAE-GANs are solved jointly for the image reconstruction streams, the image translation streams, and the cycle-reconstruction streams:

\[
\begin{align*}
\min_{E_1, E_2, G_1, G_2, D_1, D_2} & \max_{L_{VAE_1}(E_1, G_1), L_{VAE_2}(E_2, G_2), L_{GAN_1}(E_1, G_1, D_1), L_{GAN_2}(E_2, G_2, D_2), L_{CC_1}(E_1, G_1, E_2, G_2), L_{CC_2}(E_2, G_2, E_1, G_1)} \\
& L_{VAE_1}(E_1, G_1, z) + L_{GAN_1}(E_1, G_1, D_1) + L_{CC_1}(E_1, G_1, E_2, G_2) + L_{VAE_2}(E_2, G_2, z) + L_{GAN_2}(E_2, G_2, D_2) + L_{CC_2}(E_2, G_2, E_1, G_1)
\end{align*}
\]

Where, \( L_{VAE_1} \) and \( L_{VAE_2} \) are objective functions of VAEs; \( L_{GAN_1} \) and \( L_{GAN_2} \) are objective functions of GANs; \( L_{CC_1} \) and \( L_{CC_2} \) are objective functions to model the cycle-consistency constraint.

UNIT learns translation between only two domains. So, as demonstrated in figure 2(a), in order to learn mapping between k domains, \((k(k-1))/2 \) UNIT model have to be trained, where each one consists from two coupled generators.

\[\text{Figure 1. UNIT framework. E1, E2, G1 and G2 are presented using CNNs. Here, Here, } z_{1-t^1} \text{ and } z_{1-t^2}^{+2} \text{ are self-reconstructed images, and } z_{1-t^2}^{+1} \text{ and } z_{2-t^1}^{+1} \text{ are domain-translated images, z is shared-latent space assumption using a weight sharing constraint (illustrated using dashed lines).}\]

2.4 StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation

In contrast to UNIT framework, Choi et al. [4] proposed StarGAN, a novel generative approach that can perform image-to-image translations for multiple domains using only a single GAN model. As demonstrated in figure 2(b), StarGAN takes training images of multiple domains, and learns the mappings between all available domains using only one generator. To achieve this mapping, the idea is simple, instead of learning fixed translation (e.g., young-to-old face), the model takes both input image and target domain information \((x,c)\) and learns to flexibly translate the input image into an output image \(y\) on the corresponding domain. The domain information is represented by a label
(binary or one hot vector for categorical attributes). And during training, target domain labels $c$ are generated randomly so that the generator $G$ learns to flexibly translate the input image: $c, G(x, c) \rightarrow y$. Auxiliary classifier [15] is introduced on top of the discriminator to allow him to control multiple domains. So, two probability distributions are produced by discriminator over both sources and domain labels, $D: x \rightarrow \{D_{src}(x), D_{cls}(x)\}$.

To learn StarGAN, the following objective functions are used to optimize both discriminator $D$ and generator $G$, respectively:

$$L_D = -L_{adv} + \lambda_{cls} L_{cls}^{f},$$

$$L_G = L_{adv} + \lambda_{cls} L_{cls}^{r} + \lambda_{rec} L_{rec},$$

where, $L_{adv}$ is adversarial loss, to make the generated images indistinguishable from real images; $L_{rec}$ is reconstruction loss to make the generated images are realistic and classified to its correct target domain with preserving the content of its input images; $L_{cls}^{f}$ is domain classification loss for real images, while $L_{cls}^{r}$ is domain classification loss for fake images; $\lambda_{cls}$ and $\lambda_{rec}$ are hyper-parameters.

In StarGAN, the original image can be reconstructed from translated image due to applying the cycle consistency loss [14,15] to the generator.

![Figure 2. Comparison between UNIT and StarGAN frameworks [4]. a) Six UNIT models are required to learn translation task between 4 attributes. Every UNIT model contains two generators. b) One StarGAN is required only to learn any number of attributes.](image)

3. The Proposed Evaluation Method

Several measures have been introduced for evaluating and comparing GANs, but there is no consensus as to which measure best captures strengths and limitations of models. In comparative analysis between two image-to-image translation models, we must determine which model generates photorealistic images with desired attribute and which model preserves the identity. So, in this section, we review firstly some quantitative and qualitative measures for evaluating GAN models. Then we present our evaluation method to compare StarGAN and UNIT models.

