CR-GAN: Learning Complete Representations for Multi-view Generation

Yu Tian[1], **Xi Peng**[1], **Long Zhao**[1], **Shaoting Zhang**[2] and **Dimitris N. Metaxas**[1]
[1] Rutgers University
[2] University of North Carolina at Charlotte
{yt219, px13, lz311, dnm}@cs.rutgers.edu, szhang16@uncc.edu

Abstract

Generating multi-view images from a single-view input is an essential yet challenging problem. It has broad applications in vision, graphics, and robotics. Our study indicates that the widely-used generative adversarial network (GAN) may learn "incomplete" representations due to the single-pathway framework: an encoder-decoder network followed by a discriminator network. We propose CR-GAN to address this problem. In addition to the single reconstruction path, we introduce a generation sideway to maintain the completeness of the learned embedding space. The two learning pathways collaborate and compete in a parameter-sharing manner, yielding considerably improved generalization ability to "unseen" dataset. More importantly, the two-pathway framework makes it possible to combine both labeled and unlabeled data for self-supervised learning, which further enriches the embedding space for realistic generations. The experimental results prove that CR-GAN significantly outperforms state-of-the-art methods, especially when generating from "unseen" inputs in wild conditions. [1]

1 Introduction

Generating multi-view images from a single-view input is an interesting problem with broad applications in vision, graphics, and robotics. Yet, it is a challenging problem since 1) computers need to "imagine" what a given object would look like after a 3D rotation is applied; and 2) the multi-view generations should preserve the same "identity".

Generally speaking, previous solutions to this problem include model-driven synthesis [Blanz and Vetter, 1999], data-driven generation [Zhu *et al.*, 2014; Yan *et al.*, 2016], and a combination of the both [Zhu *et al.*, 2016; Rezende *et al.*, 2016]. Recently, generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] have shown impressive results in multi-view generation [Tran *et al.*, 2017; Zhao *et al.*, 2017].

These GAN-based methods usually have a single-pathway design: an encoder-decoder network is followed by a discriminator network. The encoder (E) maps input images into a latent space (Z), where the embeddings are first manipulated and then fed into the decoder (G) to generate novel views.

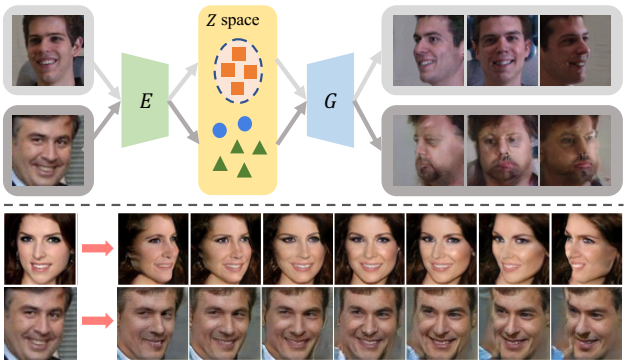


Figure 1: Top: The limitation of existing GAN-based methods. They can generate good results if the input is mapped into the learned subspace (Row 1). However, "unseen" data may be mapped out of the subspace, leading to poor results (Row 2). Bottom: Our results. By learning complete representations, the proposed CR-GAN can generate realistic, identity-preserved images from a single-view input.

However, our experiments indicate that this single-pathway design may have a severe issue: they can only learn "incomplete" representations, yielding limited generalization ability on "unseen" or unconstrained data. Take Fig. 1 as an example. During the training, the outputs of E constitute only a subspace of Z since we usually have a limited number of training samples. This would make G only "see" part of Z. During the testing, it is highly possible that E would map an "unseen" input outside the subspace. As a result, G may produce poor results due to the unexpected embedding.

To address this issue, we propose CR-GAN to learn *Complete Representations* for multi-view generation. The main idea is, in addition to the reconstruction path, we introduce another generation path to create view-specific images from embeddings that are randomly sampled from Z. Please refer to Fig. 2 for an illustration. The two paths share the same G. In other words, G learned in the generation path will guide the learning of both E and D in the reconstruction path, and vice versa. E is forced to be an inverse of G, yielding complete representations that would span the entire Z space. More importantly, the two-pathway learning can easily utilize both labeled and unlabeled data for self-supervised learning, which

[1] The code and pre-trained models are publicly available: https://github.com/bluer555/CR-GAN

can largely enrich the Z space for natural generations.

To sum up, we have following contributions:

- To the best of our knowledge, we are the first to investigate "complete representations" of GAN models;
- We propose CR-GAN that can learn "complete" representations, using a two-pathway learning scheme;
- CR-GAN can leverage unlabeled data for self-supervised learning, yielding improved generation quality;
- CR-GAN can generate high-quality multi-view images from even "unseen" dataset in wild conditions.

