International Journal of Computer Vision (2020) 128:2436–2458
https://doi.org/10.1007/s11263-020-01311-4

Discriminator Feature-Based Inference by Recycling the Discriminator of GANs

Duhyeon Bang[1] · Seoungyoon Kang[1] · Hyunjung Shim[1]

Received: 29 April 2019 / Accepted: 19 February 2020 / Published online: 4 March 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
Generative adversarial networks (GANs) successfully generate high quality data by learning a mapping from a latent vector to the data. Various studies assert that the latent space of a GAN is semantically meaningful and can be utilized for advanced data analysis and manipulation. To analyze the real data in the latent space of a GAN, it is necessary to build an inference mapping from the data to the latent vector. This paper proposes an effective algorithm to accurately infer the latent vector by utilizing GAN discriminator features. Our primary goal is to increase inference mapping accuracy with minimal training overhead. Furthermore, using the proposed algorithm, we suggest a conditional image generation algorithm, namely a spatially conditioned GAN. Extensive evaluations confirmed that the proposed inference algorithm achieved more semantically accurate inference mapping than existing methods and can be successfully applied to advanced conditional image generation tasks.

Keywords Generative adversarial networks · Inference mapping · Conditional image generation · Quality metric for inference mapping · Spatial semantic manipulation

1 Introduction

Generative adversarial networks (GANs) have demonstrated remarkable progress in successfully reproducing real data distribution, particularly for natural images. Although GANs impose few constraints or assumptions on their model definition, they are capable of producing sharp and realistic images. To this end, training GANs involves adversarial competition between a generator and discriminator: the generator learns the generation process formulated by mapping from the latent distribution P_z to the data distribution P_{data}; and the discriminator evaluates the generation quality by distinguishing generated images from real images. Goodfellow et al. (2014) formulated the objective of this adversarial training using the minimax game

$$\min_{\mathrm{G}} \max_{\mathrm{D}} \mathop{\mathbb{E}}_{x\sim P_{\text{data}}} [\log(\mathrm{D}(x))] + \mathop{\mathbb{E}}_{z\sim P_z} [\log(1-\mathrm{D}(\mathrm{G}(z))], \quad (1)$$

where $\mathbb{E}$ denotes expectation; G and D are the generator and discriminator, respectively; and z and x are samples drawn from P_z and P_{data}, respectively. Once the generator learns the mapping from the latent vector to the data (i.e., $z \to x$), it is possible to generate arbitrary data corresponding to randomly drawn z. Inspired by this pioneering work, various GAN models have been developed to improve training stability, image quality, and diversity of the generation.

In addition to image generation, GAN models are an attractive tool for building interpretable, disentangled representations. Due to their semantic power, several studies (Radford et al. 2016; Berthelot et al. 2017) show that data augmentation or editing can be achieved by simple operations in the GAN latent space. To utilize the semantic representation derived by the GAN latent space, we need to establish inference mapping from the data to the latent vector (i.e., $x \to z$). Previous studies generally adopt acyclic or cyclic inference mapping approaches to address the inference problem.

Communicated by Jun-Yan Zhu, Hongsheng Li, Eli Shechtman, Ming-Yu Liu, Jan Kautz, Antonio Torralba.

Duhyeon Bang and Seoungyoon Kang have contributed equally to this work.

✉ Hyunjung Shim
kateshim@yonsei.ac.kr

Duhyeon Bang
duhyeonbang@yonsei.ac.kr

Seoungyoon Kang
sy.kang@yonsei.ac.kr

[1] School of Integrated Technology, Yonsei Institute of Convergence Technology, Yonsei University, Seoul, Republic of Korea

Acyclic inference models develop inference mapping $x \rightarrow z$ independently from generation mapping (i.e., GAN training). Consequently, learning this inference mapping can be formulated as minimizing image reconstruction error through latent optimization. Previous studies (Liu and Tuzel 2016; Berthelot et al. 2017) solve this optimization problem by finding an inverse generation mapping, $\mathrm{G}^{-1}(x)$, using a non-convex optimizer. However, calculating this inverse path suffers from multiple local minima due to the generator's non-linear and highly complex nature; thus it is difficult to reach the global optimum. In addition, the consequentially heavy computational load at runtime limits practical applications. To alleviate computational load at runtime, iGAN (Zhu et al. 2016) first proposed a hybrid approach, estimating from $x \rightarrow z^0$ and then $z^0 \rightarrow z$, where z^0 is the initial state for z. Specifically, iGAN predicted the initial latent vector for x using an encoder model ($x \rightarrow z^0$), then used it as the initial optimizer value to compute the final estimate z ($z^0 \rightarrow z$). Although the encoder model accelerates execution time for the testing phase, this initial estimate $x \rightarrow z^0$ is often inaccurate due to disadvantage of its encoder models, and consequential image reconstruction loss presents performance limitations that result in missing important attributes of the input data. Section 3.1 presents a detailed discussion of various inference models.

Cyclic inference models (Dumoulin et al. 2017; Donahue et al. 2017) consider bidirectional mapping, $x \leftrightarrow z$. That is to say, inference learning and generation mapping are considered simultaneously. In contrast to acyclic inference, cyclic inference aims to train the generator using feedback from inference mapping. For example, Dumoulin et al. (2017) and Donahue et al. (2017) develop a cyclic inference mapping to alleviate the mode collapse problem. However, its performance is relatively poor in terms of both generation quality and inference accuracy, which leads to blurry images and the consequential poor inference results in inaccurate inference mapping.

This paper proposes a novel acyclic discriminator feature based inference (DFI) algorithm that exceeds both accuracy and efficiency of inference mapping for current techniques (Fig. 1). To improve inference accuracy, we suggest (1) replacing image reconstruction loss (evaluated with $x \sim P_{\mathrm{data}}$) with latent reconstruction loss (evaluated with $z \sim P_{\mathrm{z}}$) as an objective function for inference mapping, and (2) substituting the encoder with the discriminator as the feature extractor to prevent sample bias caused by latent reconstruction loss. Section 3.1 discusses this issue in detail.

Consequently, the proposed algorithm performs inference in the order of $x \rightarrow \mathrm{D}^{\mathrm{f}}$ and then $\mathrm{D}^{\mathrm{f}} \rightarrow z$, where D^{f} implies the discriminator feature. Fortunately, since the pre-trained discriminator reveals $x \rightarrow \mathrm{D}^{\mathrm{f}}$, we only focus on finding $\mathrm{D}^{\mathrm{f}} \rightarrow z$. Since this mapping is a low-to-low dimensional translation, it is much more efficient than direct encoder based approaches of $x \rightarrow z$ in terms of model parameters. Thus, the proposed algorithm achieves computational efficiency in training.

We need to consider two aspects to evaluate inference mapping: how accurately the reconstructed image preserves semantic attributes, i.e., fidelity, and reconstructed image quality after applying the inference mapping. To quantify these two aspects, we evaluated inference models with five metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), learned perceptual image patch similarity (LPIPS) (Zhang et al. 2018b), face attribute classification accuracy, and Fréchet inception distance (FID) (Dowson and Landau 1982). We use multiple metrics for evaluation because no single metric is sufficient to quantify both aspects simultaneously. The comparison confirmed that the proposed DFI outperformed existing cyclic and acyclic inference in terms of both fidelity and quality.

As a new and attractive application using the proposed inference mapping, we developed a spatially conditioned GAN (SCGAN) that can precisely control the spatial semantics for image generation. SCGAN successfully solves the spatially conditioned image generation problem due to the accurate and efficient latent estimation from the proposed inference model.

Extensive comparisons with current inference models and experimental analysis confirmed that the proposed inference algorithm provided accurate and efficient solutions for inference mapping.

2 Preliminaries

The following sections describe acyclic and cyclic inference models.

2.1 Acyclic Inference Models

An acyclic inference model develops an inference mapping on top of a pre-trained GAN model. Thus, it consists of two steps.

1. Generation mapping is established by training a baseline GAN model.
2. For inference mapping, the inference model is trained by minimizing the difference between x and its reconstructed image x', where x' is $\mathrm{G}(z')$, G is determined at step (1), and z' is the result of the inference model.

Since all generator and discriminator parameters are fixed during the inference mapping step, acyclic inference models leave baseline GAN performance intact.

CoGAN (Liu and Tuzel 2016) and BEGAN (Berthelot et al. 2017) formulate inference mapping through a searching

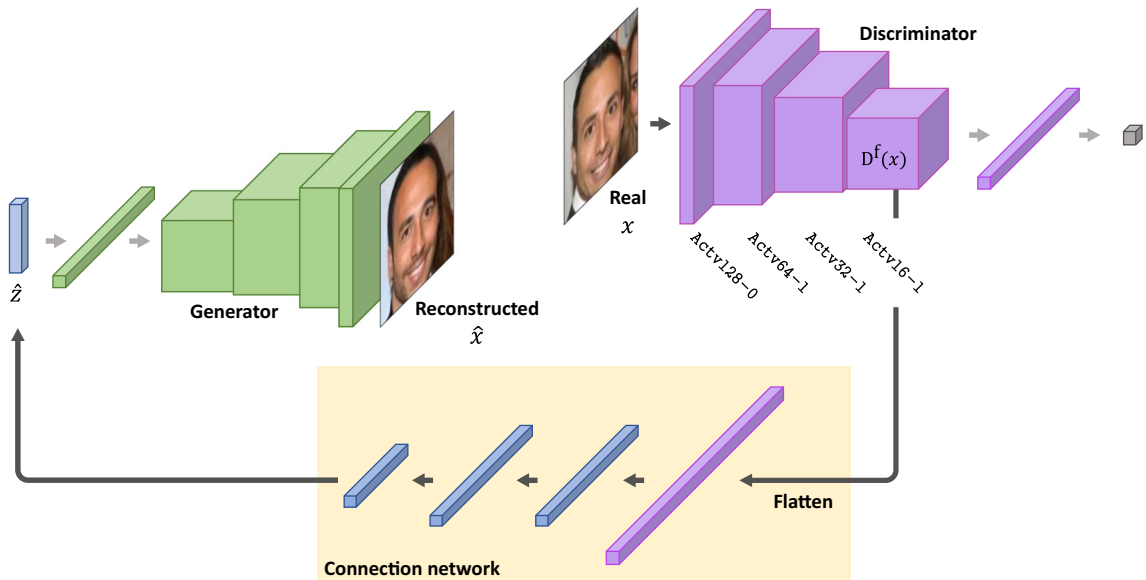

Fig. 1 Network architecture for the proposed discriminator feature based inference (DFI) model, comprising a discriminator and connection network. The discriminator extracts feature $D^f(x)$ of input image x, and then the connection network infers the latent vector $\hat{z}$ of the input image

problem. Specifically, they search latent z, which is associated with the image most similar to target image x. They use a pixel-wise distance metric to measure the similarity, and hence this problem is defined as

$$\min_z d(x,\ \mathrm{G}(z)),\ z \sim P_z,\ z^0 \in \mathcal{R}^{|z|}, \tag{2}$$

where $d(\cdot)$ is the distance metric and z^0 is the initial value for optimization. Equation 2 can be solved using advanced optimization algorithms, such as L-BFGS-B (Byrd et al. 1995u) or Adam (Kingma and Ba 2015). Although this inference process is intuitive and simple, its results are often inaccurate and generally inefficient. This non-convex optimization easily falls into spurious local minima due to the generator's non-linear and highly complex nature, and estimation results are significantly biased by the particular z^0 selected. The optimization based inference algorithm also requires intensive computational effort in the testing phase, which is prohibitive for real-time applications.