3.1 Qualitative Measures

Visual examination of samples by humans is one of the common and most intuitive ways to evaluate GANs. For instance, in [4] to evaluate StarGAN two user studies are performed in a survey format using Amazon Mechanical Turk (AMT), where the Turkers were asked to judge whether given images with single and multiple transferred attribute are real from the dataset or generated from StarGAN model. In that case, the discriminator can be viewed as a human, while the generator is a trained GAN. The problems with this approach are that Turkers may have high variance, which makes it necessary to average over a large number of them, and the experimental setup is expensive, time-consuming, difficult to reproduce, and does not fully reflect the capacity of models.

3.2 Quantitative Measures

Quantitative measures are less subjective may not directly correspond to how humans perceive and judge generated images. We dived them into two categories: indirect and direct measures. Indirect measures when the evaluation approaches rely on trained external model. For example, FaceNet
convolutional neural network in [17] is used to measure the identity-similarity between each reconstructed face image and the source face image; facial expression classifier is used in [4] to evaluate the generator of StarGAN model in synthesizing realistic facial expression, where the classifier is trained, then the model is evaluated by computing the classification error of a facial expression on reconstructed images. It achieves the lowest classification error in comparison with CycleGAN [5] and IcGAN [6]. Inception Score (IS) [18] is used as indirect metric to measure the quality and diversity of the generated images. The second category is direct measures when the evaluation of the model depends on its own structure without any external model. For example, classification accuracy of UNIT discriminator [3] is used to evaluate its performance in domain adaptation task.

Generative Adversarial Metric (GAM) [19] is the direct measure to quantify the ability of generalization and the ability of generating photorealistic images. It is used to determine which GANs generator is better, considering the performance of discriminators are equivalent. GAM compares two generative adversarial models by pitting each generator against the opponent’s discriminator. Consider two generative adversarial models $M_1=\{G_1,D_1\}$ and $M_2=\{G_2,D_2\}$, in the training phase, $G_1$ competes with $D_1$. Likewise, for $G_2$ and $D_2$. In the test phase, $G_1$ engages in a battle against $D_2$ while $G_2$ against $D_1$. The ratios of their classification errors $e$ on real test dataset and on generated samples are then calculated as $r_{\text{test}}$ and $r_{\text{samples}}$ as shown in Equations 7 and 8.

\[
r_{\text{test}} = \frac{e(D_1(x_{\text{test}}))}{e(D_2(x_{\text{test}}))}
\]

\[
r_{\text{samples}} = \frac{e(D_1(G_2(z)))}{e(D_2(G_1(z)))}
\]

The test ratio, $r_{\text{test}}$, tells us which model generalizes better since it is based on discriminating the test data. While, the sample ratio, $r_{\text{samples}}$, tells us which model can fool the other model more easily, since it is based on generating photorealistic photos. Therefore, for $G_1$ to win against $G_2$, both $r_{\text{samples}}<1$ and $r_{\text{test}}\approx1$ must be satisfied.

In [20] GAM is used in order to compare between Original GAN and CapsuleGAN, which have different discriminators, but the ratios of classification accuracies $A$ are calculated instead of errors to avoid numerical problems. So, in order to compare StarGAN and UNIT models, and Following [3,4,19,20], the new $r_{\text{test}}$ and $r_{\text{translating}}$ are shown in Equations 9 and 10.

\[
r_{\text{test}} = \frac{A(D_{\text{StarGAN}}(x_{\text{test-off}}))}{A(D_{\text{UNIT}}(x_{\text{test-off}}))}
\]

\[
r_{\text{translating}} = \frac{A(D_{\text{StarGAN}}(G_{\text{UNIT,OFF-ON}}(x_{\text{test-off}})))}{A(D_{\text{UNIT}}(G_{\text{StarGAN,OFF-ON}}(x_{\text{test-off}})))}
\]

where $D_{\text{StarGAN}}$ and $D_{\text{UNIT}}$ are discriminators of StarGAN and UNIT respectively. $G_{\text{StarGAN}}$ and $G_{\text{UNIT}}$ are generators of StarGAN and UNIT respectively. $x_{\text{test-off}}$ is a dataset of images, which has not an attribute that the generators will add. The winning model is defined as follows:

\[
\text{Winner} = \begin{cases} \text{UNIT}, & \text{if } r_{\text{translating}} > 1 \text{ and } r_{\text{test}} \approx 1 \\ \text{StarGAN}, & \text{if } r_{\text{translating}} < 1 \text{ and } r_{\text{test}} \approx 1 \\ \text{Tie}, & \text{otherwise} \end{cases}
\]

4. Experiments
In this section, we first review the training and testing datasets. Next, we perform facial attribute transfer using generators of both StarGAN and UNIT models on unseen images during the training phase. lastly, we calculate the evaluation metric GAM by conducting a classification experiment on the translated images using the discriminators of models.
Figure 3. Examples of facial attributes transfer task using StarGAN and UNIT.
4.1 Datasets
Several image datasets come with a number of labeled attributes. For instance, the CelebA [17] dataset contains 40 labels related to facial attributes such as hair color, gender, and age.