2 Related Work

Generative Adversarial Networks (GANs). Goodfellow *et al*. [Goodfellow *et al.*, 2014] introduced GAN to estimate target distribution via an adversarial process. Gulrajani *et al*. [Gulrajani *et al.*, 2017] presented a more stable approach to enforce *Lipschitz Constraint* on Wasserstein GAN [Arjovsky *et al.*, 2017]. AC-GAN *et al*. [Odena *et al.*, 2017] extended the discriminator by containing an auxiliary decoder network to estimate class labels for the training data. BiGANs [Donahue *et al.*, 2017; Dumoulin *et al.*, 2017] try to learn an inverse mapping to project data back into the latent space. Our method can also find an inverse mapping, make a balanced minimax game when training data is limited.

Multi-view Synthesis. Hinton *et al*. [Hinton *et al.*, 2011] introduced transforming auto-encoder to generate images with view variance. Yan *et al*. [Yan *et al.*, 2016] proposed Perspective Transformer Nets to find the projection transformation. Zhou *et al*. [Zhou *et al.*, 2016] propose to synthesize views by appearance flow. Very recently, GAN-based methods usually follow a single-pathway design: an encoder-decoder network [Peng *et al.*, 2016] followed by a discriminator network. For example, to normalize the viewpoint, *e.g.* face frontalization, they either combine encoder-decoder with 3DMM [Blanz and Vetter, 1999] parameters [Yin *et al.*, 2017], or use duplicates to predict global and local details [Huang *et al.*, 2017]. DR-GAN [Tran *et al.*, 2017] follows the single-pathway framework to learn identity features that are invariant to viewpoints. However, it may learn "incomplete" representations due to the single-pathway framework. In contrast, CR-GAN can learn complete representations using a two-pathway network, which guarantees high-quality generations even for "unseen" inputs.

Pose-Invariant Representation Learning. For representation learning [Li *et al.*, 2016; Fan *et al.*, 2016], early works may use *Canonical Correlation Analysis* to analyze the commonality among different pose subspaces [Hardoon *et al.*, 2004; Peng *et al.*, 2015]. Recently, deep learning based methods use synthesized images to disentangle pose and identity factors by cross-reconstruction [Zhu *et al.*, 2014; Peng *et al.*, 2017], or transfer information from pose variant inputs to a frontalized appearance [Zhu *et al.*, 2013]. However, they usually use only labeled data, leading to a limited performance. We proposed a two-pathway network to leverage both labeled and unlabeled data for self-supervised learning, which can generate realistic images in challenging conditions.

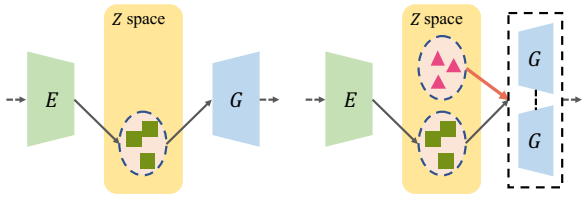


Figure 2: Left: Previous methods use a single path to learn the latent representation, but it is incomplete in the whole space. Right: We propose a two-pathway network combined with self-supervised learning, which can learn complete representations.

3 Proposed Method

3.1 A Toy Example of Incomplete Representations

A single-pathway network, *i.e.* an encoder-decoder network followed by a discriminator network, may have the issue of learning "incomplete" representations. As illustrated in Fig. 2 left, the encoder E and decoder G can "touch" only a subspace of Z since we usually have a limited number of training data. This would lead to a severe issue in testing when using "unseen" data as the input. It is highly possible that E may map the novel input out of the subspace, which inevitably leads to poor generations since G has never "seen" the embedding.

A toy example is used to explain this point. We use Multi-PIE [Gross *et al.*, 2010] to train a single-pathway network. As shown in the top of Fig. 1, the network can generate realistic results on Multi-PIE (the first row), as long as the input image is mapped into the learned subspace. However, when testing "unseen" images from IJB-A [Klare *et al.*, 2015], the network may produce unsatisfactory results (the second row). In this case, the new image is mapped out of the learned subspace.

This fact motivates us to train E and G that can "cover" the whole Z space, so we can learn complete representations. We achieve this goal by introducing a separate generation path, where the generator focuses on mapping the entire Z space to high-quality images. Fig. 2 illustrates the comparison between the single-pathway and two-pathway networks. Please refer to Fig. 3 (d) for an overview of our approach.