To mitigate these drawbacks, iGAN (Zhu et al. 2016) focused on providing a good initial z^0 to assist the optimization search in terms of both effectiveness and efficiency, proposing a hybrid method combining an encoder model and optimization module sequentially. The method first predicts z^0 for the input x using an encoder model, and the best estimate for subsequent z is approximated by minimizing pixel difference between $\mathrm{G}(z)$ and x. Thus, the first step for training the encoder model E is defined as

$$\min_{\mathrm{E}} \mathop{\mathbb{E}}_{x \sim P_{\text{data}}} [d(x,\ \mathrm{G}(\mathrm{E}(x)))]. \tag{3}$$

The second step is the same optimizing Eq. 2 except that the predicted latent vector is used as an initial value, $z^0 = \mathrm{E}(x)$. Consequently, iGAN reduces computational complexity for inference mapping at runtime. However, since the encoder training utilizes samples from the data distribution, inference accuracy is severely degraded by the pre-trained generator having a mode missing problem, i.e., the generator is incapable of representing the minor modes. Section 3.1 discusses this issue in more detail. Due to this accuracy issue, iGAN often misses important input data attributes, which are key components for interpreting the input.

2.2 Cyclic Inference Models

Cyclic inference models learn inference and generation mapping simultaneously. Variational (VAE) (Kingma and Welling 2013) and adversarial (AAE) (Makhzani et al. 2016) autoencoders are popularly employed to learn bidirectional mapping between z and x. Their model architectures are quite similar to autoencoders (Baldi 2012), comprising an encoder, i.e., the inverse generator, and a decoder, i.e., the generator. In contrast to autoencoders, VAE and AAE match latent distributions to prior distributions (Wainwright et al. 2008), enabling data generation. Whereas VAE utilizes Kullback–Leibler divergence to match latent and prior distributions, AAE utilizes adversarial learning for latent distribution matching. Although both algorithms establish bidirectional mapping between the latent and data distributions through stable training, their image quality is poorer than for unidirectional GANs. Specifically, generated images are blurry with lost details.

The ALI (Dumoulin et al. 2017) and BiGAN (Donahue et al. 2017) bidirectional GANs jointly learn bidirectional mapping between z and x in an unsupervised manner. They use a generator to construct forward mapping from z to x, and then an encoder to model inference mapping from x to z. To train the generator and the encoder simultaneously, they define a new objective function for the discriminator to distinguish the joint distribution, $\{\mathrm{G}(z), z\}$, from $\{x, \mathrm{E}(x)\}$. Thus, the ALI and BiGAN objective function is

$$\min_{\mathrm{G}}\max_{\mathrm{D}} \mathop{\mathbb{E}}_{x\sim P_{\text{data}}} \left[\log\left(\mathrm{D}(x, \mathrm{E}(x))\right)\right] + \mathop{\mathbb{E}}_{z\sim P_{\mathrm{z}}} \left[\log(1 - \mathrm{D}(\mathrm{G}(z), z))\right]. \tag{4}$$

Although these models can reconstruct the original image from the estimated latent vector, generation quality is poorer than that for unidirectional GANs due to convergence issues (Li et al. 2017). In contrast, they alleviate the unidirectional GAN mode collapse problem by utilizing inference mapping.

The VEEGAN (Srivastava et al. 2017) and ALICE (Li et al. 2017) introduce an additional constraint that enforces the reconstructed image (or the latent vector) computed from the estimated latent vector (or image) to match the original image (or latent vector). This improves either mode collapse or training instability for bidirectional GANs. Specifically, VEEGAN utilizes cross-entropy between P_{Z} and $\mathrm{E}(x)$, defined as the reconstruction penalty in the latent space, to establish joint distribution matching; whereas ALICE aims to improve GAN training instability by adopting conditional entropy, defined as cycle consistency (Zhu et al. 2017). Although both methods improve joint distribution matching performance, they still suffer from discrepancies between theoretical optimum and practical convergence (Li et al. 2017), resulting in either slightly blurred generated images or inaccurate inference mapping.

3 Discriminator Feature Based Inference

The proposed algorithm is an acyclic inference model, in that the training process is isolated from GAN training, i.e., both the generator and discriminator are updated. This implies that baseline GAN model performance is not affected by inference mapping. Our goal with the proposed pre-trained GAN model, is to (1) increase inference mapping accuracy and (2) build a real-time inference algorithm with minimal training overhead.

Therefore, we propose a discriminator feature based inference algorithm to achieve these goals. Specifically, we build a connection network that establishes the mapping from image features to the latent vector by minimizing latent reconstruction loss. We formulate the objective for learning the connection network as

$$\min_{\mathrm{CN}} \mathop{\mathbb{E}}_{z\sim P_{\mathrm{z}}} [d(z,\ \mathrm{CN}(\mathrm{D}^{\mathrm{f}}(\mathrm{G}(z))))], \tag{5}$$

where CN is the connection network, and $\mathrm{D}^{\mathrm{f}}(x)$ indicates the discriminator feature vector of x, extracted from the last layer of the discriminator.

In our framework, the generated image from z is projected onto the discriminator feature space, and this feature vector then maps to the original z using the connection network. It is important to understand that correspondences between the latent vector z and discriminator features $\mathrm{D}^{\mathrm{f}}(x)$ are automatically set for arbitrary z once both generator and discriminator training ends. Hence, the connection network is trained to minimize the difference between z and its reconstruction by the connection network.

The following sections provide the rationale for the proposed algorithm (Sect. 3.1), suggest a new metric for inference mapping (Sect. 3.2), and then introduce a spatially conditioned GAN (SCGAN) practical application of the proposed DFI (Sect. 3.3). We stress that SCGAN addresses spatial conditioning for image generation for the first time.

3.1 Rationale

Why DFI is Superior to Previous Acyclic Algorithms The classic iGAN acyclic inference algorithm uses an encoder based inference model that minimizes image reconstruction loss in Eq. 3 in the first stage. In contrast, the proposed DFI aims to minimize latent reconstruction loss for training the connection network. These approaches are identical for an ideal GAN, i.e., perfect mapping from z to x. However, practical GANs notoriously suffer from mode collapse; where the generator only covers a few major modes, ignoring the often many minor modes.

Suppose that the distribution reproduced by the generator P_{g} does not cover the entire distribution of P_{data}, i.e., mode collapse. Then, consider the sample x, where $P_{\mathrm{g}}(x) = 0$ and $P_{\text{data}}(x) \neq 0$. For such a sample, image reconstruction loss between x and $x' = \mathrm{G}(\mathrm{E}(x))$ by Eq. 3 is ill-specified (Srivastava et al. 2017), where E is an inference algorithm that maps an image to a latent vector, since x' is undefined by the generator. Any inference model trained with image reconstruction loss inevitably leads to inaccurate inference mapping, due to those undefined samples. In other words, the image reconstruction suffers from noisy annotations since it learns the mapping from the real image to its latent code, which are latent codes for real images not covered by the generator. This leads to inference accuracy degradation, e.g. attribute losses and blurry images.

In contrast, latent reconstruction loss only considers the mapping from $z' = \mathrm{E}(\mathrm{G}(z))$ to $z \sim P_{\mathrm{z}}$, i.e., latent reconstruction loss does not handle samples not covered by the generator. Thus, Eq. 5 solves a well-specified problem: a

set of accurate image-annotation pairs are used for training. This can significantly influence inference accuracy, and is critical for acyclic inference models developed with a pre-trained generator having practical limitations, such as mode collapse.

We stress that inference mapping using a fixed generator is trained via a set of image-latent pairs in a fully supervised manner. Since supervised learning performance largely depends on annotation quality, refining the training dataset to improve annotation accuracy often improves overall performance. In this regard, the proposed latent reconstruction loss can be interpreted as the improving annotation quality, because it tends to train inference mapping using correct image-latent pairs.

Why the Discriminator is a Good Feature Extractor for DFI

Although the discriminator is typically abandoned after GAN training, we claim it is a useful feature extractor for learning the connection network. The previous study (Radford et al. 2016) empirically showed that discriminator features are powerful representations for solving general classification tasks. The discriminator feature representation is even more powerful for inference mapping, for the following reasons.

To train the connection network using latent reconstruction loss, all training samples are fake samples, drawn from $z \sim P_z$, as described in Eq. 5. Although utilizing latent reconstruction loss is useful to construct a well-specified problem, this naturally leads to sample bias, i.e., a lack of real samples, $x \sim P_{\text{data}}$, during training. To mitigate training bias, we utilize the discriminator as a feature extractor, because the discriminator feature space already provides comprehensive representation for both real and fake samples. Thus, the pre-trained discriminator learns to classify real and fake samples during training. Consequently, we expect that the discriminator feature space can bridge the discrepancy between real and fake samples, helping to alleviate sample bias.

3.2 Metrics for Assessing Inference Accuracy

Although several metrics are available for evaluating GAN models, an objective metric for assessing inference models has not been established. Developing a fair metric is beneficial to encourage constructive competition, and hence escalate the advance of inference algorithms.

Two aspects should be considered to evaluate inference algorithm accuracy: semantic fidelity and reconstructed image quality. We utilize LPIPS (Zhang et al. 2018b) and face attribute classification (FAC) accuracy (Liu et al. 2019) to measure reconstructed image semantic fidelity, i.e., similarity to the original image. Section 4.2 empirically discusses the high correlation between LPIPS and FAC accuracy. Therefore, we employ LPIPS as the measure for semantic fidelity for further experiments because FAC accuracy is not flexible enough to apply on various datasets. In addition, We suggest FID (Dowson and Landau 1982) to measure the image quality, i.e. how realistic the image is. We emphasize that LPIPS is more suitable to measure the fidelity of the reconstructed image while FID is more suitable to measure the image quality of the reconstructed image.

LPIPS The learned perceptual image patch similarity (LPIPS) metric for image similarity utilizes a pre-trained image classification network e.g. AlexNet (Krizhevsky 2014), VGG (Simonyan and Zisserman 2015), and SqueezeNet (Iandola et al. 2016) to measure feature activation differences between two images, and returns a similarity score using learned linear weights. LPIPS can capture semantic fidelity because both low and high level features of the pre-trained network influence similarity.

FID Although LPIPS is a powerful metric for semantic fidelity, it does not reflect reconstructed image quality. We need to consider whether the reconstructed image is on the image manifold to measure quality. FID is a popular metric that quantifies sample quality and diversity for generative models, particularly GANs (Lucic et al. 2018; Zhang et al. 2018a; Brock et al. 2018), where smaller FID score indicates fake samples have (1) high quality (i.e., they are sharp and realistic) and (2) various modes similar to real data distribution.

FID represents the Fréchet distance (Dowson and Landau 1982) between the moments of two Gaussians, representing the feature distribution of real images and randomly drawn fake images. We also utilize FID for evaluating inference algorithms. For that, the Fréchet distance between moments of two Gaussians are measured where two Gaussians represent feature distributions for real images and their reconstructed images.

The FID for the inference algorithm can be expressed as

$$d^2((\mu, \Sigma), (\mu_{\text{R}}, \Sigma_{\text{R}})) = ||\mu - \mu_{\text{R}}||_2^2 + \text{Tr}\left(\Sigma + \Sigma_{\text{R}} - 2\left(\Sigma\Sigma_{\text{R}}\right)^{1/2}\right), \tag{6}$$

where (μ, Σ) (or $(\mu_{\text{R}}, \Sigma_{\text{R}})$) indicates the mean vector and covariance matrix for the Inception features computed from real images (or reconstructed images obtained by inference mapping).

It is important to note that the FID for the inference algorithm is an unbiased estimator since each reconstructed image has its real image pair. Thus, the FID for the inference algorithm provides a deterministic score for given real image set, reliable even for small test samples.

