Conditional Dual-Agent GANs for Photorealistic and Annotation Preserving Image Synthesis

Zhecan Wang*[1]
zhecan.wang@students.olin.edu
Jian Zhao*[23]
zhaojian90@u.nus.edu
Yu Cheng[4]
chengyu996@gmail.com
Shengtao Xiao[2]
xiao_shengtao@u.nus.edu
Jianshu Li[2]
jianshu@u.nus.edu
Fang Zhao[2]
elezhf@nus.edu.sg
Jiashi Feng[2]
elefjia@nus.edu.sg
Ashraf Kassim[2]
ashraf@nus.edu.sg

[1] Franklin. W. Olin College of Engineering
[2] National University of Singapore
[3] National University of Defense Technology
[4] Nanyang Technological University

Abstract

Conditional and semi-supervised **G**enerative **A**dversarial **N**etworks (GANs) have been proven to be effective for image synthesis with preserved annotation information. However, learning from GAN generated images may not achieve the desired performance due to the discrepancy between distributions of the synthetic and real images. To narrow this gap, we expand existing generative methods and propose a novel **C**onditional **D**ual-**A**gent **GAN** (CDA-GAN) model for photorealistic and annotation preserving image synthesis, which significantly benefits object classification and face recognition through **D**eep **C**onvolutional **N**eural **N**etworks (DCNNs) learned with such augmented data. Instead of merely distinguishing "real" or "fake" for the generated images, the proposed dual agents of the Discriminator are able to preserve both of realism and annotation information simultaneously through a standard adversarial loss and an auxiliary annotation perception loss. During the training process, the Generator is conditioned on the desired image features learned by a pre-trained CNN sharing the same architecture of the Discriminator yet different weights. Thus, CDA-GAN is flexible in terms of the scalability and able to generate photorealistic image with well preserved class labeling information for learning DCNNs in specific domains. We perform qualitative and quantitative experiments to verify the effectiveness of our proposed method, which outperforms other state-of-the-arts on MNIST hand written digits classification dataset and **N**ational **I**nstitute of **S**tandards and **T**echnology (NIST) **I**ARPA **J**anus **B**enchmark **A** (IJB-A) face recognition dataset.

* indicate equal contributions.

Furthermore, we also prove that the CDA-GAN generated data represent the distinct class relationships as well as the real data, so adding such data for training DCNN models ends up with impressive improvement in terms of overall accuracy, generalization capacity, and robustness.

1 Introduction

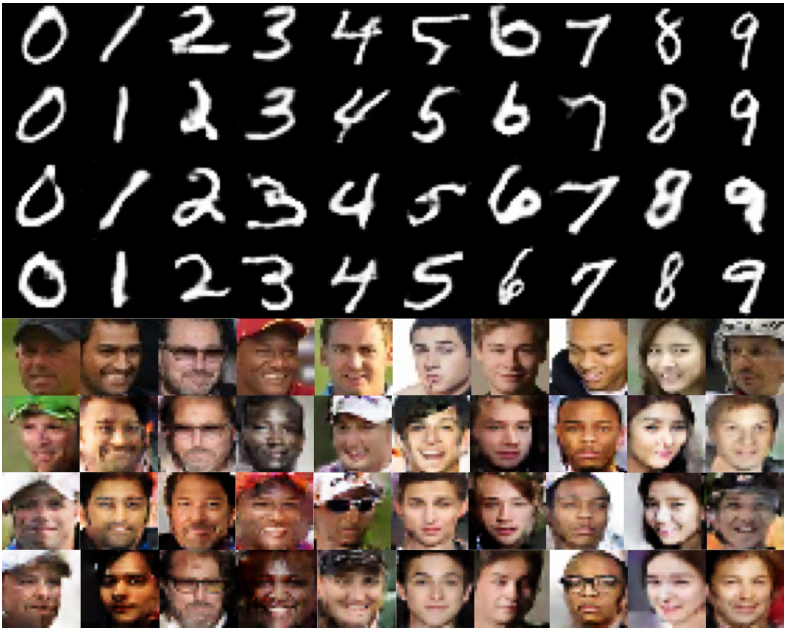


Figure 1: The top 4 rows contain the real digits from MNIST [22] handwritten digit dataset and the generated digits by the proposed CDA-GAN. The bottom 4 rows contain the real faces from NIST IJB-A [19] face recognition dataset and corresponding generated faces by CDA-GAN. We invite readers to guess which specific rows of these two types of data are generated by CDA-GAN or from real data. Please find the answer under Sec. 4.