3.2 Generation Path

The generation path trains generator G and discriminator D. Here the encoder E is not involved since G tries to generate from random noise. Given a view label v and random noise $\mathbf{z}$, G aims to produce a realistic image $G(v, \mathbf{z})$ under view v. D is trying to distinguish real data from G's output, which minimizes:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathbb{P}_\mathbf{z}}[D_s(G(v,\mathbf{z}))] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}_\mathbf{x}}[D_s(\mathbf{x})] + \\ &\lambda_1 \mathop{\mathbb{E}}_{\hat{\mathbf{x}}\sim\mathbb{P}_{\hat{\mathbf{x}}}}[(\|\bigtriangledown_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2] - \lambda_2 \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}_\mathbf{x}}[P(D_v(\mathbf{x}) = v)], \end{aligned} \tag{1}$$

where $\mathbb{P}_\mathbf{x}$ is the data distribution and $\mathbb{P}_\mathbf{z}$ is the noise uniform distribution, $\mathbb{P}_{\hat{\mathbf{x}}}$ is an interpolation between pairs of points sampled from data distribution and the generator distribution [Gulrajani *et al.*, 2017]. G tries to fool D, it maximizes:

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathbb{P}_\mathbf{z}}[D_s(G(v,\mathbf{z}))] + \lambda_3 \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathbb{P}_\mathbf{z}}[P(D_v(G(v,\mathbf{z})) = v)], \tag{2}$$

where $(D_v(\cdot), D_s(\cdot)) = D(\cdot)$ denotes pairwise outputs of the discriminator. $D_v(\cdot)$ estimates the probability of being a

Algorithm 1: Supervised training with two paths

Input: Sets of view labeled images X, max number of steps T, and batch size m.
Output: Trained network E, G and D.
for $t = 1$ **to** T **do**
  for $i = 1$ **to** m **do**
    1. Sample $\mathbf{z} \sim P_{\mathbf{z}}$ and $\mathbf{x}_i \sim P_{\mathbf{x}}$ with v_i;
    2. $\bar{\mathbf{x}} \leftarrow G(v_i, \mathbf{z})$;
    3. Update D by Eq. 1, and G by Eq. 2;
    4. Sample $\mathbf{x}_j \sim P_{\mathbf{x}}$ with v_j (where $\mathbf{x}_j$ and $\mathbf{x}_i$ share the same identity);
    5. $(\bar{\mathbf{v}}, \bar{\mathbf{z}}) \leftarrow E(\mathbf{x}_i)$;
    6. $\tilde{\mathbf{x}}_j \leftarrow G(v_j, \bar{\mathbf{z}})$;
    7. Update D by Eq. 3, and E by Eq. 4;
  end
end

specific view, $D_s(\cdot)$ describes the image quality, *i.e.*, how real the image is. Note that in Eq. 1, D learns how to estimate the correct view of a real image [Odena *et al.*, 2017], while G tries to produce an image with that view in order to get a high score from D in Eq. 2.

3.3 Reconstruction Path

The reconstruction path trains E and D but keeping G fixed. E tries to reconstruct training samples, this would guarantee that E will be learned as an inverse of G, yielding complete representations in the latent embedding space.

The output of E should be identity-preserved so the multi-view images will present the same identity. We propose a cross reconstruction task to make E disentangle the viewpoint from the identity. More specifically, we sample a real image pair $(\mathbf{x}_i, \mathbf{x}_j)$ that share the same identity but different views v_i and v_j. The goal is to reconstruct x_j from x_i. To achieve this, E takes $\mathbf{x}_i$ as input and outputs an identity-preserved representation $\bar{\mathbf{z}}$ together with the view estimation $\bar{\mathbf{v}}$: $(\bar{\mathbf{v}}, \bar{\mathbf{z}}) = (E_v(\mathbf{x}_i), E_z(\mathbf{x}_i)) = E(\mathbf{x}_i)$. Note that $\bar{\mathbf{v}}$ is learned for further self-supervised training as shown in Sec. 3.4.

G takes $\bar{\mathbf{z}}$ and view v_j as the input. As $\bar{\mathbf{z}}$ is expected to carry the identity information of this person, with view v_j's help, G should produce $\tilde{\mathbf{x}}_j$, the reconstruction of $\mathbf{x}_j$. D is trained to distinguish the fake image $\tilde{\mathbf{x}}_j$ from the real one $\mathbf{x}_i$. Thus D minimizes:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{x}_i, \mathbf{x}_j \sim \mathbb{P}_{\mathbf{x}}} [D_s(\tilde{\mathbf{x}}_j) - D_s(\mathbf{x}_i)] + \\ &\lambda_1 \mathop{\mathbb{E}}_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}} [(\|\bigtriangledown_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2] - \lambda_2 \mathop{\mathbb{E}}_{\mathbf{x}_i \sim \mathbb{P}_x} [P(D_v(\mathbf{x}_i) = v_i)], \end{aligned} \tag{3}$$

where $\tilde{\mathbf{x}}_j = G(v_j, E_z(\mathbf{x}_i))$. E helps G to generate high quality image with view v_j, so E maximizes:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{x}_i, \mathbf{x}_j \sim \mathbb{P}_{\mathbf{x}}} [D_s(\tilde{\mathbf{x}}_j) + \lambda_3 P(D_v(\tilde{\mathbf{x}}_j) = v_j) - \\ &\lambda_4 L_1(\tilde{\mathbf{x}}_j, \mathbf{x}_j) - \lambda_5 L_v(E_v(\mathbf{x}_i), v_i)], \end{aligned} \tag{4}$$

where L_1 loss is utilized to enforce that $\tilde{x}_j$ is the reconstruction of x_j. L_v is the cross-entropy loss of estimated and ground truth views, to let E be a good view estimator.