- **CelebA.** The CelebFaces Attributes (CelebA) dataset [21] contains 202,599 face images of celebrities, each annotated with 40 binary attributes. We use the MTCNN [24] detector to align the face images and scale them to a size of 128x128. We select all images for training step. Five attributes are chosen as follows: eye glasses, blond hair, beard, smiling and young. We construct ten domains using the following attributes: eye glasses (yes/no), blond hair (yes/no), beard (yes/no), smiling (yes/no) and age (young/old).

- **LFW+.** The extended LFW database (LFW+) [25] contains 15,699 unconstrained face images of about 8,000 subjects, each image is annotated with 3 categorical attributes (age, gender, race). We use LFW+ version with extended annotations, where there are more 15 categorical attributes. We perform preprocessing similarly to CelebA. Then, we choose images with the following attributes: eye glasses, blond hair, beard, smiling and age. Next, we delete images with attribute value (age=baby, age=children) to ensure that the model correctly learns the remaining values. Lastly, we convert the attributes to binary form.

4.2 Training on CelebA
Both models StarGAN and UNIT are trained on NVIDIA GeForce GTX 1080Ti using all images in CelebA. In more detail, one StarGAN model is trained for 2 days, while 5 models of UNIT are trained for 4 days, one UNIT model for each attribute. For data augmentation we flip the images horizontally with a probability of 0.5. The batch size is set to 16 for training StarGAN and 1 for training UNIT.

4.3 Experimental Results on LFW+
During training UNIT model, we notice that the training could be unstable. This problem is mentioned in the original work [3]. So, after more than one million iterations, we choose the best models that give high classification accuracy. Next, to perform facial attribute transfer task, we select from LFW+ images with no targeted attribute, in order to add them by generators. Ten experiments are conducted to add 5 attributes for both models.

- **Qualitative evaluation.** Figure 3, shows the facial attribute transfer results on LFW+ dataset. In contrast to results of StarGAN work, we observed that UNIT method provides a higher visual quality of translation results on test data compared to the multi-domain model StarGAN. We noticed also disadvantage in StarGAN, where we can not turn off the translation of the existed attributes. For example, if the source image is labeled with two attributes: age (old) and eye glasses (yes), and we want to change age from old to young. The translated image will be with eye glasses surly, but in the fact, eye glasses will be added also into the translated image. Whilst, in UNIT, only the targeted attribute will be added. See figure 4.

![Figure 4. Translating image example with attribute age (young)](image)

- **Quantitative evaluation.** In order to evaluate the ability of generators to synthesize photorealistic images, GAM is used as discussed in 5th section. StarGAN discriminator is used to classify translated images by UNIT, while UNIT discriminator is used to classify the translated images by StarGAN. Sigmoid function is used to obtain a probability for sample classification. Table 1, shows the results of classification testing images that have the targeted attribute. The reported numbers are classification accuracies.
Table 1. Classification accuracy of testing images (XTEST-ON)

| Facial attribute | StarGAN | UNIT | Number of images | r_{test} |
|------------------|---------|------|------------------|----------|
| Eye glasses      | 93 %    | 80 % | 2348            | 1.16     |
| Blond hair       | 91 %    | 41 % | 1064            | 2.22     |
| Beard            | 8 %     | 72 % | 3486            | 0.11     |
| Smiling          | 5 %     | 90 % | 5064            | 0.05     |
| Age (young)      | 18 %    | 25 % | 3336            | 0.72     |

We notice that the discriminator of StarGAN is superior to discriminator of UNIT in classifying blond hair attribute. While the discriminator of UNIT achieves better performance in classifying attributes: beard and smiling. But, it does not mean the generator of StarGAN or UNIT is better. So, according to our metric, both models, StarGAN and UNIT on these three attributes are equivalent, because the r_{test} is not equal approximately to one. While we can continue comparing on attributes: eye glasses and age (young). Table 2, shows the classification accuracies of translated images and its ratio r_{translating}.