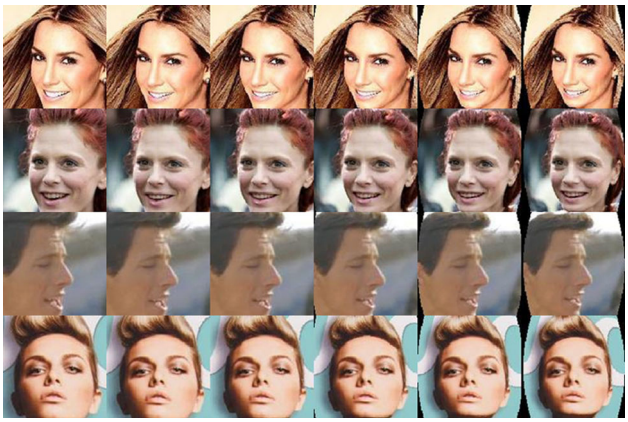

Fig. 2 Fish-eye distortion examples. Each column depicts distorted results with distortion coefficient 0.0 (original), 0.1, 0.2, 0.3, 0.4, and 0.5

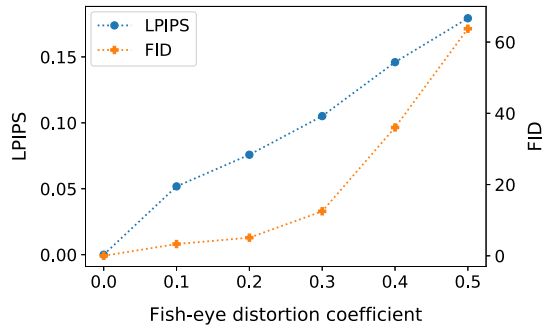

Fig. 3 Quantitative comparison between LPIPS and FID. Each score is computed using the real images and their distorted images where the fish-eye distortion coefficient gradually increases from 0 to 0.5

Fig. 4 Translation examples with two different padding strategies. First image is an original image. From the second to fourth, each image depicts the result of translation (0, 0), (21, 21) with raw padding, and (21, 21) with reflection padding respectively.

Rationale of Using Both Metrics To justify the above mentioned properties of LPIPS and FID, we provide one exemplar case and two empirical studies. First, the advantage of LPIPS can be clearly demonstrated by the following example. Note that LPIPS guarantees the ideal reconstruction if its score is zero. Meanwhile, any permutation of perfect reconstruction can yield zero FID. This indicates that LPIPS is reliable to measure faithful reconstruction; FID is not.

Contrary, LPIPS is overly sensitive to structural perturbations between the two images, thus not suitable to assess the general image quality. In fact, such a sensitivity is natural because LPIPS directly measures the pixel-level difference between two feature activations across all scales. It should be noted that FID is robust against the structural perturbations because it does not evaluate the pixel-level difference between the feature maps of the two images, but evaluates the statistical differences of the two high-level feature distributions. To demonstrate the advantage of FID, we carry out two experiments; measuring LPIPS and FID between (1) the real images and their fish-eye distorted images, and (2) the real images and their translated images. The experiment utilizing fish-eye distortions is also conducted in Zhang et al. (2018b). Figure 2 depicts several distorted images. From the left to the right, the fish-eye distortion parameter increases(the larger the parameters, the harsher the distortion). Figure 3 shows LPIPS and FID scores when distortion parameters increases. We observe that FID does not change much for the images with small distortions while the score exponentially increase for the images with large distortions. This makes sense and is analogous to how human evaluates the difference between the two images; the three images corresponding to small distortions in Fig. 3 (parameter 0.1, 0.2 and 0.3) are more similar to the original while the last two images (parameter 0.4 and 0.5) are clearly different from the original. Unlike FID, LPIPS are linearly increases as the distortion parameter increases. That means, LPIPS is not robust against small structural perturbations.

We further investigate the property of FID and LPIPS by applying random translation in real images. For padding after translation, we select two strategies; raw padding and reflection padding. For raw padding, we center crop image after shifting the original real image. For reflection padding, we center-crop image first and shift the cropped image with reflection padding. As seen from Fig. 4, raw padding results in realistic images whereas reflection padding creates creepy and unrealistic faces. We apply random shift for both vertical and horizontal axis of the image within the range $(-t, -t) \sim (t, t)$ where t is a translation coefficient. Figure 5 describes LPIPS and FID score as the translation coefficient increases. Interestingly, we observe that the difference between LPIPS scores for the two padding strategies are marginal. Contrary, the difference between the two FID scores for the two different padding strategies is considerable. Specifically, the translation using raw padding leads extremely small FID scores (FID less than 2 is almost negligible) while the translation using reflection padding yields meaningful difference in FID scores. These results present that the FID is more suitable to measure image quality, i.e., how realistic the generated samples are, than LPIPS.

From two empirical studies, we conclude that FID is more robust to small structural perturbations in images than LPIPS. Owing to this attractive properties, we confirm that FID better evaluates the image quality than LPIPS. Considering the

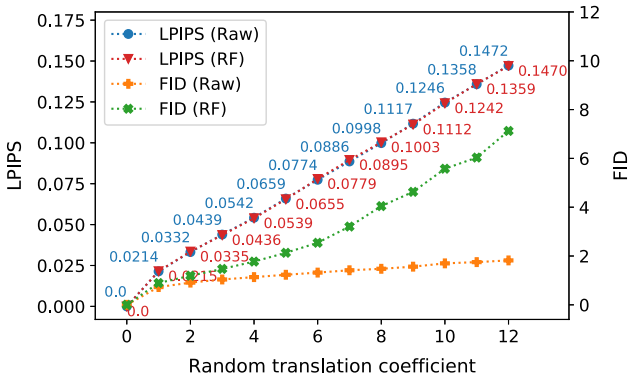

Fig. 5 Quantitative results of LPIPS and FID score between the real images and their translated images according to the random translation coefficient. To clearly visualize the difference in LPIPS scores for two padding strategies, LPIPS scores are overlaid to each point

advantages of FID and LPIPS in different aspects, we claim that both FID and LPIPS should be used for assessing inference algorithms. For this reason, we report both scores as quantitative measures for various inference algorithms.

Although we include PSNR and SSIM metrics, their scores do not reflect perceptual quality well. We argue that LPIPS and FID can better assess inference algorithm modeling power. Section 4.2 empirically shows PSNR and SSIM demerits as accuracy measures for inference algorithms.

3.3 Spatially Conditioned Image Generation

Semantic features are key components for understanding and reflecting human intentions because they are closely related to human interpretation. Indeed, the way humans define tasks is never specific but is rather abstract or only describes semantic characteristics. For example, human facial memorizing does not rely on local details, such as skin color or roughness, but focuses more on facial shape, hair color, presence of eyeglasses, etc. Therefore, from the human viewpoint, useful image analysis and manipulation should be associated with extracting semantic attributes of the data and modifying them effectively. Since the proposed inference algorithm developed by the connection network establishes semantically accurate inference mapping, combining this inference algorithm with standard GANs can provide strong baseline models for data manipulation and analysis applications.

Therefore, we suggest a new conditional image generation algorithm: spatially conditioned GAN (SCGAN). SCGAN extracts the latent vector of input using the proposed inference algorithm and uses it for spatially conditioned image generation.

In particular, we specify the position input image position, and then generate the surroundings using SCGAN. In this process, the generated surrounding region should naturally and seamlessly match the input image. Among the infinite methods to generate the outside regions, our goal is to achieve semantically seamless results. Therefore, SCGAN first maps the input image to its latent vector using DFI, which encodes the semantic attributes. Given the latent vector of input, spatially conditioned image generation is conducted by generating the large image (full size) such that the image region at the input position is the reconstructed input and its surroundings are newly generated. The generated surroundings should seamlessly match the semantics of the input with reasonably visual quality. Since many possible surroundings can match the input, we formulate the latent vector of the generated image by concatenating the random vector with the latent vector of input. Thus, SCGAN maintains input semantic attributes while allowing diverse image surroundings.

Figure 6 illustrates the proposed SCGAN architecture. To extract the latent vector for input image x_{center}, we first train baseline GANs, comprising a generator G_{center} and discriminator D_{center}, and then fix the GANs and train the connection network (CN) to utilize DFI. Given the fixed D_{center} and CN, we compute $\hat{z}_{center}$, the estimated latent vector for x_{center}. To account for diverse surroundings, we concatenate a random latent vector z_{edge} with $\hat{z}_{center}$ and feed this into the generator G_{full}. This network learns to map the concatenated latent vector to full size image y_{full}, which is the final output image.

We train G_{full} to satisfy y_{crop}: the image center of y_{full} should reconstruct x_{center}; and y_{full} should have a diverse boundary region and sufficiently high overall quality. To meet the first objective, the naïve solution is to minimize L1/L2 distance between y_{crop} and x_{center}. However, as reported previously (Larsen et al. 2015), combining image-level loss with adversarial loss increases GAN training instability, resulting in quality degradation. Hence, we define reconstruction loss in the latent space, i.e., we map y_{crop} onto its latent vector via DFI (D_{center} and CN), then force it to match $\hat{z}_{center}$. Thus, the semantic similarity between the input and its reconstruction is preserved.

To ensure seamless composition between reconstructed and generated regions, adversarial loss for G_{full} consists of feedback from y_{full} and y_{glue}. y_{glue} is obtained by substituting the generated image center y_{crop} with the reconstructed input y_{center}. This term for y_{glue} helps generate visually pleasing images, i.e., reconstructed input and its surroundings are seamlessly matched. Thus, generator loss includes two adversarial losses and latent reconstruction loss,

$$\min_{\mathrm{G}_{full}} 0.5\, L_{\mathrm{G}}^{adv} + \alpha\, L^{recon}, \tag{7}$$

$$L^{recon} = \left\| \hat{z}_{center} - \mathrm{CN}(\mathrm{D}_{center}(y_{crop})) \right\|_1,$$

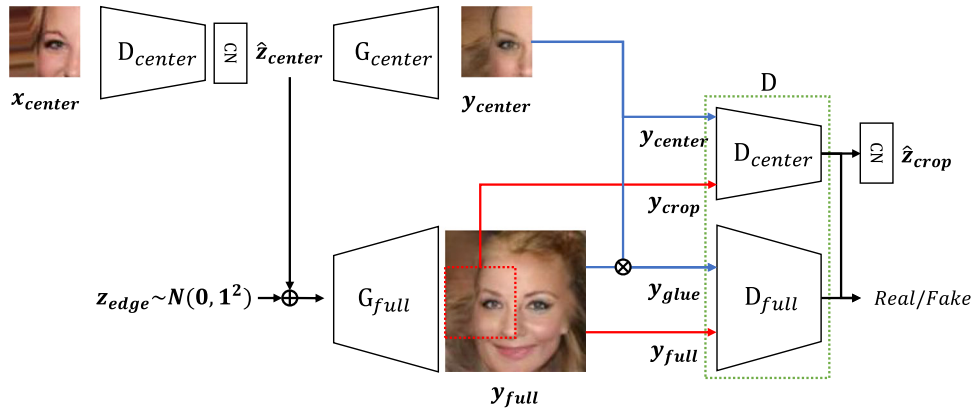

Fig. 6 Network architecture for the proposed SCGAN. $\oplus$ denotes concatenation of latent vectors. $\otimes$ denotes image replacement. y_{glue} is identical to y_{full} except the image center (square area outlined by red dots), which was replaced with y_{center}. The design choice for D was motivated by Iizuka et al. (2017), and includes a global discriminator D_{full} and local discriminator D_{center} (Color figure online)

and

$$L_G^{adv} = \mathop{\mathbb{E}}_{z_{edge} \sim P_z} \big[\log(1 - D(y_{full})) + \log(1 - D(y_{glue}))\big],$$

respectively.