Generative **A**dversarial **N**etworks (GANs) are derivatives from game-theoretic formulation and first introduced by [14] for image synthesis. GAN is a ground-breaking and distinctive method for generative models, which significantly helps address the problem of modeling data with intractable probability distribution $p(x)$. Different from traditional Markov-based models [37] [30] which typically has inevitable complexity during training, evolving lots of states or components to achieve satisfactory performance, GANs can be effortlessly optimized via the **B**ack-**P**ropagation (BP) algorithm.

Recent progress on GAN-based methods (*e.g.*, conditional GAN [28], semi-supervised GAN [29], BE-GAN [3], and TP-GAN [17]) have been proven to be effective for image synthesis with preserved annotation information, which significantly benefits modern data driven deep learning techniques [2] [11] [31]. The generative methods based on GANs have spot light on many higher-level computer vision applications, such as image generation [35] [32], image-to-image translation [18] [13], semantic image inpainting [36] and semantic image segmentation [23].

However, GANs still face the difficulties on stabilizing and constraining the optimization process to achieve satisfactory results. Thus, naively learning **D**eep **C**onvolutional **N**eural

Networks (DCNNs) models from such augmented data in specific domains (*e.g.*, object classification, face recognition, *etc.*) may even hurt the final performance.

In order to effectively address the above-mentioned challenge and narrow the gap between synthetic and real images while preserving the annotation information simultaneously, we propose a novel image generation method by extending the state-of-the-art GAN framework with dual agents for the Discriminator, focusing on adding realism and preserving annotation information for the generated images, respectively. The Generator is conditioned on the desired image features instead of a one-dimensional annotation label to gain more flexibility in terms of scalability for learning specific and robust DCNNs. Thus, we term our proposed method as **C**onditional **D**ual-**A**gent **GAN** (CDA-GAN).

In particular, a desired image is selected based on the priori knowledge and actual demand, which is forwarded through a pre-trained CNN sharing same network architecture with the Discriminator yet different weights to learn the corresponding deep features as the condition to the Generator. This condition is then concatenated with a random noise as the input of the Generator to synthesize photorealistic images. The Discriminator contains dual agents for distinguishing "real" or "fake" and identifying annotation information for the generated images. The dual agents are initialized with a standard Discriminator and a multi-class classifier with shared weights. The proposed CDA-GAN can be optimized via a standard training procedure by a combination of an annotation perception loss for annotation preserving, an adversarial loss for realism preserving and artifact repelling, and a regularization term for fast and stable convergence. The generated images present photorealistic quality with well preserved annotation information, which are used as augmented data together with real images for robust deep feature learning.

Experimental results demonstrate that our method not only presents compelling perceptual results but also significantly outperforms state-of-the-arts on the MNIST [22] hand written digits classification dataset and **N**ational **I**nstitute of **S**tandards and **T**echnology (NIST) **I**ARPA **J**anus **B**enchmark **A** (IJB-A) [19] face recognition dataset.

Our contributions are summarized as follows.

- We propose a novel **C**onditional **D**ual-**A**gent **G**enerative **A**dversarial **N**etwork (CDA-GAN) for photorealistic and annotation preserving image synthesis.
- The proposed dual-agent architecture effectively combines priori knowledge from data distribution (adversarial training) and domain knowledge of annotations (annotation perception) to exactly synthesize images in the 2D space.
- We impose a regularization term to train the conditioned dual-agent model that balances the power of the multiple Discriminators against the Generator for fast and stable convergence.
- We present qualitative and quantitative experiments showing the possibility of a "recognition via generation" framework and achieve the top performance on the the MNIST [22] hand written digits classification dataset and NIST IJB-A [19] face recognition dataset without extra human annotation efforts by training **D**eep **C**onvolutional **N**eural **N**etworks (DCNNs) on the generated images together with real images.

To facilitate future research, we will release both source code and trained models for our CDA-GAN upon acceptance.

2 Related Works

The vanilla GAN framework, first introduced in [14], consists of two sub-networks, Generator and Discriminator. The Generator is pitted against an adversary: a Discriminator that learns