The two-pathway network learns complete representations: First, in the generation path, G learns how to produce real images from *any* inputs in the latent space. Then, in the reconstruction path, G retains the generative ability since it keeps unchanged. The alternative training details of the two pathways are summarized in Algorithm 1.

Algorithm 2: Self-supervised training with two paths

Input: Sets of view labeled and unlabeled images X, max number of steps T, and batch size m.
Output: Trained network E, G and D.
Pre-train E, G and D according to Algorithm 1;
for $t = 1$ **to** T **do**
  for $i = 1$ **to** m **do**
    Sample $\mathbf{z} \sim P_{\mathbf{z}}$ and $\mathbf{x} \sim P_{\mathbf{x}}$;
    if $\mathbf{x}$ *is labeled* **then**
      1. $\mathbf{x}_i \leftarrow \mathbf{x}$;
      2. Get the label v_i of $\mathbf{x}_i$;
      3. Repeat the step 2 to 7 in Algorithm 1;
    else
      4. $(\bar{\mathbf{v}}, \bar{\mathbf{z}}) \leftarrow E(\mathbf{x})$;
      5. Compute $\hat{v}$ (the estimation of $\bar{\mathbf{v}}$);
      6. Update D by Eq. 5 and E by Eq. 6;
      7. Update D by Eq. 7 and G by Eq. 8;
    end
  end
end

3.4 Self-supervised Learning

Labeled datasets are usually limited and biased. For example, Multi-PIE [Gross *et al.*, 2010] is collected in a constrained setting, while large-pose images in 300wLP [Zhu *et al.*, 2016] are distorted. As a result, G would generate low-quality images since it has only "seen" poor and limited examples.

To solve this issue, we further improve the proposed CR-GAN with self-supervised learning. The key idea is to use a pre-trained model to estimate viewpoints for unlabeled images. Accordingly, we modify the supervised training algorithm into two phases. In the first stage, we pre-train the network on labeled data to let E be a good view estimator. In the second stage, both labeled and unlabeled data are utilized to boost G. When an unlabeled image $\mathbf{x}$ is fed to the network, a view estimation $\bar{\mathbf{v}}$ is obtained by $E_v(\cdot)$. Denote $\hat{v}$ to be the closest one-hot vector of $\bar{\mathbf{v}}$, in the reconstruction path, we let E minimize $L_v(\bar{\mathbf{v}}, \hat{v})$ and then reconstruct $\mathbf{x}$ to itself. Similar to Eq. 3, D minimizes:

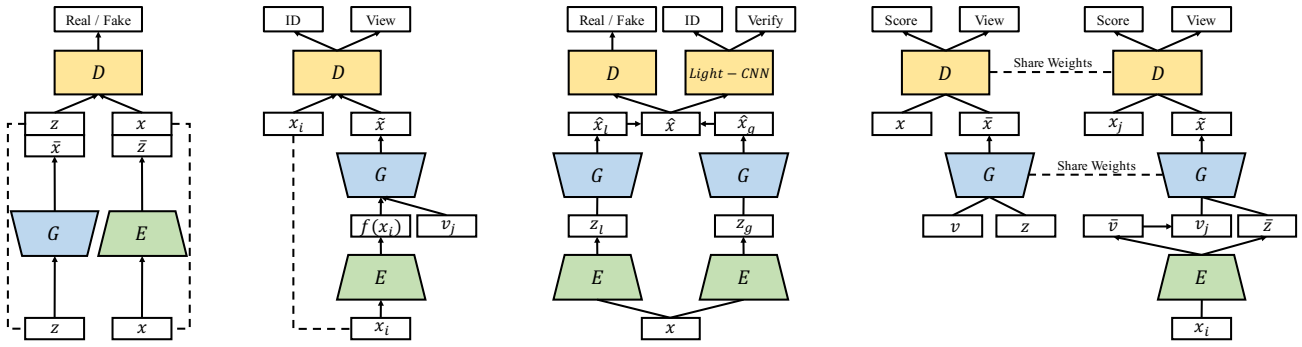
$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [D_s(G(\hat{v}, E_z(\mathbf{x}))) - D_s(\mathbf{x})] + \\ &\lambda_1 \mathop{\mathbb{E}}_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}} [(\|\bigtriangledown_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2] - \lambda_2 \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [P(D_v(\mathbf{x}) = \hat{v})], \end{aligned} \tag{5}$$

similar to Eq. 4, E maximizes:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [D_s(G(\hat{v}, E_z(\mathbf{x}))) + \lambda_3 P(D_v(G(\hat{v}, E_z(\mathbf{x}))) = \hat{v}) - \\ &\lambda_4 L_1(G(\hat{v}, E_z(\mathbf{x})), \mathbf{x}) - \lambda_5 L_v(E_v(\mathbf{x}), \hat{v})]. \end{aligned} \tag{6}$$