Semantic consistency between reconstructed and generated regions is important to create natural images. To obtain locally and globally consistent images, we utilize local and global discriminator D (Iizuka et al. 2017) architecture that uses discriminator features from both D_{center} and D_{full}. We also employ PatchGAN (Isola et al. 2017) architecture to strengthen the discriminator, accounting for semantic information from patches in the input, and apply the zero-centered gradient penalty (0GP) (Mescheder et al. 2018) to D_{full} to facilitate high resolution image generation. Considering adversarial loss and zero-centered gradient penalty, discriminator loss can be expressed as

$$\max_{D_{full}} L_D^{adv} + 0.5\, L_G^{adv} + L^{GP} \tag{8}$$

$$L_D^{adv} = \mathop{\mathbb{E}}_{x \sim P_{data}} [\log D(x)], \; and \; L^{GP} = \frac{\gamma}{2} \mathop{\mathbb{E}}_{x \sim P_{data}} [\|\nabla D(x)\|^2].$$

4 Experimental Results

For a concise expression, we use the abbreviation for network combinations for the rest of the paper. Table 1 summarizes the component of each network model and its abbreviation. For additional optimization, each baseline model first infers initial z_0 and then optimize z by following Eq. 2 for 50 iterations (Zhu et al. 2016).

Metrics for Quantitative Evaluation We employed PSNR, SSIM, LPIPS, face attribute classification (FAC) accuracy, FID, and a user study to quantitatively evaluate various inference algorithms. For the user study, 150 participants compared real images with their reconstruction from all inference models to select the one that most similar to the real image. Each participant then responded to three questions.

1. We provided seven images: the original and reconstructed images from (a) ENC_{image}, (b) ENC_{image}^{opt} (iGAN), (c) ENC_{latent}, (d) ENC_{latent}^{opt}, (e) DFI, and (f) DFI^{opt}. We asked the participant to select the image most similar to the original image from among the six reconstructed images.
2. The participant was asked to explain the reason for their choice.
3. We provided DFI and DFI − VGG16 (discussed in Sect. 4.6) images, and asked participants to select the one most similar to the original.

This was repeated 25 times using different input images.

State-of-the-Art Inference Algorithms for Comparison Experimental comparisons are conducted for acyclic and cyclic inference models. First, we compare the proposed inference algorithm with three acyclic inference algorithms: naïve encoder (ENC_{image} and ENC_{latent}), hybrid inference by iGAN (Zhu et al. 2016) (ENC_{image}^{opt}), and hybrid inference combined with DFI (DFI^{opt}). The proposed DFI model outperformed all three acyclic models for all four evaluation methods (LPIPS, FAC accuracy, FID, and user study).

We then compared current cyclic models (VAE, ALI/BiGAN, and ALICE) with the proposed DFI based model upon various baseline GAN models. Cyclic model inference mapping influences baseline GAN performance, whereas acyclic model (i.e., DFI) inference mapping does not. We combined six different baseline GANs with DFI for this evaluation: DCGAN (Radford et al. 2016), LSGAN (Mao et al. 2017), DFM (Warde-Farley and Bengio 2017), RFGAN (Bang and Shim 2018), SNGAN (Miyato et al. 2018), and WGAN-GP (Gulrajani et al. 2017). These six were selected because they are significantly different from

Table 1 The abbreviation for various baseline models and the variants of the proposed models

	Training loss		Architecture		Additional optimization
	Image reconstruction loss	Latent reconstruction loss	Encoder trained from scratch	Discriminator (D^f) + CN network	
ENC_{image}	✓		✓		
ENC_{image}^{opt} (iGAN)	✓		✓		✓
ENC_{latent}		✓	✓		
ENC_{latent}^{opt}		✓	✓		✓
DFI_{image}	✓			✓	
DFI_{image}^{opt}	✓			✓	✓
DFI (proposed)		✓		✓	
DFI^{opt}		✓		✓	✓

each other in terms of loss functions or network architectures. We evaluated all results with $\leq (64, 64, 3)$ resolution since cyclic models are unstable for high resolution images. To illustrate DFI scalability, we build inference mapping with high resolution GANs (Mescheder et al. 2018; Miyato et al. 2018) combined with DFI, and observed similar tendency in terms of inference accuracy for $(128, 128, 3)$ resolution images.

Qualitative Evaluation for DFI Generators learn rich linear structure in representation space due to the power of semantic representations of GAN latent space (Radford et al. 2016). To qualitatively evaluate semantic accuracy for the proposed DFI, we conducted two simple image manipulation tasks: latent space walking and vector arithmetic.

Model Architecture for Fair Comparison To ensure fair evaluation, we based baseline GAN architectures on DCGAN for low resolution and SNGAN for high resolution experiments, i.e., number of layers, filter size, hyper-parameters, etc. The connection network included just two fully connected (FC) layers: 1024—group normalization (GN) (Wu and He 2018)—leaky rectified linear unit (Leaky ReLU)—1024 FC—GN—Leaky ReLU—dimension of P_z FC.

Datasets One synthetic and three real datasets were used for both qualitative and quantitative evaluations. We generated eight Gaussian spreads for the synthetic dataset distribution. Real datasets included Fashion MNIST (Xiao et al. 2017), CIFAR10 (Krizhevsky and Hinton 2009), and CelebA (Liu et al. 2015), and were all normalized on $[-1, 1]$. Input dimensionality for Fashion MNIST $= (28, 28, 1)$; CIFAR10 $= (32, 32, 3)$; and CelebA $= (64, 64, 3)$ and $(128, 128, 3)$ for low and high resolution GANs, respectively. Quantitative experiments for high resolution GANs included 10,000 images in the test set.

4.1 DFI Verification Using the Synthetic Dataset

Figure 7 (left) compares performance for the acyclic inference algorithms using the synthetic dataset. The dataset consisted of eight Gaussian spreads with standard deviation $= 0.1$. We reduced the number of samples from two Gaussian spreads at the second quadrant to induce minor data distribution modes, and then trained the GANs using real samples (green dots). The generator and discriminator included three FC layers with batch normalization. Subsequently, we obtained generated samples (orange dots) by randomly producing samples using the generator. The distributions confirm that GAN training was successful, with generated samples covering all data distribution modes.

Although the pre-trained GANs covered all modes, two modes on the second quadrant were rarely reproducible. This commonly incurs in GAN training, leading to poor diversity in sample generation. Using this pre-trained GANs, we trained (1) ENC_{image}, (2) ENC_{latent} (the degenerated version of the proposed algorithm), and (3) DFI (the proposed algorithm). Hyper-parameters and network architecture were identical for all models, i.e., DFI included the discriminator (two FC layers without the final FC layer) and the connection network (two FC layers), whereas the encoders (ENC_{image} and ENC_{latent}) included four FC layers with the same architecture and model parameters as DFI. Each inference algorithm calculated corresponding latent vectors from the test samples (gray dots), and then regenerating the test samples from the latent vectors. For sufficient training, we extract the results after 50 K iterations.

Figure 7 (right) compares performance for the inference algorithms with sample reconstruction results. The ENC_{image} (the red dots) tends to recover the right side of test samples but is incapable of recovering samples on the left side, and only five modes were recovered in this experiment; whereas ENC_{latent} (cyan and blue dots) recover many more modes

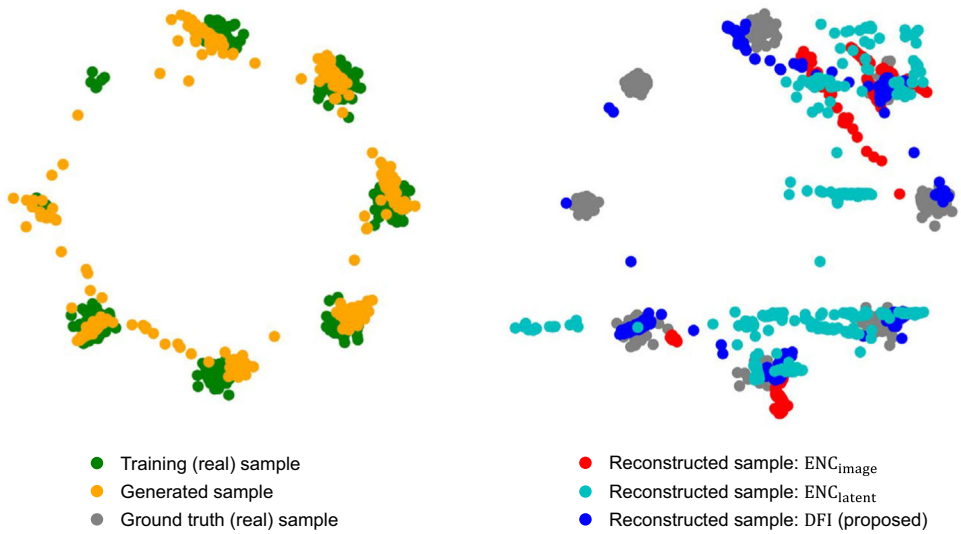

Fig. 7 Inference algorithm performances using a synthetic dataset with eight Gaussian spreads: (left) green dots are training (real) samples, orange dots are generated samples from baseline GAN generators; (right) gray dots are ground truth (real) test samples, red dots are reconstructed samples by the $\mathrm{ENC}_{\mathrm{image}}$, the cyan dots are reconstructed samples by the $\mathrm{ENC}_{\mathrm{latent}}$, and the blue dots are reconstructed samples by the proposed DFI (Color figure online)

after reconstruction. This visual comparison clearly demonstrates the $\mathrm{ENC}_{\mathrm{image}}$ drawbacks.

For inference algorithms with the same latent reconstruction loss, DFI significantly outperforms the algorithm using the $\mathrm{ENC}_{\mathrm{latent}}$. In particular, the reconstructed samples using the $\mathrm{ENC}_{\mathrm{latent}}$ are inaccurate in terms of reconstruction accuracy because considerable portions of reconstructed samples (e.g. cyan dots in the middle) are far from all eight Gaussian spreads. DFI reconstructed samples are much closer to the original Gaussian spreads, i.e., more accurate results.

Thus, latent reconstruction loss was more effective than image reconstruction loss to derive accurate acyclic inference algorithms. Utilizing the pre-trained discriminator as a feature extractor also helped to further increase inference mapping accuracy. Therefore, the proposed approach to employ latent reconstruction loss with the discriminator as a feature extractor is an effective and efficient solution for inference algorithms.

4.2 Comparison with Acyclic Inference Models

In Fig. 8, we use various objective metrics for quantitatively evaluating the inference algorithms. Specifically, PSNR, SSIM, LPIPS, face attribute classification (FAC) accuracy, FID and user study results are reported for comparing DFI with the other acyclic models. For the FAC accuracy, we utilize the same classifier as STGAN (Liu et al. 2019), that uses 13 attributes in CelebA dataset to measure accuracy. For the experimental results in CelebA, LPIPS exhibits similar tendency to FAC accuracy. Therefore, we choose LPIPS to assess inference algorithm semantic similarity for the remaining experiments since if can measure semantic fidelity on various datasets.

LPIPS, FAC accuracy, FID and the user study scores indicate DFI based models to be significantly superior. Although PSNR and SSIM scores from methods using image reconstruction loss are significantly higher than for DFI models, significant gaps in the user study confirm that PSNR and SSIM are not reliable metrics for this application. Inference algorithms with image reconstruction loss are expected to have higher PSNR and SSIM scores, simply because their objectives, i.e., minimizing pixel-level difference exactly match the metrics.