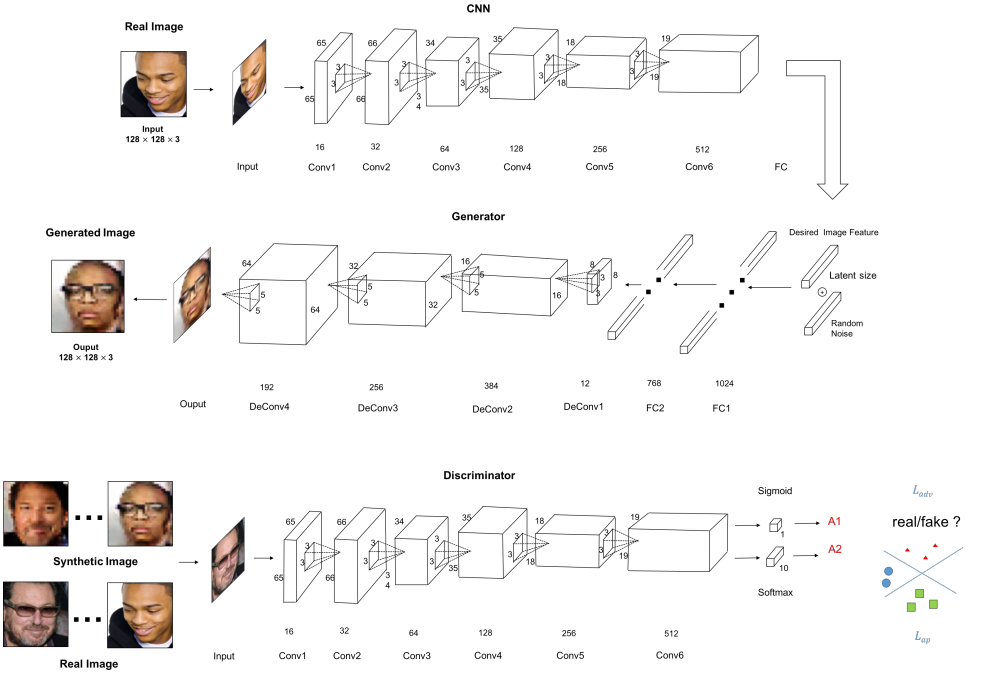


Figure 2: Overview of the proposed CDA-GAN architecture. The CNN module automatically extracts the desired image features as the condition to the Generator, which is further concatenated with a random noise as the input of the Generator to synthesize photorealistic images. The Discriminator contains dual agents (A_1, A_2) for distinguishing "real" or "fake" and identifying annotation information for the generated images. CDA-GAN can be optimized in an end-to-end way by minimizing a specially designed combination of adversarial loss and annotation perception loss.

to determine whether a sample is from the model distribution or the data distribution. The Generator can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the Discriminator is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles. Mirza and Osindero [28] introduce the conditional version of GAN, to condition on to both the Generator and Discriminator for effective image tagging. Berthelot et al. [3] propose a new **B**oundary **E**quilibrium **GAN** (BE-GAN) framework paired with a loss derived from the Wasserstein distance for training auto-encoder based GAN, which derives a way of controlling the trade-off between image diversity and visual quality. These successful applications of GAN motivate us to develop photorealistic and annotation preserving image synthesis method based on GAN.

However, in the real-world scenario, good police should not only detect fake articles, but also distinguish different categories, *e.g.*, different denominations. This in turn encourages the Generator to produce natural looking counterfeits with well preserved category information that are not only indistinguishable from the genuine articles but also applicable to actual demands. Inspired by this process, we propose a novel **C**onditional **D**ual-**A**gent **GAN** (CDA-GAN) for photorealistic and annotation preserving image synthesis, where the dual

agents focus on discriminating the realism of synthetic images using unlabeled real data and perceiving the annotation information, respectively. During the training process, the Generator is conditioned on the desired image features learned by a pre-trained CNN sharing the same architecture with the Discriminator yet different weights. Thus, CDA-GAN is flexible in terms of the scalability and able to generate photorealistic image with well preserved class labeling information. Such synthetic data can be effortlessly injected into limited real annotated data for learning DCNNs in specific domains. This separates us well with previous GAN-based attempts.

3 Conditional Dual-Agent GANs

Our proposed CDA-GAN is able to 1) generate photorealistic and annotation preserving images with fast and stable training, and 2) achieve the state-of-the-art performance for object classification and face recognition by learning deep invariant features through a "recognition via generation" framework without extra human annotation. As shown in Figure. 2, the CNN module automatically extracts the desired image features as the condition to the Generator, which is further concatenated with a random noise as the input of the Generator to synthesize photorealistic images. The Discriminator contains dual agents for distinguishing "real" or "fake" and identifying annotation information for the generated images. CDA-GAN can be optimized in an end-to-end way by minimizing a specially designed combination of adversarial loss and annotation perception loss. The convergence is fast and stable by imposing a regularization term for balancing the power of the Discriminator against the Generator. We now present each component in detail.

3.1 CNN Module

CNNs have achieved very good performance on various computer vision tasks, from object classification [20] [9] to face recognition [21] [27]. The performance gain roots in the multiple layered model and a large amount of available training data. The layered architecture enables the model to extract high level visual patterns for describing the visual properties of images and a large number of training data provide supervision for optimizing the huge number of inherent parameters. In order to achieve flexibility in terms of scalability, generalization capacity, and robustness, we propose to use a CNN module to learn the desired image features as the condition to the Generator. The CNN module share the same network architecture with the Discriminator yet different weights, and it is pre-trained on a large-scale dataset (*i.e.*, ImageNet [12]) as an initialization.