In the generation path, we let $\hat{v}$ be the ground truth of $\mathbf{x}$, and generate an image in view $\hat{v}$. So similar to Eq. 1, D minimizes:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{z} \sim \mathbb{P}_{\mathbf{z}}} [D_s(G(\hat{v}, \mathbf{z}))] - \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [D_s(\mathbf{x})] + \\ &\lambda_1 \mathop{\mathbb{E}}_{\hat{\mathbf{x}} \sim \mathbb{P}_{\hat{\mathbf{x}}}} [(\|\bigtriangledown_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2] - \lambda_2 \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}} [P(D_v(\mathbf{x}) = \hat{v})], \end{aligned} \tag{7}$$



(a) BiGAN (b) DR-GAN (c) TP-GAN (d) CR-GAN (Ours)

Figure 3: Comparison of BiGAN [Donahue *et al.*, 2017], DR-GAN [Tran *et al.*, 2017], TP-GAN [Huang *et al.*, 2017] and our method.

similar to Eq. 2, G maximizes:

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim\mathbb{P}_\mathbf{z}}[D_s(G(\hat{v},\mathbf{z}))] + \lambda_3 \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathbb{P}_\mathbf{z}}[P(D_v(G(\mathbf{z})) = \hat{v})]. \quad (8)$$

Once we get the pre-trained model, the encoder E predicts the probabilities of the input image belonging to different views. We choose the view with the highest probability as the estimation. Our strategy is similar to RANSAC algorithm [Fischler and Bolles, 1981], where we treat the estimations with higher confidence as "inliers" and use them to make view estimation more accurate. We summarize the self-supervised training in Algorithm 2.

Compared with the single-pathway solution, the proposed two-pathway network boosts the self-supervised learning in two aspects: 1) it provides a better pre-trained model for viewpoint estimation as a byproduct; and 2) it guarantees that we can take full advantage of unlabeled data in training since CR-GAN learns complete representations.

3.5 Discussion

To highlight the novelty of our method, we compare CR-GAN with the following three GANs. In Fig. 3, we show their network structures as well as ours for visual comparison.

BiGAN [Donahue *et al.*, 2017; Dumoulin *et al.*, 2017] jointly learns a generation network G and an inference network E. The authors proved that E is an inverse network of G. However, in practice, BiGAN produces poor reconstructions due to finite data and limited network capacity. Instead, CR-GAN uses explicit reconstruction loss to solve this issue.

DR-GAN [Tran *et al.*, 2017] also tries to learn an identity preserved representation to synthesize multi-view images. But we have two distinct differences. First, the output of its encoder, also acts as the decoder's input, completely depends on the training dataset. Therefore, it can not deal with new data. Instead, we use the generation path to make sure that the learning of our G is "complete". Second, we don't let D estimate the identity for training data, because we employ unlabeled dataset in self-supervised learning which has no identity information. The involvement of unlabeled dataset also makes our model more robust for "unseen" data.

TP-GAN [Huang *et al.*, 2017] uses two pathway GANs for frontal view synthesis. Their framework is different from ours: First, they use two distinct encoder-decoder networks, while CR-GAN shares all modules in the two pathways. Besides, they use two pathways to capture global features and local details, while we focus on learning complete representations in multi-view generation.

4 Experiments

CR-GAN aims to learn complete representations in the embedding space. We achieve this goal by combining the two-pathway architecture with self-supervised learning. We conduct experiments to evaluate these two contributions respectively. Then we compare our CR-GAN with DR-GAN [Tran *et al.*, 2017], both the visual results and t-SNE visualization in the embedding space are shown. We also compare CR-GAN and BiGAN with an image reconstruction task.

4.1 Experimental Settings

Datasets. We evaluate CR-GAN on datasets with and without view labels. Multi-PIE [Gross *et al.*, 2010] is a labeled dataset collected under constrained environment. We use 250 subjects from the first session with 9 poses within $\pm 60^\circ$, 20 illuminations, and two expressions. The first 200 subjects are for training and the rest 50 for testing. 300wLP [Zhu *et al.*, 2016] is augmented from 300W [Sagonas *et al.*, 2013] by the face profiling approach [Zhu *et al.*, 2016], which contains view labels as well. We employ images with yaw angles ranging from -60° to $+60^\circ$, and discretize them into 9 intervals.