$\mathrm{ENC}_{\mathrm{latent}}$ and $\mathrm{ENC}^{\mathrm{opt}}_{\mathrm{latent}}$ results do not provide accurate fidelity (lower LPIPS). The $\mathrm{ENC}_{\mathrm{latent}}$ utilizes only fake samples for training the feature extractor, i.e., convolutional layers, whereas DFI exploits the discriminator feature extractor,which was trained with real and fake samples. Thus, the $\mathrm{ENC}_{\mathrm{latent}}$ model is incapable of capturing a common feature to represent real and fake images. Consequently, reconstruction fidelity is significantly degraded. On the other hand, their image quality, i.e., realistic and sharp, exceeds other methods using image reconstruction loss, because the inference algorithm learns to reduce image level distance regardless of the image manifold. Consequently, it tends to produce blurry images without distinct attributes, leading to quality degradation. In contrast, inference algorithms with latent reconstruction loss generally provide high quality images after inference mapping. Thus, latent distance is more favorable to retain samples onto the image manifold, helping to improve image quality.

All LPIPS, FID assessments, and user study scores confirm that DFI and $\mathrm{DFI}^{\mathrm{opt}}$ outperform the other models. Other inference mappings are particularly degraded when the input images include distinctive attributes, such as eyeglasses or a mustache; whereas the proposed DFI inference mapping consistently performs well, increasing the performance gap

Fig. 8 Qualitative and quantitative comparison of various inference algorithms. Column 1 includes target (real) images and the remaining columns include reconstructed images by each method in the order shown in the table. All images were computed after 40 K training steps. The table summarizes quantitative metrics. We used *AlexNet(lin)* (Zhang et al. 2018b) environment for LPIPS perceptual loss. FAC indicates face attribute classification accuracy using 13 attributes in CelebA dataset. Smaller LPIPS and FID indicate more accurate and realistic results. User study participants selected the image most similar to the target among six reconstructed images. The number of votes in percentage is reported in the table

	Image reconstruction loss		Latent reconstruction loss			
	$\text{ENC}_{\text{image}}$	$\text{ENC}^{\text{opt}}_{\text{image}}$ (iGAN)	$\text{ENC}_{\text{latent}}$	$\text{ENC}^{\text{opt}}_{\text{latent}}$	DFI (proposed)	DFI^{opt}
PSNR (dB)	19.15	19.43	13.57	14.15	14.77	15.66
SSIM	0.5440	0.5519	0.4271	0.4353	0.4841	0.4956
LPIPS	0.1968	0.1949	0.2279	0.2228	**0.1931**	**0.1848**
FAC (%)	87.92	88.04	87.67	87.87	**89.31**	**89.55**
FID	23.38	22.68	12.90	12.22	**8.91**	**8.12**
User study(%)	6.83	8.33	6.33	4.17	**27.83**	**46.50**

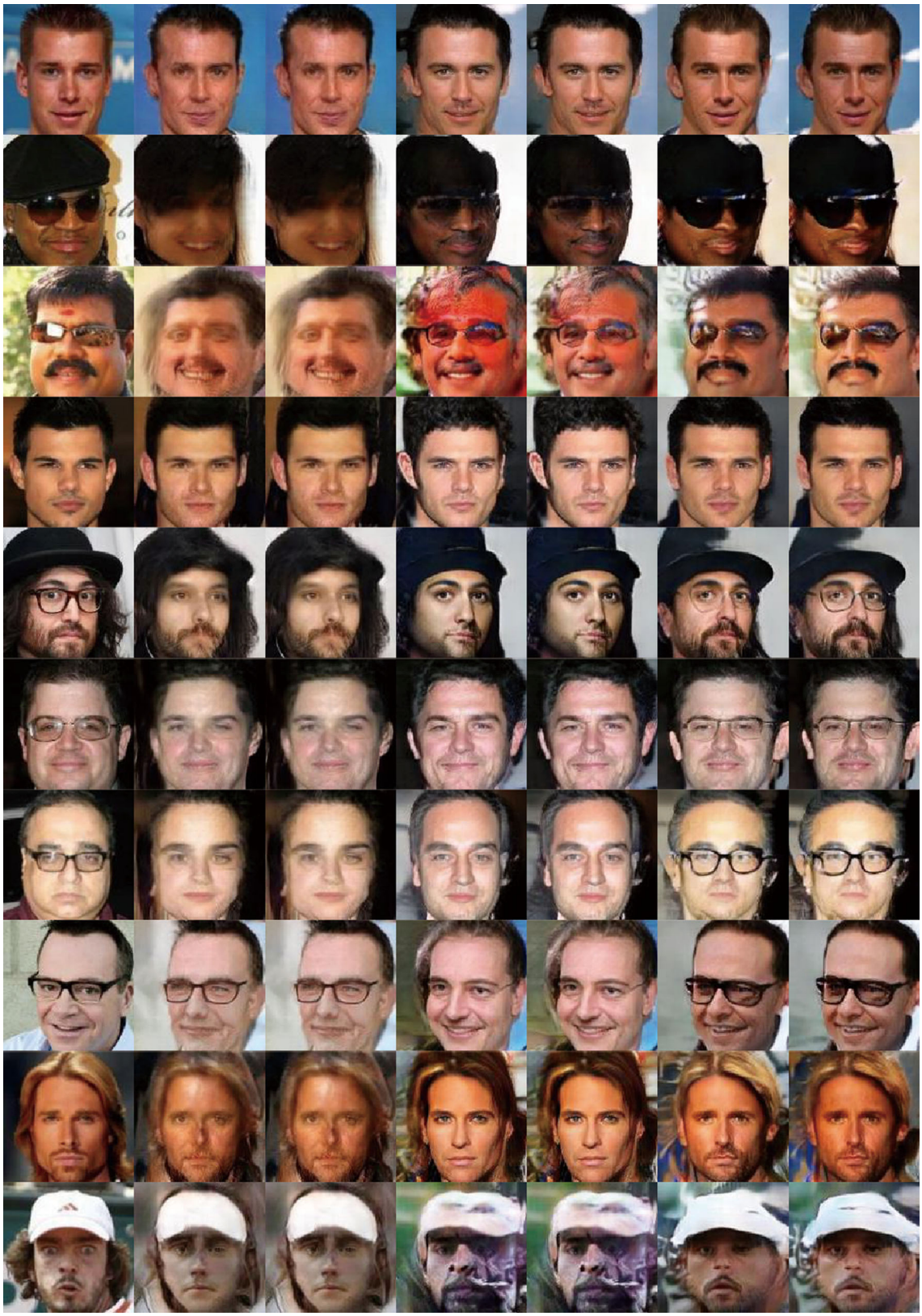

Fig. 9 Qualitative comparisons for various inference algorithms. Column 1 includes target (real) images and remaining columns include reconstructed images by each method in the order of Fig. 8. All images were computed after 40K training steps

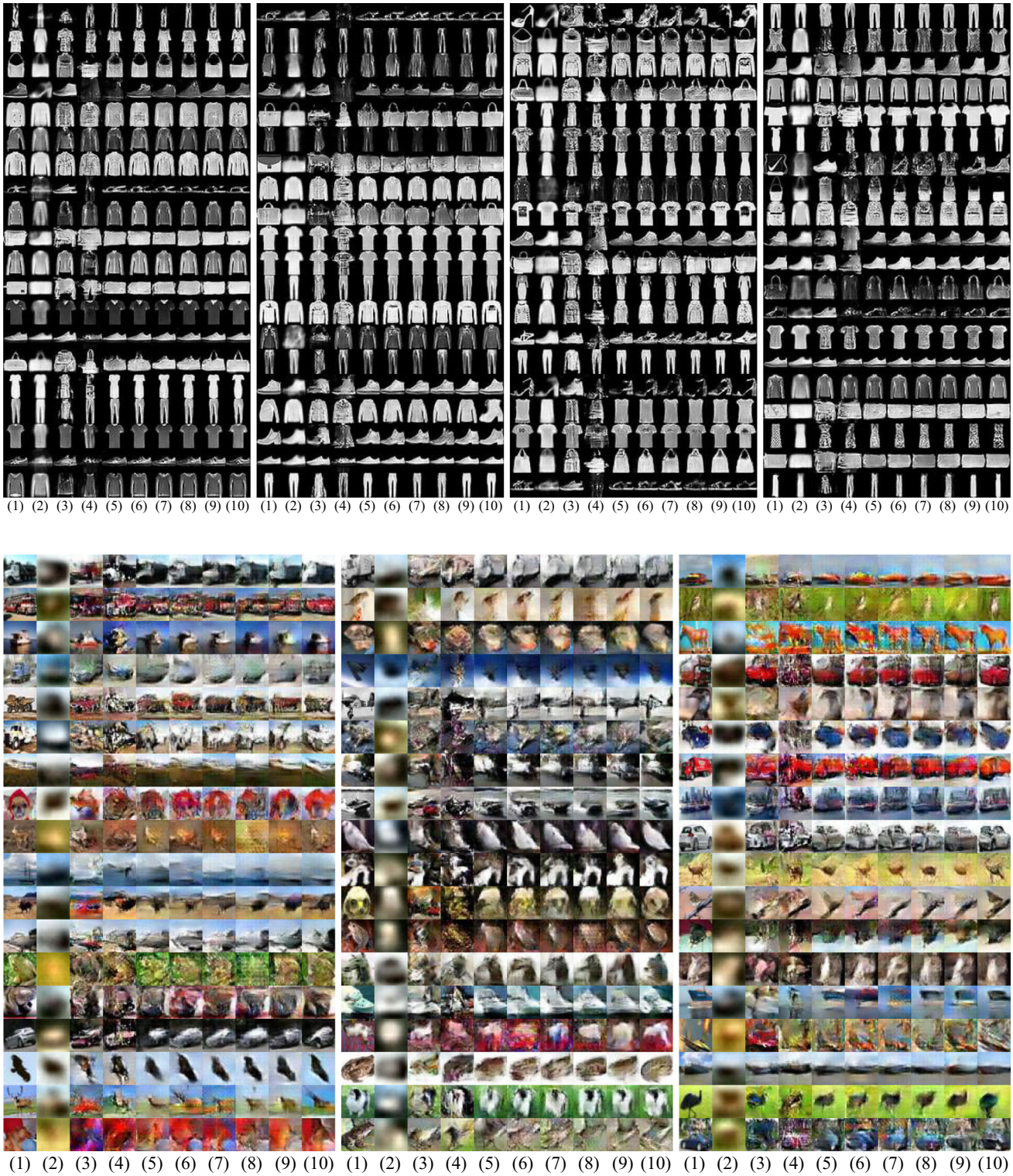

Fig. 10 Qualitative comparison with cyclic inference algorithms and DFI variants using FashionMNIST and CIFAR-10 datasets. Column (1) includes target (real) images, and the remainder include reconstructed images by (2) VAE, (3) ALI/BiGAN, (4) ALICE, DFI with {(5) DCGAN, (6) LSGAN, (7) DFM, (8) RFGAN, (9) SNGAN, and (10) WGAN-GP}

Fig. 11 Qualitative comparison with cyclic inference algorithms and DFI variants using the CelebA dataset. Column (1) includes target (real) images and the remainder include reconstructed images by (2) VAE, (3) ALI/BiGAN, (4) ALICE, DFI with {(5) DCGAN, (6) LSGAN, (7) DFM, (8) RFGAN, (9) SNGAN, and (10) WGAN-GP}

between the proposed DFI mapping and others approaches for samples with distinctive attributes. Therefore, the proposed inference mapping was effective in restoring semantic attributes and reconstruction results were semantically more accurate than other inference mappings.