More formally, let the CNN module be denoted by F_{θ_1} (θ_1 are the learnable parameters of F), the learned deep features be denoted by f, where $f \in \mathbb{R}^m$ and m is defined as the latent size, and the real image with the desired label be denoted by x, then

$$f := F_{\theta_1}(x). \tag{1}$$

3.2 Generator

In order to achieve flexibility in terms of scalability, generalization capacity and robustness, and generate photorealistic and annotation preserving images which are truly beneficial for DCNNs learning in specific domains, we concatenate the desired image features (*i.e.*, condition) with a random noise with the latent size m as the input to the Generator.

More formally, let the random noise be denoted by z, where $z \in \mathbb{R}^m$, and the synthetic image be denoted by $\tilde{x}$, then

$$\tilde{x} := G_{\theta_2}(f + z), \tag{2}$$

where G represents the Generator, θ_2 are the learnable parameters of G, $+$ represents the concat operation and $f+z \in \mathbb{R}^{2m}$.

The key requirements for CDA-GAN are that the synthetic image $\tilde{x}$ should look like a real image in appearance while preserving the intrinsic annotation information from the desired image features.

To this end, we propose to learn θ_2 by minimizing a combination of following terms:

$$\mathcal{L}_{G_{\theta_2}} = (\mathcal{L}_{adv} + \lambda \mathcal{L}_{ap} - \delta I(f, G(z, f))) \tag{3}$$

where $\mathcal{L}_{adv}$ is the **adv**ersarial loss for adding realism to the synthetic images and alleviating artifacts, and $\mathcal{L}_{ap}$ is the **a**nnotation **p**erception loss for preserving the annotation information. The last term, $\delta I(\mathrm{f}, G(\mathrm{z},\mathrm{f}))$ is the regularization term using mutual information between the input noise z and the condition, learned deep features f to constraint the optimization of CDA-GAN for a fast and stable convergence (δ is a hyper-parameter for constraining).

The learning of the generator to transform from noise input z and conditional information, f, extracted image feature to synthesized output, $\tilde{x}$ or $G(z, f)$ is complicated and entangled inside the network structure. By maximizing the mutual information term and thus minimizing $\mathcal{L}_{G_{\theta_2}}$, we force the generator to learn from the conditional information and use extracted knowledge, f to directly apply in generating $G(z, f)$. This solution allows us to better control the use of input noise and conditional information under iterations. In practice, we would use the term $\mathcal{L}_{ap}(G, D_{\theta_3})$($D_{\theta_3}$ is the auxiliary classifier) to approximate the value of I(f, G(z,f)) [6].

To add realism to the synthetic images to really benefit DCNN performance in specific domains, we need to narrow the gap between the distributions of synthetic and real images. An ideal Generator will make it impossible to classify a given image as real or synthesized with high confidence. Meanwhile, preserving the annotation information is the essential and critical part for object classification and face recognition, *etc*. An ideal Generator will generate the images that have small intra-class distance and large inter-class distance in the feature space spanned by the DCNNs. These motivate the use of an adversarial Discriminator with dual agents.

3.3 Dual-Agent Discriminator

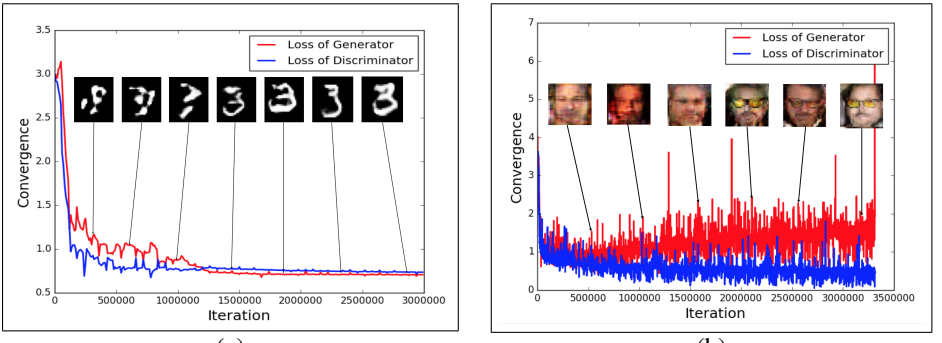
To incorporate the priori knowledge from the desired image distribution and domain knowledge of classes' distribution, we herein introduce a Discriminator with dual agents for distinguishing "real" *v.s.* "fake" and annotations simultaneously. To facilitate this process, we leverage a CNN as the Discriminator D_{θ_3} to be as simple as possible to avoid typical GAN tricks, which projects the input real / fake image into high-dimensional feature space through several **Conv**olution (Conv) and **F**ully **C**onnected (FC) layers, as shown in Figure. 2. θ_3 are the learnable parameters of the Discriminator.