For evaluation on unlabeled datasets, we use CelebA [Liu *et al.*, 2015] and IJB-A [Klare *et al.*, 2015]. CelebA contains a large amount of celebrity images with unbalanced viewpoint distributions. Thus, we collect a subset of 72,000 images from it, which uniformly ranging from -60° to $+60^\circ$. Notice that the view labels of the images in CelebA are only utilized to collect the subset, while no view or identity labels are employed in the training process. We also use IJB-A which contains 5,396 images for evaluation. This dataset is challenging, since there are extensive identity and pose variations.

Implementation Details. Our network implementation is modified from the residual networks in WGAN-GP [Gulrajani *et al.*, 2017], where E shares a similar network structure with D. During training, we set v to be a one-hot vector with 9 dimensions and $z \in [-1, 1]^{119}$ in the latent space. The batch size is 64. Adam optimizer [Kingma and Ba, 2015] is used with the learning rate of 0.0005 and momentum of $[0, 0.9]$. According to the setting of WGAN-GP, we let $\lambda_1 =$

Figure 4: Results generated by the single-pathway and two-pathway from (a) Multi-PIE [Gross *et al.*, 2010] and (b) IJB-A [Klare *et al.*, 2015]. In each case, the images generated by the two-pathway (Row 2) outperform the ones produced by the single-pathway (Row 1).

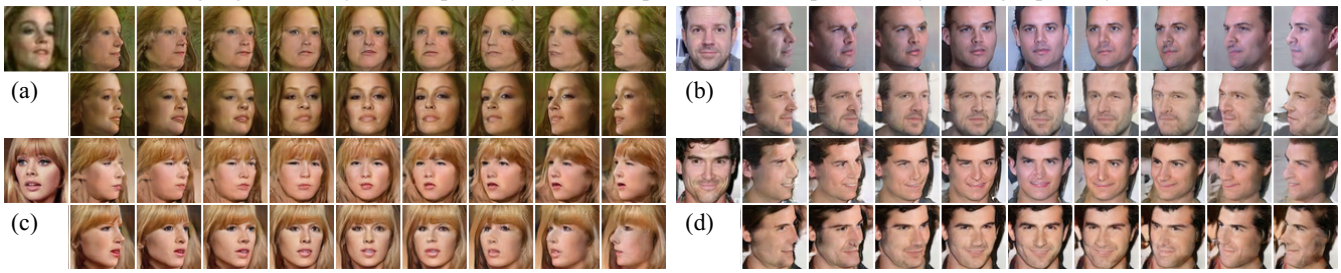


Figure 5: Multi-view face generation results on CelebA [Liu *et al.*, 2015]. In each case, self-supervised learning (Row 2) generates more realistic images than supervised learning (Row 1). Note that in (b) and (d), the beard and eyebrows are well-kept in different views.

10, $\lambda_2 \sim \lambda_4 = 1$, $\lambda_5 = 0.01$. Moreover, all the networks are trained after 25 epochs in supervised learning; we train 10 more epochs in self-supervised learning.

4.2 Single-pathway *vs.* Two-pathway

We compare two-pathway network with the one using a single reconstruction path. All the networks are trained on Multi-PIE. When test with Muli-PIE, as shown in Fig. 4 (a), both models produce desirable results. In each view, facial attributes like glasses are kept well. However, single-pathway model gets unsatisfactory results on IJB-A, which is an "unseen" dataset. As shown in Fig. 4 (b), two-pathway model consistently produce natural images with more details and fewer artifacts. Instead, the single-pathway model cannot generate images with good quality. This result prove that our two-pathway network handles "unseen" data well by learning a complete representation in the embedding space.

4.3 Supervised *vs.* Self-supervised Learning

The two-pathway network is employed in the following evaluations. We use Multi-PIE and 300wLP in supervised learning. For self-supervised learning, in addition to the above datasets, CelebA is employed as well. Note that we don't use view or identity labels in CelebA during training.

Evaluation on CelebA. Fig. 5 shows the results on CelebA. In Fig. 5 (a), although the supervised model generates favorable results, there are artifacts in all views. As the supervised model is only trained on Multi-PIE and 300wLP, it is difficult to "approximate" the data in the wild. Instead, the self-supervised model has learned a latent representation where richer features are embedded, so it generates more realistic results while the identities are well preserved. We can make the similar observation in Fig. 5 (b). The supervised model can only generate images that are similar to Multi-PIE, while the self-supervised model can generate novel identities. In Fig. 5 (c) and (d), the self-supervised model preserve identity and attributes in a better way than others.

Evaluation on IJB-A. Fig. 6 shows more results on IJB-A. We find that our self-supervised model successfully generalize what it has learned from CelebA to IJB-A. Note that it is our self-supervised learning approach that makes it possible to train the network on unlabeled datasets.