Figure 9 compares the proposed DFI method with (1) encoder mapping ($\text{ENC}_{\text{image}}$ and $\text{ENC}_{\text{latent}}$), (2) hybrid inference as suggested by iGAN (Zhu et al. 2016) ($\text{ENC}^{\text{opt}}_{\text{image}}$ and $\text{ENC}^{\text{opt}}_{\text{latent}}$), and (3) DFI^{opt}. To investigate the effect of latent reconstruction loss, we modified the encoder objective function in (1) and (2) from image reconstruction loss to latent reconstruction loss.

Reconstruction results using image reconstruction loss (Columns 2 and 3 from Figs. 8, 9) are generally blurred or have missing attributes, e.g. eyeglasses, mustache, gender, wrinkles, etc., compared with DFI reconstruction results. These results support our argument in Sect. 3.1: latent reconstruction loss provides more accurate inference mapping than image reconstruction loss. Previous iGAN studies have shown that additional latent optimization after inference mapping (in both $\text{ENC}^{\text{opt}}_{\text{image}}$ and DFI^{opt}) effectively improves inference accuracy. The current study found that optimization was useful to better restore the original color distribution, based on feedback from the user study.

However, although the additional optimization fine tunes the inference mapping, it still has computational efficiency limitations. Therefore, we chose DFI without additional optimization for subsequent experiments to trade-off between accuracy and computational efficiency.

The last row in Figs. 8 and 9 present examples where all inference methods performed poorly. These poor results were due to baseline GAN performance limitations rather than the inference algorithms. However, despite the inaccurate reconstruction, the proposed DFI approach recovered many original semantic attributes, e.g. glasses on the right side and mustache on the left.

4.3 Comparison with Cyclic Inference Models

Figures 10 and 11 compare the proposed DFI approach with VAE, ALI/BiGAN, and ALICE representative generative models that allow inference mapping adopting the six baseline GANs discussed above. Table 2 shows corresponding reconstruction accuracy in terms of LPIPS and FID.

Table 2 Quantitative comparisons with cyclic inference algorithms and DFI variants for Fashion MNIST, CIFAR10, and CelebA datasets using LPIPS and FID

Dataset	Metric	Cyclic models			DFI with various GAN models					
		VAE	ALI/BiGAN	ALICE	DCGAN	LSGAN	DFM	RFGAN	SNGAN	WGAN-GP
Fashion MNIST	LPIPS	0.1287 ± 0.0017	0.0869 ± 0.0097	0.0898 ± 0.0177	0.0351 ± 0.0014	**0.0269** ± 0.0009	0.0352 ± 0.0013	0.0345 ± 0.0017	**0.0306** ± 0.0004	0.0366 ± 0.0012
	FID	90.94 ± 2.78	69.99 ± 23.73	62.10 ± 13.77	25.73 ± 1.27	19.76 ± 0.60	25.88 ± 1.00	25.26 ± 3.06	**15.53** ± 1.00	**11.85** ± 0.28
CIFAR10	LPIPS	0.4839 ± 0.0014	0.1094 ± 0.0081	0.1108 ± 0.0061	0.0744 ± 0.0033	**0.0716** ± 0.0032	0.0745 ± 0.0022	**0.0735** ± 0.0021	0.2799 ± 0.0164	0.0758 ± 0.0019
	FID	234.03 ± 1.30	54.04 ± 5.66	56.43 ± 8.44	44.83 ± 1.92	**39.67** ± 1.62	46.23 ± 2.83	46.79 ± 5.97	61.88 ± 7.75	**31.54** ± 1.49
CelebA	LPIPS	0.3535 ± 0.0026	0.1573 ± 0.0213	0.1424 ± 0.0052	**0.0980** ± 0.0012	0.1237 ± 0.0185	**0.0980** ± 0.0005	0.0994 ± 0.0009	0.0997 ± 0.0087	0.1033 ± 0.0011
	FID	94.96 ± 0.52	19.56 ± 5.61	16.36 ± 1.92	12.86 ± 0.15	24.99 ± 14.17	13.24 ± 0.16	14.02 ± 1.08	**11.92** ± 3.44	**8.73** ± 0.78

Bold values indicate Top-2 scores for each metric and dataset

Table 3 Quantitative comparison of cyclic inference algorithms and DFI variants for the CelebA dataset using LPIPS and FID: top score among high resolution (HR) GANs

Metric	DCGAN-0GP	SNGAN	SNGAN-0GP
LPIPS	0.2147 ± 0.0014	0.2116 ± 0.0002	**0.1931** ± 0.0015
FID	15.10 ± 0.54	16.09 ± 0.30	**8.91** ± 0.12

Bold values indicate Top-1 scores for each metric

Reconstructed images from VAE are blurry and lose detailed structures because it was trained with image reconstruction loss. Less frequently appearing training dataset attributes, e.g. mustache or baldness, were rarely recovered due to popularity bias. ALI/BiGAN and ALICE restore sharper images than VAE, but do not effectively recover important input image characteristics, e.g. identity, and occasionally generate completely different images from the inputs.

In contrast, reconstructed images from DFI variants exhibit consistently better visual quality than VAE, ALI/BiGAN, and ALICE. DFI training focused on accurate inference mapping, without influencing baseline GAN performance. Hence, reconstructed image quality from DFI models is identical to that of the baseline unidirectional GANs: sharp and realistic. DFI variants consistently provide more accurate reconstructions, i.e., faithfully reconstruct the input images including various facial attributes; whereas VAE, ALI/BiGAN, and ALICE often fail to handle these aspects. Thus, the proposed algorithm accurately estimates the latent vector corresponding to the input image and retains image quality better than competitors.

Table 2 confirms that inference accuracy for DFI based models significantly outperform VAE, ALI/BiGAN, and ALICE for LPIPS and FID metrics, similar to the case for qualitative comparisons. In addition, Table 3 supports the scalability of DFI for high resolution GANs. Unlike other cyclic inference algorithms, our DFI does not influence (degrade) the generation quality of baseline GANs and still provides the robust and consistent performance in inference mapping.

4.4 Ablation Study on DFI

To understand the effect of latent reconstruction on DFI, we conduct two experiments; (1) $\mathrm{DFI}_{\mathrm{image}}$ and (2) $\mathrm{DFI}^{\mathrm{opt}}_{\mathrm{image}}$. For both experiments, the training strategy is identical to DFI, i.e. a fixed discriminator for D^{f} and a trainable CN network. $\mathrm{DFI}_{\mathrm{image}}$ utilizes the image reconstruction loss instead of the latent reconstruction loss. $\mathrm{DFI}^{\mathrm{opt}}_{\mathrm{image}}$ performs an additional optimization on top of $\mathrm{DFI}_{\mathrm{image}}$.

Figure 12 demonstrates qualitative and quantitative comparisons. Compared to the results with the latent recon-

	DFI	DFI^{opt}	$\text{DFI}_{\text{image}}$	$\text{DFI}^{\text{opt}}_{\text{image}}$
PSNR (dB)	14.77	15.66	19.23	**19.28**
SSIM	0.4841	0.4956	0.5493	**0.5501**
LPIPS	0.1931	**0.1848**	0.1917	0.1913
FAC (%)	89.31	**89.55**	88.30	88.31
FID	8.91	**8.12**	21.33	21.39

Fig. 12 Ablation study on proposed DFI. The first column includes the target (real) images, (1) includes DFI reconstructed images, (2) includes DFI^{opt} reconstructed images, (3) includes $\text{DFI}_{\text{image}}$ reconstructed images, and (4) includes $\text{DFI}^{\text{opt}}_{\text{image}}$ reconstructed images, respectively. Experimental setting and metrics are identical to those for Fig. 8

struction loss, the results from $\text{DFI}_{\text{image}}$ and $\text{DFI}^{\text{opt}}_{\text{image}}$ lose semantic details and quality. Even though some samples show reasonable quality, they generally lose details such as facial expressions and glasses. For example, in the fourth row in Fig. 12, the results with the image reconstruction loss do not preserve details, whereas the results with the latent reconstruction loss do so. In table in Fig. 12, the LPIPS score of $\text{DFI}_{\text{image}}$ is better than the proposed DFI. However, its FID score is worse than DFI. This is because the methods with the image reconstruction loss are optimized to reduce the pixel-level distance that leads high structural similarity regardless of its quality. Meanwhile, FID is more robust to small structural difference than LPIPS, thereby more appropriate to measure semantic similarity. This is analogous when the examples using the image reconstruction loss are compared with the examples using the latent reconstruction loss; the method using the latent reconstruction loss preserves image quality better. Similarly, despite $\text{DFI}_{\text{image}}$ achieves the best LPIPS score among all methods that do not utilize the optimization, the image quality of $\text{DFI}_{\text{image}}$ is worse than that of DFI. Comparing $\text{DFI}_{\text{image}}$ and $\text{ENC}_{\text{image}}$, we observe similar visual quality and tendency. This result is consistent with our statement in Sect. 3.1 and the simulation experiment in Sect. 4.1. Because the image reconstruction loss utilizes real data for training the inference model although the generator may not be able to create them (i.e. undefined data), both $\text{DFI}_{\text{image}}$ and $\text{ENC}_{\text{image}}$ suffer from the inevitable errors caused by those undefined data. Despite the limitation of the image reconstruction loss, we observe that $\text{DFI}_{\text{image}}$ enjoys the quantitative improvement over $\text{ENC}_{\text{image}}$ owing to the effective feature extractor (i.e. a discriminator).

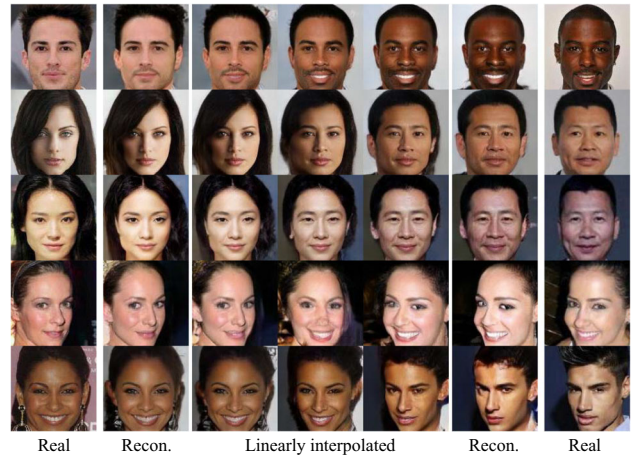

Fig. 13 Column (1) includes input images, (2)–(6) include generated images using linearly interpolated latent vector, and (7) include latent space walking results for two inferred latent spaces using column(1) images

4.5 DFI Qualitative Evaluation

To verify that DFI produced semantically accurate inference mapping, we applied latent space walking on the inferred latent vector. For two real images x_1 and x_2, we obtained inferred latent vectors z'_1 and z'_2 using DFI. Then we linearly interpolated $z'_L = \alpha z'_1 + (1-\alpha) z'_2$, where $\alpha \in [0, 1]$. Figure 13 shows images generated using z'_L, where columns (2)–(6) include interpolated images for $\alpha = 0.00, 0.25, 0.50, 0.75, 1.00$, respectively. If DFI incorrectly mapped the real images to the latent manifold, reconstructed images would exhibit transitions or unrealistic images. However, all reconstructed images exhibit semantically smooth transitions on the image space, e.g. skin color, hair shape, face orientation and expressions all change smoothly.