One agent of D_{θ_3} is trained with $\mathcal{L}_{adv}$ to minimize the standard dual-class cross-entropy loss based on the output from the bottleneck layer of D_{θ_3},

$$\mathcal{L}_{adv} = \mathbb{E}_{x \sim p_{real}(x)}[log D_{\theta_3}(x)] + \mathbb{E}_{(f+z) \sim p_{syn}(f+z)}[log(1 - D_{\theta_3}(G_{\theta_2}(f+z)))]. \tag{4}$$

$\mathcal{L}_{adv}$ serves as a supervision to push the synthetic image to reside in the manifold of real images. It can prevent the blurry effect, alleviate artifacts and produce visually pleasing results.

The other agent of D_{θ_3} is trained with $\mathcal{L}_{ap}$ to preserve the annotation discriminability of the synthetic images. Specially, we define $\mathcal{L}_{ap}$ with the multi-class cross-entropy loss based



(a) (b)

Figure 3: Quality of synthetic results *w.r.t.* the network convergence measurement on MNIST [22] (a) and IJB-A [19] (b).

on another branch of output from the bottleneck layer of D_{θ_3}.

$$\begin{aligned}\mathcal{L}_{ap} = &\frac{1}{N}\sum_j(-(y_j log(D_{\theta_3}(x_j))) + (1-y_j log(1-D_{\theta_3}(x_j))))\\ &+\frac{1}{N}\sum_i((-y_i log(D_{\theta_3}(\tilde{x}_i))) + (1-y_i log(1-D_{\theta_3}(\tilde{x}_i)))),\end{aligned} \tag{5}$$

where y is the class ground truth, it is exactly the same as that of desired image.

Thus, minimizing $\mathcal{L}_{ap}$ would encourage deep features of the synthetic images belonging to the same class to be close to each other. If one visualizes the learned deep features in the high-dimensional space, the learned deep features of synthetic image set form several compact clusters and each cluster may be far away from others. Each cluster has a small variance. In this way, the synthetic images are enforced with well preserved annotation information. We also conduct experiments for illustration.

Using $\mathcal{L}_{ap}$ alone makes the results prone to annoying artifacts, because the search for a local minimum of $\mathcal{L}_{ap}$ may go through a path that resides outside the manifold of natural images. Thus, we combine $\mathcal{L}_{ap}$ with $\mathcal{L}_{adv}$ as the final objective function for D_{θ_3} to ensure that the search resides in that manifold and produces photorealistic and annotation preserving image:

$$\mathcal{L}_{D_{\theta_3}} = \mathcal{L}_{adv} + \lambda \mathcal{L}_{ap}. \tag{6}$$

3.4 Optimization

The goal of CDA-GAN is to use a set of desired real images x to learn a generator G_{θ_2} that adaptively generate photorealistic images $\tilde{x}$ with well preserved annotation information. We optimize CDA-GAN by alternatively optimizing D_{θ_3} and G_{θ_2} for each training iteration. We measure the convergence of CDA-GAN by using the above loss functions for D_{θ_3} and G_{θ_2}, respectively. Such measurement can be used to determine when the network has reached its final state or if the model has collapsed. We also conduct experiments for illustration.

4 Experiments

4.1 Experimental Settings

Benchmark Dataset: Except for synthesizing natural looking images, the proposed CDA-GAN also aims to generate annotation preserving images for accurate object- / face-centric

Operation	Kernel	Strides	Feature maps	Dropout	Reshape	Nonlinearity
$G_{\theta_2}(f+z)-2\times100$ input						
Linear	N/A	N/A	1024	N/A	N/A	ReLU
Linear	N/A	N/A	6272	N/A	(128, 7, 7)	ReLU
Transposed Convolution	5×5	2×2	256	N/A	N/A	ReLU
Transposed Convolution	5×5	2×2	128	N/A	N/A	ReLU
Convolution	2×2	N/A	3	N/A	N/A	Tanh
$D_{\theta_3}(x)-28\times28\times3$ input						
Convolution	3×3	2×2	32	0.3	N/A	Leaky ReLU
Convolution	3×3	1×1	64	0.3	N/A	Leaky ReLU
Convolution	3×3	2×2	128	0.3	N/A	Leaky ReLU
Convolution	3×3	1×1	256	0.3	N/A	Leaky ReLU
Linear	N/A	N/A	1	0.0	N/A	Sigmoid
Linear	N/A	N/A	10	0.0	N/A	Softmax
$G_{\theta_2}(f+z)-2\times100$ input						
Linear	N/A	N/A	1024	N/A	N/A	ReLU
Linear	N/A	N/A	768	N/A	(12, 8, 8)	ReLU
Transposed Convolution	5×5	2×2	384	N/A	N/A	ReLU
Transposed Convolution	5×5	2×2	256	N/A	N/A	ReLU
Transposed Convolution	5×5	2×2	192	N/A	N/A	ReLU
Transposed Convolution	5×5	2×2	3	N/A	N/A	Tanh
$D_{\theta_3}(x)-128\times128\times3$ input						
Convolution	3×3	2×2	16	0.3	N/A	Leaky ReLU
Convolution	3×3	1×1	32	0.3	N/A	Leaky ReLU
Convolution	3×3	2×2	64	0.3	N/A	Leaky ReLU
Convolution	3×3	1×1	128	0.3	N/A	Leaky ReLU
Convolution	3×3	2×2	256	0.3	N/A	Leaky ReLU
Convolution	3×3	1×1	512	0.3	N/A	Leaky ReLU
Linear	N/A	N/A	1	0.0	N/A	Sigmoid
Linear	N/A	N/A	10	0.0	N/A	Softmax
Latent size	100					
Generator Optimizer	Adam ($\alpha=0.00002$, $\beta_1=0.5$), $\lambda=0.25$					
Discriminator Optimizer	Adam ($\alpha=0.00002$, $\beta_1=0.5$), $\lambda=0.25$					
Batch size	10 for MNIST [22], 2 for IJB-A [19]					
Iterations	30000 for MNIST [22], 350000 for IJB-A [19]					