4.4 Comparison with DR-GAN

Furthermore, we compare our self-supervised CR-GAN with DR-GAN [Tran *et al.*, 2017]. We replace DC-GAN [Radford *et al.*, 2016] network architecture used in DR-GAN with WGAN-GP for a fair comparison.

Evaluation on IJB-A. We show the results of DR-GAN and CR-GAN in Fig. 6 respectively. DR-GAN produces sharp images, but the facial identities are not well-kept. By contrast, in Fig. 6 (a) and (b), CR-GAN produces face images with similar identities. In all cases, DR-GAN fails to produce high-quality images with large poses. Although not perfect enough, CR-GAN can synthesize reasonable profile images.

Identity Similarities on IJB-A. We generate 9 views for each image in IJB-A both using DR-GAN and CR-GAN. Then we obtain a 128-dim feature for each view by FaceNet [Schroff *et al.*, 2015]. We evaluate the identity similarities between the real and generated images by feeding them to FaceNet. The squared L2 distances of the features directly corresponding to the face similarity: faces of the same subjects have small distances, while faces of different subjects have large distances. Table 1 shows the results of the average L2 distance of CR-GAN and DR-GAN in different datasets. Our method outperforms DR-GAN on all datasets, especially on IJB-A which contains unseen data. Fig. 7 shows the t-SNE visualization in the embedding space of DR-GAN and CR-GAN respectively. For clarity, we only visualize 10 randomly selected subjects along with 9 generated views of each. Compared with DR-GAN, CR-GAN produces tighter clusterings: multi-view images of the same subject are embedded close to each other. It means the identities are better preserved.

Generative Ability. We utilize DR-GAN and CR-GAN to generate images from random noises. In Fig. 8, CR-GAN can produce images with different styles, while DR-GAN leads to blurry results. This is because the single-pathway generator of DR-GAN learns incomplete representations in the embedding

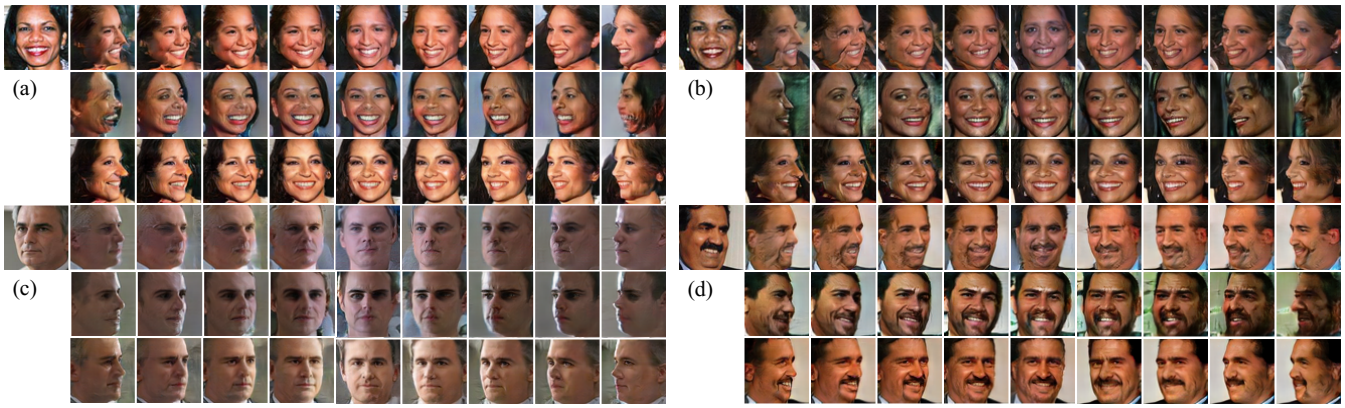


Figure 6: Multi-view face generation results on IJB-A [Klare *et al.*, 2015]. In each case, from top to bottom: the results generated by our supervised model, DR-GAN [Tran *et al.*, 2017] and our self-supervised model. DR-GAN fails to produce favourable images of large poses, while our method can synthesize reasonable profile images.

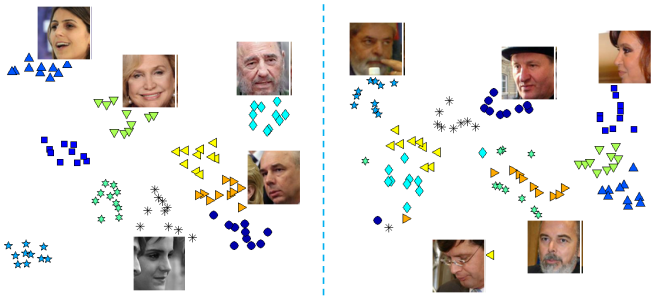


Figure 7: t-SNE visualization for the embedding space of CR-GAN (**left**) and DR-GAN (**right**), with 10 subjects from IJB-A [Klare *et al.*, 2015]. The same marker shape (color) indicates the same subject. For CR-GAN, multi-view images of the same subject are embedded close to each other, which means the identities are better preserved.