Figure 14 show vector arithmetic results for adding eyeglasses and mustache vector attributes (v_E and v_M, respectively):

$$v_E = \frac{1}{2}\left(v^{male}_{EO} - v^{male}_{OO}\right) + \frac{1}{2}\left(v^{female}_{EO} - v^{female}_{OO}\right), \tag{9}$$

$$v_M = v^{male}_{OM} - v^{male}_{OO},$$

where v with any superscripts and subscripts are mean sample vectors inferred by DFI; E and M in subscripts indicate eyeglasses and mustache attributes presence, respectively, in

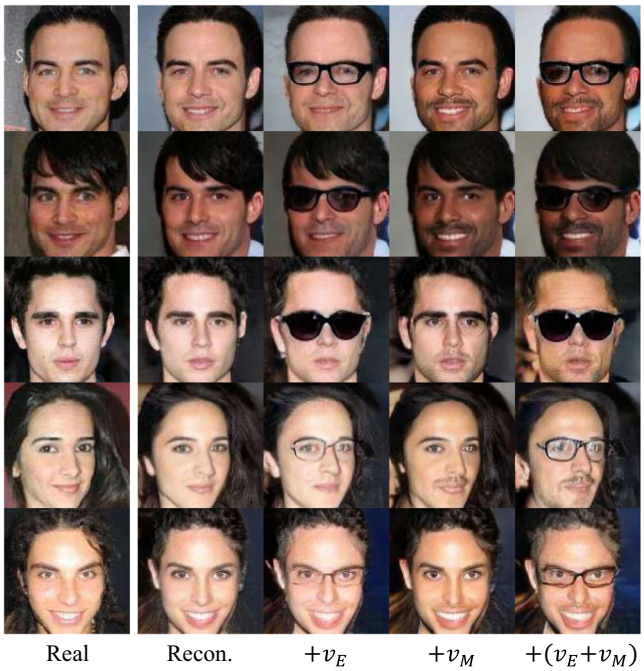

Fig. 14 Semantic image editing results using vector arithmetic on GAN latent space. Column (1) includes the original input image, (2) includes the reconstructed image using inferred latent vector by DFI, (3)–(5) include results from adding eyeglasses, mustache, and both vectors to the latent vector, respectively

sample images, and O indicates non-presence of an attribute. We used 20 images to obtain the mean inferred vector for each group. Thus, Simple vector arithmetic on the latent vector can manipulate images, e.g. adding eyeglasses, mustache, or both. Therefore, DFI successfully establishes semantically accurate mapping from image to latent space.

4.6 Feature Extractor Effects

To confirm the discriminator as a good feature extractor, we compared several DFI versions: original DFI (discriminator with CN network), DFI − VGG16 (pre-trained VGG16 (Simonyan and Zisserman 2015) using `Pool5` feature as CN network input), and DFI − ResNet (pre-trained residual based network (He et al. 2016) using features before GAP as CN network input). Among the pre-trained models, we empirically observe that DFI − VGG16 outperforms DFI − ResNet (e.g. ResNet34, ResNet50, and ResNet101) in all quantitative metrics, thus we mainly report DFI − VGG16 results. Please note that VGG16 and ResNet are well-known, powerful feature extractors with an excessive number of parameters, and should be much more powerful feature extractors for general purposes.

Figure 15 shows several reconstruction examples with quantitative evaluation results (after 40K training iteration

	DFI	DFI − VGG16
LPIPS	**0.1931**	0.1935
FAC (%)	**89.31**	88.44
FID	**8.91**	9.01
User study (%)	**78.67**	21.33

Fig. 15 Proposed DFI feature extractor effects. Column (1) includes the target (real) images, (2) includes DFI reconstructed images, and (3) includes images reconstructed from a DFI − VGG16 using VGG16 as the feature extractor. Experimental setting and metrics were identical to those for Fig. 8

steps) using LPIPS, FID and the user study. Surprisingly, the original DFI produces more accurate reconstructions than the DFI − VGG16 in both qualitative and quantitative comparisons. DFI − VGG16 results are sharp and realistic, similar to the proposed DFI alone approach. However, considering semantic similarity, the original DFI can restore unique attributes, e.g. mustache, race, age, etc., better than the DFI − VGG16. Although LPIPS and FID scores from the two methods are quite close, the original DFI significantly outperforms DFI − VGG16 in user study results.

Although the pre-trained VGG16 is a powerful feature extractor in general, the deep generalized strong feature extractor might not outperform the shallow but data specific and well-designed feature extractor for inference mapping using the specific training dataset (CelebA). Most impor-

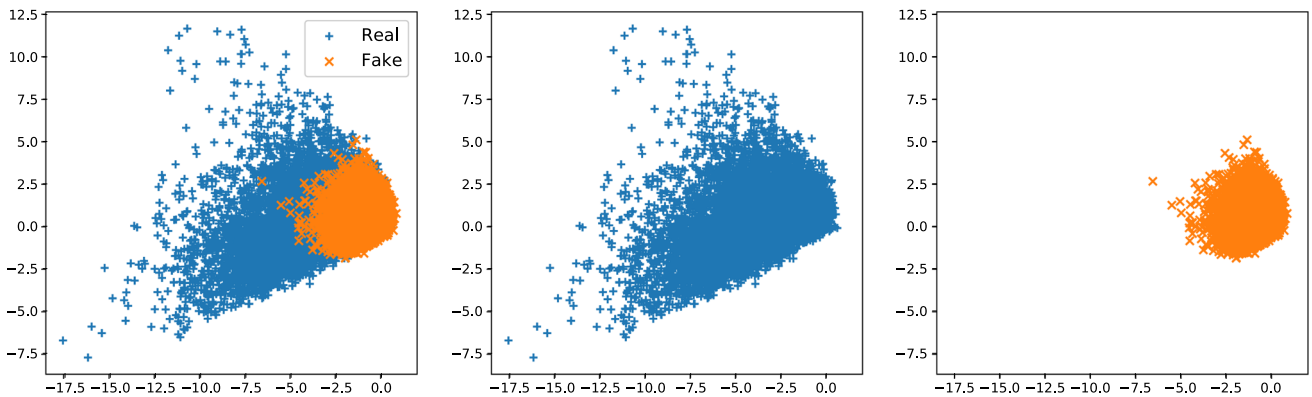

Fig. 16 Visualizations for $D^f(x)$ and $D^f(G(z))$ using two most significant principal component axis projection. Columns (2) and (3) show real and fake samples separately, respectively, with the same axis scale as first column to more easily visualize the overlap area (Color figure online)

tantly, the pre-trained classifier never experiences the GAN training dataset, and hence cannot exploit training data characteristics. If the VGG16 model was finetuned with GAN training data, we would expect it to exhibit more accurate inference mapping. However, that would be beyond the scope of the current paper because VGG16 already requires many more parameters than the proposed DFI approach. Our purpose was to show that DFI was as powerful as VGG16 although requiring significantly less computing resources without additional overheads required for feature extraction. Quantitative comparisons confirm that the original DFI (utilizing discriminator features) performs better than the DFI − VGG16 (utilizing VGG16 features) when the same training iterations are set. Thus, the original DFI is more efficient than the DFI − VGG16 for inference mapping.

One might consider that discriminator feature D^f distributions for real and fake images should not overlap because the discriminator objective is to separate fake images from generated and real images. The distributions may not overlap if the discriminator was trained in a stationary environment or the discriminator defeats the generator, i.e., the generator fails. However, the proposed approach simultaneously trains the generator to deceive the discriminator, hence the GAN training is not stationary. Therefore, if the generator is successfully trained, the generated sample distribution will significantly overlap the real sample distribution, i.e., the generator produces realistic samples. Ideally, training is terminated when the discriminator cannot tell the difference between real and fake images, but for practical GANs, the discriminator is not completely deceived.

Suppose the generator produces highly realistic fake samples, indistinguishable from real samples. Then D^f for fake samples will significantly overlap with D^f for real samples. If the generator is not performing well, e.g. under-training, or small network capacity, D^f for real and fake samples will not overlap because the discriminator defeated the generator. However, in this situation GAN training fails, i.e., none of the inference algorithms can reconstruct the given image.

To empirically show that D^f for real and fake images overlap, Fig. 16 projects D^f on to the two most significant principal component axes using the LSGAN discriminator. The D^f for real (blue) and fake images (orange) have significant overlap, with the real sample distribution having wider coverage than for the fake samples due to limited diversity, i.e., mode collapse. Therefore, the discriminator offers a meaningful feature extractor for both real and fake images.

4.7 Toward a High Quality DFI

To improve inference mapping accuracy, we modified the DFI by selecting the layer for extracting discriminator features D^f; and increasing the connection network capacity. We first introduce a method to improve D^f by using a middle level discriminator feature, improving DFI accuracy. Then we investigated inference accuracy with respect to connection network capacity, confirming that higher connection network capacity does not degrade DFI accuracy.

Since the discriminator feature is extracted from the last layer of the discriminator, it corresponds to a large receptive field. This is advantageous to learn high level information, but incapable of capturing low level details, such as wrinkles, curls, etc. For reconstruction purposes, this choice is clearly disadvantageous to achieve high quality reconstruction. To resolve this limitation, we transfer knowledge from the intermediate feature map discriminator to the connection network.

In particular, we calculated global average pooling (GAP) (Zhou et al. 2016) for the intermediate feature map as the compact representation for the intermediate feature map to achieve computational efficiency. We then concatenated GAP outputs extracted from specific layers of the discriminator with the last discriminator feature. We utilized SNGAN architecture (Miyato et al. 2018) for the experiments.

Table 4 shows the network architecture and feature map names, Table 5 shows LPIPS and FID scores for several combinations of extracted GAP layers, and Fig. 17 shows

Table 4 Network architecture for SNGAN, using architecture is based on Table 3 of Miyato et al. (2018).

Generator	
$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$	
Fully connected layer → 16 × 16 × 512	
4×4, stride=2 transposed conv. BN 256 ReLU	
4×4, stride=2 transposed conv. BN 128 ReLU	
4×4, stride=2 transposed conv. BN 64 ReLU	
3×3, stride=1 conv. 3 Tanh	
Discriminator	
	RGB images $\in \mathbb{R}^{128\times128\times3}$
`Actv128-0`	3×3, stride=1 conv 64 Leaky ReLU
`Actv64-0`	4×4, stride=2 conv 64 Leaky ReLU
`Actv64-1`	3×3, stride=1 conv 128 Leaky ReLU
`Actv32-0`	4×4, stride=2 conv 128 Leaky ReLU
`Actv32-1`	3×3, stride=1 conv 256 Leaky ReLU
`Actv16-0`	4×4, stride=2 conv 256 Leaky ReLU
`Actv16-1` (last feature)	3×3, stride=1 conv 512 Leaky ReLU
	Fully connected layer → 1

Table 5 Reconstruction performance for combinations of extracted layers of GAP (top)

DFI	`Actv128-0`	`Actv64-1`	`Actv32-1`	LPIPS	FID
✓				0.1931	8.91
✓	✓			0.1920	10.34
✓		✓		0.1903	**8.54**
✓			✓	0.1917	8.83
✓	✓	✓		0.1902	9.78
✓		✓	✓	0.1897	9.35
✓	✓	✓	✓	**0.1887**	10.32

LPIPS and FID scores are the average of the best performance among 50K training iterations with 10K steps each trial
Bold values indicate Top-1 scores for each metric

Fig. 17 Qualitative comparison between DFI and DFI with GAP. Column (1) includes target (real) images, (2) and (3) include images reconstructed by DFI and DFI with GAP on the `Actv64-1` layer, respectively

several reconstruction examples for DFI with the GAP layer. Reconstructions from DFI with the GAP layer preserve more attributes attribute, e.g. expressions, eyeglasses, etc. When utilizing features from a single layer, we found that applying `Actv64-1` produced the best accuracy in terms of both LPIPS and FID. Combining features from multiple layers, accuracy (LPIPS) increases with increasing number of combinations, whereas FID decreases. Considering fidelity, quality, and computational efficiency, we suggest applying `Actv64-1` to obtain additional accuracy.

