Contextual Distortion Information Compensation for GAN inversion image editing

Ziqiang Zhang[1], Yeyu Zhu[1], Shang Wang[2], Xiuhuai Xie[1], Hailong Li[1]

[1]Class of School of Informatics
[2]Class of Artificial Intelligence Research Institute
23020231154258, 23020231154266, 31520231154270, 23020231154238, 23020231154145

Abstract

Cross-modal generation, involving the transformation of information across different modalities while preserving semantic consistency, has been a focus of research. Text-to-image generation, in particular, presents a significant challenge due to the abstract nature of text descriptions. This study addresses the task of text-driven image attribute editing within the context of Generative Adversarial Networks Inversion. The primary aim is to develop a method that faithfully reconstructs images while allowing for flexible and precise attribute editing driven by textual descriptions.We proposes an innovative approach that aims to overcome the Distortion-Edit trade-off dilemma. The key idea is to leverage contextual information completion during image reconstruction. the study introduces two key components: the Contextual Distortion Information Prediction Network and a Context Information Fusion module. The former predicts rich contextual and geometric information between two images, while the latter uses a gating mechanism to integrate distortion information into the image generation process. Ensuring both editability and significantly improved image fidelity. The research also leverages the latest Contrastive Language-Image Pre-training (CLIP) model, specifically the StyleCLIP variant, to modify input latent vectors and achieve attribute editing aligned with user-provided text descriptions.

Introduction

Visual mental imagery plays a crucial role in human cognitive processes, impacting memory, spatial navigation, and reasoning. The rise of deep learning, particularly the emergence of Generative Adversarial Networks (GANs), has revolutionized computer vision and image processing (Goodfellow et al. 2014). One of the applications of GANs is image synthesis, a technique that allows the generation of new images or modifications to existing ones, providing extensive creative opportunities in fields such as art, design, and game development. GANs have been successfully employed in high-resolution face synthesis, image super-resolution, restoration of images with an oil-painting effect, style transfer, image-to-image translation, and representation learning, among others.

However, existing GAN inversion methods either perform per-image optimization for higher reconstruction quality (Abdal, Qin, and Wonka 2019; Kang, Kim, and Cho 2021), which may lead to deviations from the GAN manifold and reduced editing quality, or utilize encoder-based approaches for faster inference and superior editing performance (Alaluf et al. 2022; Richardson et al. 2021) but often sacrifice reconstruction accuracy and fidelity. These methods can capture a coarse layout (low-frequency patterns) but tend to overlook image-specific details (high-frequency patterns), resulting in distortions. For example, reconstructed face images may exhibit prevalent patterns found in the training data, like typical poses or expressions, while details appearing infrequently in the training data, such as background, lighting, or accessories, are distorted. Preserving image-specific details with high fidelity is crucial for reconstruction and editing.

Although efforts have been made to enhance the reconstruction accuracy of encoder-based methods, their editing performance often degrades (Alaluf et al. 2022). From one perspective, the GAN inversion problem can be viewed as a lossy data compression system with a fixed-parameter decoder, which involves a trade-off between universal information and the retention of image-specific details. Consequently, balancing this trade-off is essential.

With this background in mind, this paper introduces a novel Contextual Distortion Information Complementary GAN Inversion for Image Editing (CDIC) approach. CDIC enhances image fidelity and perceptual quality while retaining editable attributes for image editing. Specifically, CDIC leverages contextual distortion information between the original image and the initial reconstruction to supplement detailed knowledge. A Contextual Distortion Information Prediction (CDIP) network is designed, utilizing a weighted stacked hourglass structure with spatial attention mechanisms to encode contextual information from the image. The CDIP network accurately predicts distortion information between the original and reconstructed images.