	Multi-PIE	CelebA	IJB-A
DR-GAN	1.073 ± 0.013	1.281 ± 0.007	1.295 ± 0.008
CR-GAN	$\mathbf{1.018 \pm 0.019}$	$\mathbf{1.214 \pm 0.009}$	$\mathbf{1.217 \pm 0.010}$

Table 1: Identity similarities between real and generated images.

space, which fails to handle random inputs. Instead, CR-GAN produces favorable results with complete embeddings.

4.5 Comparison with BiGAN

111 To compare our method with BiGAN, we qualitatively show the image reconstruction results of both methods on CelebA in Fig. 9. We can find that as demonstrated by [Donahue *et al.*, 2017; Dumoulin *et al.*, 2017], BiGAN cannot reconstruct the data correctly, while CR-GAN keeps identities well due to the explicit reconstruction loss we employed.

5 Conclusion

In this paper, we investigate learning "complete representations" of GAN models. We propose CR-GAN that uses a two-pathway framework to achieve the goal. Our method can leverage both labeled and unlabeled data for self-supervised learning, yielding high-quality multi-view image generations from even "unseen" data in wild conditions.

Figure 8: Generating multi-view images from the random noise. (a) DR-GAN [Tran *et al.*, 2017] generates blurry results and many artifacts. (b) CR-GAN generates realistic images of different styles.

Figure 9: Reconstruction results on CelebA. BiGAN (Row 2) cannot keep identity well. Ours (Row 3) produces better results.

Acknowledgements

This work is partly supported by the Air Force Office of Scientific Research (AFOSR) under the Dynamic Data-Driven Application Systems program, NSF CISE program, and NSF grant CCF 1733843.

References

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *ICML*, pages 214–223, 2017.

[Blanz and Vetter, 1999] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999.

[Donahue *et al.*, 2017] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017.

[Dumoulin *et al.*, 2017] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. In *ICLR*, 2017.

[Fan *et al.*, 2016] Miao Fan, Qiang Zhou, and Thomas Fang Zheng. Learning embedding representations for knowledge inference on imperfect and incomplete repositories. In *Web Intelligence (WI)*, pages 42–48, 2016.

[Fischler and Bolles, 1981] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

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

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

[Gulrajani *et al.*, 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In *NIPS*, pages 5769–5779, 2017.

[Hardoon *et al.*, 2004] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[Hinton *et al.*, 2011] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *ICANN*, pages 44–51, 2011.

[Huang *et al.*, 2017] 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. In *ICCV*, 2017.

[Kingma and Ba, 2015] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Klare *et al.*, 2015] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, pages 1931–1939, 2015.

[Li *et al.*, 2016] Yingming Li, Ming Yang, and Zhongfei Zhang. Multi-view representation learning: A survey from shallow methods to deep methods. *arXiv preprint arXiv:1610.01206*, 2016.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[Odena *et al.*, 2017] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017.

[Peng *et al.*, 2015] Xi Peng, Junzhou Huang, Qiong Hu, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. From circle to 3-sphere: Head pose estimation by instance parameterization. *Computer Vision and Image Understanding*, 136:92–102, 2015.

[Peng *et al.*, 2016] Xi Peng, Rogerio S Feris, Xiaoyu Wang, and Dimitris N Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *ECCV*, pages 38–56, 2016.

[Peng *et al.*, 2017] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017.

[Radford *et al.*, 2016] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[Rezende *et al.*, 2016] Danilo Jimenez Rezende, SM Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. In *NIPS*, pages 4996–5004, 2016.

[Sagonas *et al.*, 2013] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, pages 397–403, 2013.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet": A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[Tran *et al.*, 2017] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition. In *CVPR*, 2017.

[Yan *et al.*, 2016] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision. In *NIPS*, pages 1696–1704, 2016.

[Yin *et al.*, 2017] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.

[Zhao *et al.*, 2017] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, and Jiashi Feng. Multi-view image generation from a single-view. *arXiv preprint arXiv:1704.04886*, 2017.

[Zhou *et al.*, 2016] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, pages 286–301, 2016.

[Zhu *et al.*, 2013] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning identity-preserving face space. In *ICCV*, pages 113–120, 2013.

[Zhu *et al.*, 2014] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*, pages 217–225, 2014.

[Zhu *et al.*, 2016] X. Zhu, Z. Lei, X. Liu, H. Shi, and S.Z. Li. Face Alignment Across Large Poses: A 3D Solution. In *CVPR*, pages 146–155, 2016.