Table 1: CDA-GAN network architecture.

analysis with state-of-the-art DCNNs. Therefore, we evaluate the possibility of "recognition via generation" of CDA-GAN on the MNIST [22] hand written digits classification dataset and the challenging unconstrained face recognition dataset IJB-A [19].

The MNIST [22] handwritten digit dataset contains 60K training samples and 10K testing samples. The digits have been size-normalized and centered in a fixed-size (*i.e.*, 28×28) image.

IJB-A [19] contains both images and video frames from 500 subjects with 5,397 images and 2,042 videos that are split into 20,412 frames, 11.4 images and 4.2 videos per subject, captured from in-the-wild environment to avoid the near frontal bias, along with protocols for evaluation of both *verification* (1:1 comparison) and *identification* (1:N search) tasks. For training and testing, 10 random splits are provided by each protocol, respectively.

Implementation Details: Full details on the proposed CDA-GAN network architectures and training procedures for MNIST [22] and IJB-A [19] are summarized in Table.1 upper and lower panels, respectively. The hyper-parameters (such as α, β_1 and λ) are selected through cross-validated experiments.

Reproducibility: The proposed method is implementated by extending the Keras framework [7]. All networks are trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory for each. The source code and trained models for our CDA-GAN will be released upon acceptance.

4.2 Results and Discussions

Qualitative Results: In order to illustrate the compelling perceptual results generated by the proposed CDA-GAN, we first visualize the quality of refined results *w.r.t.* the network convergence measurement on MNIST [22] and IJB-A [19], as shown in Figure. 3. As can be seen, our CDA-GAN ensures a fast yet stable convergence through the carefully designed optimization scheme. The network convergence measurement correlates well with image fidelity. The proposed CDA-GAN generalize well from modelling hand written digits to human face data with enough details, diverse modalities and minimum artifacts, despite that

Method	verification			identification			
	TAR @ FAR=0.10 (%)	TAR @ FAR=0.01 (%)	TAR @ FAR=0.001 (%)	FNIR @ FPIR=0.10 (%)	FNIR @ FPIR=0.01 (%)	Rank1 (%)	Rank5 (%)
OpenBR [24]	43.30±0.006	23.60±0.009	10.40±0.014	85.10±0.028	93.40±0.017	24.60±0.011	37.50±0.008
GOTS [24]	62.70±0.012	40.60±0.014	19.80±0.008	76.50±0.033	95.30±0.024	44.30±0.021	59.50±0.020
BCNNs [8]	-	-	-	65.90±0.032	85.70±0.024	55.80±0.020	79.60±0.017
Pooling Faces [24]	81.90	63.10	-	-	-	84.60	93.30
Sankaranarayanan *et al.* [26]	94.50±0.005	79.00±0.010	59.00±0.020	-	-	88.00±0.010	95.00±0.005
PAMs [24]	-	82.60±0.018	65.20±0.037	-	-	84.00±0.012	92.50±0.008
Deep Multi-Pose [1]	95.40	87.60	-	**25.00**	**48.00**	84.60	92.70
Chellappa *et al.* [2]	96.34±0.005	83.10±0.035	-	-	-	89.90±0.011	**97.00**±0.075
DCNN [1]	**96.70**±0.009	83.80±0.042	-	-	-	90.30±0.012	96.50±0.008
Masi *et al.* [20]	-	**88.60**	**72.50**	-	-	**90.60**	96.20
CDA-GAN	96.81±0.009	89.06±0.014	73.12±0.031	22.63±0.011	41.97±0.045	91.02±0.010	96.73±0.008
CDA-GAN$_{more}$	97.83±0.005	90.76±0.013	75.64±0.021	19.81±0.009	36.96±0.042	92.39±0.008	97.21±0.005