Related Work

GAN Inversion. The existing GAN inversion methods can be classified into optimization-based methods, encoder-based methods and hybrid methods. Optimization methods can achieve higher reconstruction quality, but the inference speed is slow. (Abdal, Qin, and Wonka 2019) used ADAM to solve the optimization problem. Minyoung et al used covariance matrix adaptation for gradient-free optimization. Instead of per-image optimization, Junyan Zhu et al. learned an encoder to transform the image. pSp (Richardson et al. 2021) and GHFeat (Xu et al. 2021) proposed to embed latent codes in a hierarchical manner. Furthermore, e4e (Tov et al. 2021) analyzes the trade-off between reconstruction and editing capabilities. (Wei et al. 2021) improved the inversion efficiency by a shallow network with efficient heads. (Alaluf, Patashnik, and Cohen-Or 2021) projected latent code with iterative refinement. These methods are more efficient, but do not allow for high-fidelity reconstruction. Hybrid methods make a compromise. Junyan Zhu et al. initialize the optimization with the encoder output for speedup. Guan et al. designed a collaborative learning scheme for the encoder and optimization iterator. Daniel et al. fine-tuned the StyleGAN parameters for each image after predicting the initial latent code, which takes several minutes for a single image. Compared to previous methods, the approach in this paper greatly improves the reconstruction quality of the encoder model and does not increase inference time.

GAN inversion methods can also be categorized according to the latent space used. Z-space (Karras, Laine, and Aila 2019) is simple but has feature entanglement. The W (Karras, Laine, and Aila 2019) and $W+$ (Abdal, Qin, and Wonka 2019) spaces in Style GAN are more de-entangled, where the $W+$ space extends the W space by using a different additional latent code.The S space (Wu, Lischinski, and Shechtman 2021) is proposed by transforming the $W+$ by affine layers.The P space (Zhu et al. 2020) inverts the image to the last activation layer in the nonlinear mapping network. In addition to StyleGAN, some works (Gu, Shen, and Zhou 2020) also used multiscale latent codes for ProgressGAN (Karras et al. 2017). However, these latent spaces inevitably lose details when reconstructing the image due to bit rate constraints. For high-fidelity inversion, this paper proposes contextual information complementation to convey information specific to the high-frequency domain of an image.

Latent Space Editing. A number of supervised and unsupervised methods explore vector operations for semantic orientations in the GAN latent space. Supervised methods require off-the-shelf attribute classifiers or images annotated with specific attributes. InterfaceGAN (Shen et al. 2020) trains SVMs to learn the boundary hyperplane for each binary attribute. StyleFlow (Abdal et al. 2021) learns invertible mappings via normalized flows and off-the-shelf classifiers. Other work (Jahanian, Chai, and Isola 2019; Plumerault, Borgne, and Hudelot 2020) explores simple geometric transformations through self-supervised learning. The unsupervised approach does not require a pre-trained classifiers. GANspace (Härkönen et al. 2020) performs PCA on early feature layers. similarly, SeFa (Shen et al. 2020) performs feature vector decomposition on affine layers. Some works (Lu et al. 2020; Voynov and Babenko 2020) find distinguishable attribute directions based on mutual information. LatentCLR (Yüksel et al. 2021) explored attribute directions through contrast learning.

Recent work has demonstrated the existence of a distortion-editability trade-off: inverting images into well-behaved regions of StyleGAN's latent space yields better editability. However, these regions are typically less expressive, resulting in a reconstructed image that is less faithful to the original. Img2Style proposes that extending the input latent code from the W space to the $W+$ space can achieve higher fidelity, but accordingly, the editability of the image is greatly reduced.

Method

Overview

We propose utilizing contextual information to enhance the network's understanding of the relationship between local and global image features, aiming to reduce artifacts. Our approach introduces a new encoding process consisting (as shown in Figure 1) of two main components: Contextual Distortion Information Prediction (CDIP) and Contextual Distortion Information Fusion (CDIF). In contrast to conventional encoder-based methods that employ a single input, we utilize both the original image I and the initial generated image R_o from the pre-trained e4e model as inputs.

During the inversion process, CDIP leverages the geometric information from two images and incorporates contextual information from the original image to derive distortion information D. This distortion information represents the fine-grained details lost during the e4e encoding process while considering the contextual relationships. The obtained distortion information is subsequently fed into CDIF to generate the latent code w, and be fused with the feature of the generator.