Table 2. Classification accuracy of translated images (XTEST-OFF→ON)

| Facial attribute | D_{StarGAN}(G_{UNIT}) | D_{UNIT}(G_{StarGAN}) | r_{translating} |
|------------------|------------------------|-----------------------|-----------------|
| Eye glasses      | 89.40 %                | 39 %                  | 2.30            |
| Age (young)      | 54.32 %                | 7.43 %                | 7.31            |

So, from Table. II, we notice high classification accuracy in StarGAN discriminator, while low value in UNIT discriminator. It means that the generator of UNIT superior to StarGAN generator, where the 89.4% of total testing images have eye glasses, while only 22.20% that generated by StarGAN have eye glasses. We refer to the translated xtest-off with targeted attributes as xtest-off→on. The number of xtest-off images is 9263 and 8275 for eye glasses attribute and age attribute respectively. Figure 5, shows examples of generated images for age (young) and eye glasses attributes.

Figure 5. Examples of facial attributes transfer task using StarGAN and UNIT, the left column for adding eye glasses attribute and the right column to add age (young) attribute.
5. Conclusion

In this paper, the comparative analysis between two recent image-to-image translation models based on Generative Adversarial Networks, multi-domain model StarGAN and cross-domain model UNIT provided. To quantify the ability of generating photorealistic photos and the ability to generalize, GAM metric is used to compare these models by having them engage in a “battle” against each other. Since, both StarGAN and UNIT models consist of a discriminator and a generator in pairs, firstly we trained them on CelebA dataset to learn facial attribute transfer task on five attributes: eye glasses, blond hair, beard, smiling and age. Then we exchanged the pairs of generators and discriminators and have them play the generative adversarial game against each other by generating images from testing dataset LFW+ with targeted attributes. The experimental results show the superiority of UNIT model over StarGAN, although they are equivalents in generating: beard, blond hair, smiling, but generator of UNIT performs better than StarGAN generator in translating images with age and eye glasses attributes. Next, we will conduct experiments to measure the quality and diversity of the generated images using Inception Score (IS), and to measure the identity-preserving using FaceNet classifier.

Acknowledgments

This work was financially supported by the Government of the Russian Federation (Grant 08-08).

References

[1] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” International Conference on Computer Vision, 2015.
[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” Conference on Computer Vision and Pattern Recognition, 2017.
[3] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” arXiv preprint arXiv:1703.00848, 2017.
[4] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, “StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[5] J.-Y. Zhu, P. Isola, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” Conference on Computer Vision and Pattern Recognition, 2017.
[6] iGAN: G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Alvarez, “Invertible conditional gans for image editing,” arXiv preprint arXiv:1611.06355, 2016.
[7] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” ECCV, 2016.
[8] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, “Removing camera shake from a single photograph. ACM Transactions on Graphics (TOG), 2006.
[9] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, “Sketch2photo: internet image montage,” ACM Transactions on Graphics (TOG), 28(5):124, 2009.
[10] G.A. Kukharev, Yu.N. Matveev, A.L. Oleinik, “Mutual image transformation algorithms for visual information processing and retrieval,” Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2017, vol. 17, no. 1, pp. 62–74. doi: 10.17586/2226-1494-2017-17-1-62-74 (in Russian).
[11] Y. Shih, S. Paris, F. Durand, and W. T. Freeman, “Data-driven hallucination of different times of day from a single outdoor photo,” ACM Transactions on Graphics (TOG), 32(6):200, 2013.
[12] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” Advances in Neural Information Processing Systems workshop, 2011.
[13] M.-Y. Liu and O. Tuzel, “Coupled generative adversarial networks,” Advances in Neural Information Processing Systems, 2016.
[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” International Conference on Computer Vision, 2017.
[15] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” International Conference on Machine Learning, 2017.
[16] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier gans,”
arXiv preprint arXiv:1610.09585, 2016.

[17] L. Zhigang, and L. Yupin, “Generate Identity-Preserving Faces by Generative Adversarial Networks,” arXiv preprint arXiv: 1706.03227. 2017

[18] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” In Advances in Neural Information Processing Systems, pp. 2226–2234, 2016.

[19] D. J. Im, C. D. Kim, H. Jiang, and R. Memisevic, “Generative adversarial metric,” 2016.

[20] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan, “CapsuleGAN: Generative Adversarial Capsule Network,” arXiv: 1802.06167, 2018

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

[22] D. Kingma, and M. Welling. “Auto-encoding variational bayes,” 12 2013.

[23] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” In Advances in Neural Information Processing Systems (NIPS), pages 2672–2680, 2014.

[24] K. Zhang, Zhang, Z. Zhang, Z. Li and Y. Qiao, (2016), “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Processing Letters, 23(10):1499–1503.

[25] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, “Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach,” To appear in IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), 2017, DOI: 10.1109/TPAMI.2017.2738004.