Table 2: Performance comparison of CDA-GAN "recognition via generation" framework with state-of-the-arts on IJB-A [19] verification and identification protocols. For FNIR metric, a lower number means better performance. For the other metrics, a higher number means better performance. We use the modern ResNeXt-50 [34] as the backbone model with template adapted **S**upport **V**ector **M**athine (SVM) [10] as metric learning. We report CDA-GAN w/o or w/ (denoted by the subscript "*more*") the synthetic faces as the augmented data. The results are averaged over 10 testing splits. Symbol "-" implies that the result is not reported for that method. Standard deviation is not available for some methods. The results offered by our proposed CDA-GAN model are highlighted in blue. The second best results inferior to our models are highlighted in bold.

the convergence for human face data is relatively slower due to the complexility and difficulty to model the challenging human face data distribution. A comparison between the final output from our CDA-GAN and the original data from both MNIST [22] and IJB-A [19] could also be found in Figure. 1. Only the 3_{rd}, 4_{th}, 5_{th} and 6_{th} rows are the original data and all the others are synthesized.

Moreover, to gain insights into the effectivenss of annotation preserving quality of our CDA-GAN, we further use t-SNE [24] to visualize the deep features of both real images and generated counterparts for both MNIST [22] and IJB-A [19] by the dual-agent Discriminator of CDA-GAN on a 2D space in Figure. 4. As can be seen, the synthetic images present small intra-class distance and large inter-class distance, which is similar to (even better than) those of real images. This reveals that CDA-GAN ensures well preserved annotation information with the auxiliary agent for $\mathcal{L}_{ap}$.

Quantitative Results: We conduct classification on MNIST [22] with two different settings, directly using our discriminator network on original data (baseline: CDA-GAN for short) and augmenting generated handwritten digits to training data (CDA-GAN$_{more}$ for short). Based on our experiments, we observe that our baseline method CDA-GAN outperforms all the publicly accepted algorithm records by test error rate of 0.0073%. However, CDA-GAN$_{more}$ even performs better with test error rate, 0.0069% which proves that our augmented data could help to increase the accuracy for other potential classification tasks.

Furthermore, we also conduct unconstrained face recognition (*i.e.*, verification and identification) on IJB-A [19] with two different settings. In the two settings, the "recognition via generation" frameworks are respectively ResNeXt-50 [34] with template adapted SVM [10] pre-trained on MS-Celeb-1M [15] and fine-tuned on the original training data of each split without extra data (baseline: CDA-GAN for short), the original training data of each split with extra synthetic faces (our method: "recognition via generation" framework based on CDA-GAN, CDA-GAN$_{more}$ for short). The performance comparison of CDA-GAN with the two settings and other state-of-the-arts on IJB-A [19] are given in Table. 2. We can observe that with the injection of photorealistic and annotation preserving faces generated by CDA-GAN without extra human annotation efforts, our method CDA-GAN$_{more}$ outperforms

the baseline method CDA-GAN by 2.52% for TAR @ FAR=0.001 of verification and 5.01% for FNIR @ FPIR=0.01, 1.37% for Rank-1 of identification. Our method CDA-GAN$_{more}$ also outperforms the 2*nd*-best method [26] by 3.14% for TAR @ FAR=0.001 of verification and 1.79% for Rank-1 of identification. This well verified the promising potential of synthetic faces by our CDA-GAN on the challenging unconstrained face recognition problem.

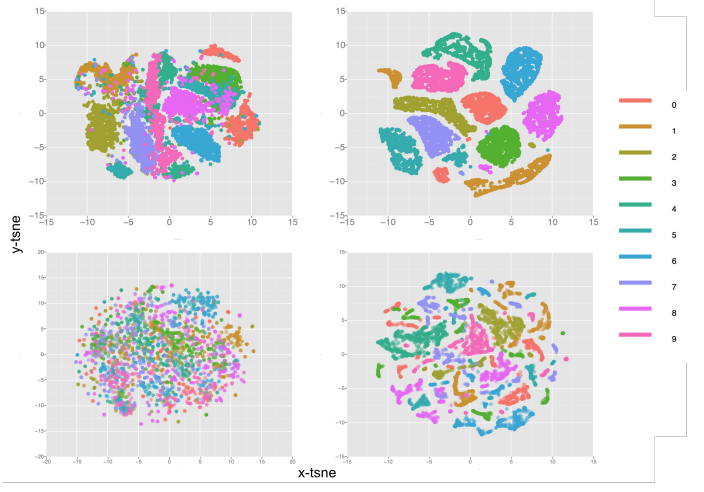


Figure 4: Feature space of the real hand written digits (top left), synthetic counterparts (top right), real human faces (bottom left), and synthetic counterparts (bottom right) by the dual-agent Discriminator of CDA-GAN using t-SNE [24], from the training data of MNIST [22] and IJB-A [19] split1. Each visually colored cluster shows a distinct class. Best viewed in color.