The same process applies to attribute editing, except the initial edited image E_o is used instead of R_o obtained through pre-trained e4e .

Contextual distortion information prediction

The network can be conceptualized as two hourglass structures with multi-scale learning, enabling the acquisition of information at different levels of granularity. The learning process involves extracting image features from fine to coarse and then from coarse to fine. The larger-sized layers contain a greater amount of information and can capture localized image attributes such as texture and details, whereas the smaller-sized layers capture holistic information about the entire image, including background and layout.

Context hourglass The network's first component, called the contextual hourglass, is used to encode the input images I and R_o and extract contextual information. It is represented by the blue section preceding the yellow cube in Figure 1. Both images are fed into the contextual hourglass, resulting in output image features of size $(B, 24, H, W)$, where B represents the batch size, H denotes the image height, and W represents the image width.To form the geometric feature G, the features of the two input images are stitched together. G has dimensions (B, 1, 48, H, W). Regarding contextual information, we utilize image I as the base and extract three different scales of contextual information (C_1, C_2, C_3) during the encoding process. These contextual information scales are then fed into CDIF, which incorporates the contextual information into the fusion process.

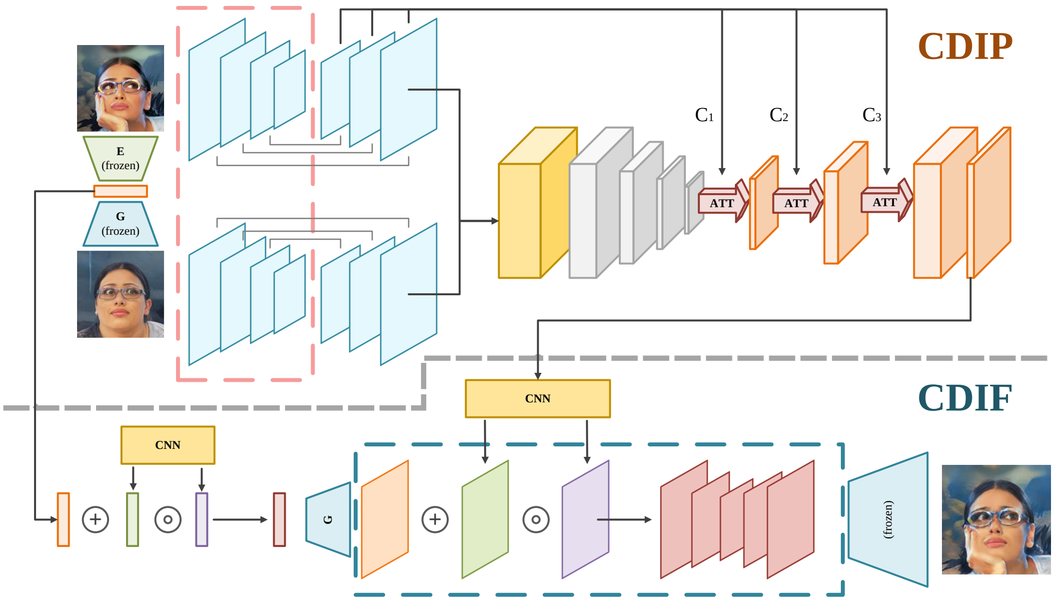

Figure 1: The pipeline of our method. Firstly, the original image I and the initial inverse image R_o (obtained through pre-training e4e) are input into CDIP. CDIP incorporates spatial attention mechanism to integrate contextual and geometric information, generating the distortion information D. Subsequently, CDIF combines the initial latent code wo with D, obtained from e4e, to form a new input w fed into the generator G. Additionally, D is fused in an early layer of the generator.

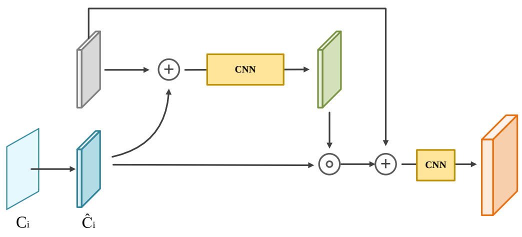

Figure 2: Integrate contextual and geometric information. We apply the spatial attention mechanism to process the contextual information. The processed contextual and geometric information are then combined through summation and convolution operations.