Although the GAP requires low computational cost, spatial information about the feature is completely missing because GAP reduces the feature map spatial dimension 1×1. Therefore, we should consider average pooling layer variants, considering feature map spatial information. To this end, we designed an average pooling to output (R, R, C) feature map, with $R \times R$ final feature map resolution and C is the channel dimension for the intermediate feature map. Larger R preserves more feature map spatial information, and it is equivalent to GAP when $R = 1$, i.e., $1 \times 1 \times C$. We used the `Actv64-1` layer in this experiment, since that provided the highest score in single layer combination as well as the FID score.

Table 6 shows LPIPS and FID scores corresponding to the average pooling layer using the final $R \times R$ resolution feature map. Thus, average pooling preserving spatial information can empirically improve both fidelity and quality compared with GAP. However, both scores increase when $R > 4$. We suggest this is due to the large number of parameters, which leads to DFI overfitting the training data.

The DFI modeling power solely depends on the connection network capacity because both the generator and discriminator are fixed when training the connection network. Training high capacity networks commonly suffer from overfitting with limited datasets. Therefore, the proposed inference algorithm may also experience overfitting on training data if the high capacity model was selected for the connection network. Fortunately, in the training scenario

Table 6 Reconstruction performance with respect to average pooled feature map size, R

DFI	Average pooling	LPIPS	FID
✓	1×1	0.1903	8.54
✓	2×2	0.1894	8.29
✓	4×4	0.1835	8.26
✓	8×8	0.1858	8.77

Average pooling was conducted on feature map `Actv64-1`. 1×1 denotes GAP and $R \times R$ denotes average pooling with the feature map. FID scores are average of the best performance among 50 K training iterations with 10 K steps each trial

Table 7 Reconstruction performance with respect to number of FC layers in the connection network

No. FC layers	LPIPS	FID
1	0.1942	8.96
2 (default)	0.1931	8.91
3	0.1943	8.92
4	0.1948	8.85
5	0.1946	8.73

LPIPS and FID scores are average score of best performance among 50K training iterations with 10K steps each trial

using the proposed latent reconstruction loss, we can utilize unlimited training samples because their seed, i.e., a latent code, can be drawn from a continuous prior distribution and their images can be created by the generator. Thus, regardless of network capacity, we will have sufficient training data to avoid overfitting. Consequently, the network capacity (provided it includes more than two FC layers) does not affect inference mapping accuracy.

To verify this, we investigated inference accuracy with respect to the number of connection network layers, i.e., connection network capacity. The default setting for other experiments reported here was two FC layers. Table 7 summarizes LPIPS and FID scores for various numbers of FC layers in the connection network. Thus we experimentally verify that connection network complexity does not significantly influence inference accuracy.

4.8 SCGAN Experimental Results

We verified spatially conditioned image generation feasibility using the proposed SCGAN approach for the CelebA (Liu et al. 2015) and cat head (Zhang et al. 2008) dataset. All experiments set center image size (input) = (64, 64, 3) and full image size = (128, 128, 3). We assigned the input patch location to the middle left for the CelebA dataset and top left for the cat head dataset. Latent vector dimension = 128 for z_{full} and 64 for both z_{center} and z_{edge}. SCGAN baseline architecture was built upon SNGAN (Miyato et al. 2018), where only spectral normalization was applied to the discriminator. Throughout all SCGAN experiments, we used hyperparameter $\alpha = 10$ for L^{recon} and $\gamma = 10$ for L^{GP}.

Two evaluation criteria were employed for spatially conditioned image generation: reconstruction accuracy and generation quality. To assess reconstruction quality we adopted

Fig. 18 Spatially conditioned image generation for the CelebA dataset using the proposed SCGAN approach. Row (1) includes input images (inside box) and outer images (outside box), (2) and (3) include SCGAN generated images

Fig. 19 Spatially conditioned image generation using the proposed SCGAN approach for the cat head dataset. Row (1) includes input images (inside box) and their original outer images (outside box), (2) and (3) include SCGAN generated images

Fig. 20 Proposed SCGAN compared with PICNet state-of-the-art image completion technique. Column (1) and (4) include input images, (2), (3), (5), and (6) include PICNet and SCGAN generations, respectively

LPIPS and FID. First, we measured LPIPS and FID scores between x_{center} and y_{center}, reconstructed by the proposed DFI inference algorithm, using 10k test images from CelebA and 1k test images from the cat head dataset. These scores, (0.1673, 31.24) and (0.1669, 32.64), respectively, served as the baseline for SCGAN reconstruction quality. We then calculated both scores between x_{center} and y_{crop} (reconstructed by SCGAN), achieving (0.1646, 31.70) and (0.1653, 33.03) respectively, which are comparable with the baseline LPIPS and FID scores. Hence SCGAN reconstruction ability is similar to the proposed inference algorithm.

To qualitatively assess generation quality, we examined whether generated images were diverse, semantically consistent with the reconstructed image, and visually pleasing. Figures 18 and 19 show example spatially conditioned images using SCGAN. Row (1) includes input images (inside box) with their surrounding regions, and rows (2) and (3) include various image generation results from the same input, i.e., the same input latent vector, $\hat{z}_{center}$, but with a different z_{edge} latent vectors. Figure 18 shows six generated results for different z_{edge} are clearly different from each other, presenting various facial shapes, hairstyles, or lips for the same

input. However, all reconstructions have acceptable visual quality and match input image semantics well in terms hair color, skin tone, or eye and eyebrow shape. Figure 19 shows four generated cat head dataset reconstructions with similar tendencies to CelebA results. Each cat has a different face shape, hair color, and expression, with reasonable visual quality. However, the input is correctly reconstructed, and the generated surroundings are semantically seamless with the input.

Thus, SCGAN successfully controlled spatial conditions by assigning input position, producing various high quality images.

Finally, we compared the proposed approach with the PICNet state-of-the-art image completion technique (Zheng et al. 2019) under the same conditions, as shown in Fig. 20. SCGAN can generate realistic entire faces, whereas PICNet cannot maintain consistent quality across the entire image region. This is due to the surrounding regions requiring extrapolation, whereas PICNet image completion is designed to solve image interpolation. Unlike various image completion models such as PICNet, SCGAN possesses the strong generation capability of GANs, producing the images from the latent codes, despite it can faithfully keep the input patch by utilizing inference mapping. As a result, SCGAN solves image extrapolation, which is not possible by previous image completion models.

5 Conclusion

This study proposed an acyclic inference algorithm to improve inference accuracy with minimal training overhead. We introduced discriminator feature based inference (DFI) to map discriminator features to the latent vectors. Extensive experimental evaluations demonstrated that the proposed DFI approach outperforms current methods, accomplishing semantically accurate and computationally efficient inference mapping.

We believe the accuracy gain is achieved by the well-defined objective function, i.e., latent reconstruction loss; and the powerful feature representation from the discriminator. The computational problem was simplified into deriving the mapping from low dimensional representation to another low dimensional representation by adopting discriminator features. Consequently, the proposed approach also provides computational efficiency in training by significantly reducing training parameters.

We also introduced a novel conditional image generation algorithm (SCGAN), incorporating the proposed DFI approach. SCGAN can generate spatially conditioned images using accurate semantic information inferred from the proposed inference mapping. We experimentally demonstrated that spatial information about the image can be used as a conditional prior, in contrast to traditional priors, e.g., class labels or text. We expect the proposed model architecture can be extended to solve image extrapolation and editing problems.

Acknowledgements This research was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Korean Government (Grant NRF-2019R1A2C2006123), the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2016-0-00288) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), and also by ICT R&D program of MSIP/IITP. [R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding].

References

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning* (pp. 37–49).

Bang, D., & Shim, H. (2018). Improved training of generative adversarial networks using representative features. In *International conference on machine learning*.

Berthelot, D., Schumm, T., & Metz, L. (2017). *Began: Boundary equilibrium generative adversarial networks*. arXiv preprint arXiv:1703.10717.

Brock, A., Donahue, J., & Simonyan, K. (2018). *Large scale GAN training for high fidelity natural image synthesis*. arXiv preprint arXiv:1809.11096.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Donahue, J., Krähenbühl, P., & Darrell, T. (2017). Adversarial feature learning. In *International conference on learning representations*.

Dowson, D., & Landau, B. (1982). The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, *12*(3), 450–455.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., et al. (2017). Adversarially learned inference. In *International conference on learning representations*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein GANs. In *Advances in neural information processing systems* (pp. 5769–5779).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). *Squeezenet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 mb model size*. arXiv preprint arXiv:1602.07360.

Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, *36*(4), 107.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5967–5976). IEEE.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *International conference on learning representations*.

Krizhevsky, A. (2014). *One weird trick for parallelizing convolutional neural networks*. arXiv preprint arXiv:1404.5997.

Krizhevsky, A., & Hinton, G. (2009). *Learning multiple layers of features from tiny images*. Technical report. Citeseer.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2015). *Autoencoding beyond pixels using a learned similarity metric*. arXiv preprint arXiv:1512.09300.

Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., & Carin, L. (2017). Alice: Towards understanding adversarial learning for joint distribution matching. In *Advances in neural information processing systems* (pp. 5495–5503).

Liu, M. Y., & Tuzel, O. (2016). Coupled generative adversarial networks. In *Advances in neural information processing systems* (pp. 469–477).

Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., & Wen, S. (2019). STGAN: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3673–3682).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730–3738).

Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2018). Are GANs created equal? A large-scale study. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31, pp. 700–709). Red Hook: Curran Associates, Inc.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2016). Adversarial autoencoders. *International conference on learning representations*.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Smolley, S. P. (2017). Least squares generative adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 2813–2821). IEEE.

Mescheder, L., Geiger, A., & Nowozin, S. (2018). Which training methods for GANs do actually converge? In *International conference on machine learning* (pp. 3478–3487).

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). *Spectral normalization for generative adversarial networks*. arXiv preprint arXiv:1802.05957.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International conference on learning representations*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference learning representations*.

Srivastava, A., Valkoz, L., Russell, C., Gutmann, M. U., & Sutton, C. (2017). Veegan: Reducing mode collapse in GANs using implicit variational learning. In *Advances in neural information processing systems* (pp. 3310–3320).

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, *1*(1–2), 1–305.

Warde-Farley, D., & Bengio, Y. (2017). Improving generative adversarial networks with denoising feature matching. In *International conference on learning representations*.

Wu, Y., & He, K. (2018). Group normalization. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19).

Xiao, H., Rasul, K., & Vollgraf, R. (2017). *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms*. arXiv preprint arXiv:1708.07747.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2018a). *Self-attention generative adversarial networks*. arXiv preprint arXiv:1805.08318.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018b) The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586–595).

Zhang, W., Sun, J., & Tang, X. (2008). Cat head detection-how to effectively exploit shape and texture features. In *European conference on computer vision* (pp. 802–816). Berlin: Springer.

Zheng, C., Cham, T. J., & Cai, J. (2019). *Pluralistic image completion*. arXiv preprint arXiv:1903.04227.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929).

Zhu, J. Y., Krähenbühl, P., Shechtman, E., & Efros, A. A. (2016). Generative visual manipulation on the natural image manifold. In *European conference on computer vision*. Berlin: Springer.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.