5 Conclusion

We propose a novel **C**onditional **D**ual-**A**gent **G**enerative **A**dversarial **N**etwork (CDA-GAN) for photorealistic and annotation preserving image synthesis. CDA-GAN effectively combines priori knowledge from data distribution (adversarial training) and domain knowledge of classes (annotation perception) to exactly synthesize images in the 2D space. CDA-GAN can be optimized in a fast yet stable way with the carefully designed loss functions and optimization strategy. One promising potential application of the proposed CDA-GAN is for solving transfer learning problems more effectively. Qualitative and quantitative experiments verify the possibility of our "recognition via generation" framework, which achieved the top performance on the challenging NIST IJB-A unconstrained face recognition benchmark without extra human annotation efforts.

Acknowledgement

The work of Jian Zhao was partially supported by **C**hina **S**cholarship **C**ouncil (CSC) grant 201503170248.

The work of Jiashi Feng was partially supported by National University of Singapore startup grant R-263-000-C08-133 and Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112.

References

[1] Wael AbdAlmageed, Yue Wu, Stephen Rawls, Shai Harel, Tal Hassner, Iacopo Masi, Jongmoo Choi, Jatuporn Lekust, Jungyeon Kim, Prem Natarajan, et al. Face recognition using deep multi-pose representations. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[2] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. *arXiv preprint arXiv:1702.01983*, 2017.

[3] David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[4] Rama Chellappa, Jun-Cheng Chen, Rajeev Ranjan, Swami Sankaranarayanan, Amit Kumar, Vishal M Patel, and Carlos D Castillo. Towards the design of an end-to-end automated system for image and video-based recognition. *arXiv preprint arXiv:1601.07883*, 2016.

[5] Jun-Cheng Chen, Vishal M Patel, and Rama Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[7] François Chollet. keras. https://github.com/fchollet/keras, 2015.

[8] Aruni Roy Chowdhury, Tsung-Yu Lin, Subhransu Maji, and Erik Learned-Miller. One-to-many face recognition with bilinear cnns. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[9] Dan C Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1237. Barcelona, Spain, 2011.

[10] Nate Crosswhite, Jeffrey Byrne, Omkar M Parkhi, Chris Stauffer, Qiong Cao, and Andrew Zisserman. Template adaptation for face verification and identification. *arXiv preprint arXiv:1603.03958*, 2016.

[11] Luke de Oliveira, Michela Paganini, and Benjamin Nachman. Learning particle physics by example: Location-aware generative adversarial networks for physics synthesis. *arXiv preprint arXiv:1701.05927*, 2017.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[13] Hao Dong, Paarth Neekhara, Chao Wu, and Yike Guo. Unsupervised image-to-image translation with generative adversarial networks. *arXiv preprint arXiv:1701.02676*, 2017.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[15] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision*. Springer, 2016.

[16] Tal Hassner, Iacopo Masi, Jungyeon Kim, Jongmoo Choi, Shai Harel, Prem Natarajan, and Gérard Medioni. Pooling faces: template based face recognition with pooled face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 59–67, 2016.

[17] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[19] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939. IEEE, 2015.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[21] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.

[22] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[23] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[25] Iacopo Masi, Stephen Rawls, Gerard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4838–4846, 2016.

[26] Iacopo Masi, Anh Tuan Tran, Jatuporn Toy Leksut, Tal Hassner, and Gerard Medioni. Do we really need to collect millions of faces for effective face recognition? *arXiv preprint arXiv:1603.07057*, 2016.

[27] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5):555–559, 2003.

[28] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[29] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[30] Simon Osindero and Geoffrey E Hinton. Modeling image patches with a directed hierarchy of markov random fields. In *Advances in neural information processing systems*, pages 1121–1128, 2008.

[31] Siamak Ravanbakhsh, Francois Lanusse, Rachel Mandelbaum, Jeff G Schneider, and Barnabas Poczos. Enabling dark energy science with deep generative models of galaxy images. In *AAAI*, pages 1488–1494, 2017.

[32] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[33] Swami Sankaranarayanan, Azadeh Alavi, and Rama Chellappa. Triplet similarity embedding for face verification. *arXiv preprint arXiv:1602.03418*, 2016.

[34] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[35] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. *arXiv preprint arXiv:1512.00570*, 2015.

[36] Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016.

[37] Xiaojin Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2(3):4, 2006.