It is important to note that we also need to compensate for the lost information during attribute editing. After attribute editing, the initial edited image E_o may deviate from the original image I. This deviation can impact CDIP's ability to capture the lost detail information of the image accurately. To enable the network to handle this deviation, we share the weights of the first half of the contextual hourglass. This allows the network to learn information from both images simultaneously, facilitating its understanding of the distortion and correction process.

Geometry hourglass We employ a geometric hourglass to decode the geometric information from the geometric feature G obtained through the Context hourglass. The decoding process involves aggregating the image features and then performing upsampling to obtain high-resolution geometric features.To increase the perceptual field of the features and reduce computational effort, we utilize three downsampling modules. Each downsampling module consists of two 3D convolution layers with size of 3x3x3. These downsampling modules help to expand the sensory field to aggregate image features. The resulting geometric features after downsampling are denoted as $G_1 \in \mathbb{R}^{B\times 8\times 48\times H/8\times W/8}$.To decode the high-resolution geometric features, we employ alternating Att (attention) and upsampling modules. Each upsampling module consists of a 3D transposed convolution with a size of 4x4x4, followed by two 3D convolutions with a size of 3x3x3 and. This upsampling process helps restore the fine details in the geometric features.

In order to leverage contextual information to enhance the accuracy of detailed information, we incorporate a spatial attention mechanism (Woo et al. 2018) for feature fusion rather than directly adding the geometric and contextual features together, as shown in Figure 2. The spatial attention weights enable the adaptive selection of ”important” regions for the fusion of geometric and contextual features. The fusion process involves first summing the extended $\widehat{C}_i$ (contextual information) and G_i and then convolving them as,

$$W_i = \sigma(f^{5\times 5}(G_i + \widehat{C}_i)), \tag{1}$$

where σ denotes the sigmoid function, and $f^{5\times 5}$ is a convolution operation with a convolution kernel size of 5x5. The attention weights obtained contain information about the locations that need to be emphasized or suppressed. Finally, we fuse the contextual and geometric features as,

$$G_{i+1} = f^{5\times 5}(G_i + W_i \odot \widehat{C}_i). \tag{2}$$

where i denotes the output of the i-th fusion and $\odot$ denotes the Hadamard product.

Contextual distortion information Fusion

The size of w, which is only (18, 256), has a low bit rate, limiting the amount of information it can carry. Therefore, compensating for information solely within w is not sufficient. To address this limitation, we also introduce compensation at the early layers of the generator. The generator consists of 18 layers, and to prevent overfitting, we specifically choose to compensate for distortion information at the 7th layer. This layer has a size of (512, 64, 64) and can accommodate more information. To fuse the distortion information D with the initial latent code w and the generator, we employ an affine transformation process. As depicted in Figure 1, we utilize two convolutional networks to process the distortion information obtained from the CDIP output. These networks are responsible for predicting the scaling parameter γ and the displacement parameter θ, which are used for compensating the information,

$$\begin{aligned} \gamma_w = f_w^g(D), \theta_w = f_w^t(D), \\ \gamma_F = f_F^g(D), \theta_F = f_F^t(D), \end{aligned} \tag{3}$$

where mapping functions f^g and f^t are convolution layers.For the original feature map F from the generator, we apply a channel scaling operation using the scaling parameter γ and then a channel displacement operation using the displacement parameter θ. This process helps filter out unwanted features and complement detailed features, thereby facilitating the generation of high-fidelity features in StyleGAN,

$$\begin{aligned} AFF(w_i|D) = \gamma_{wi} \times w_i + \theta_{wi}, \\ AFF(F_i|D) = \gamma_{Fi} \times F_i + \theta_{Fi}, \end{aligned} \tag{4}$$

where AFF represents the affine transformation. F_i represents the i-th channel of the feature map, while D refers to the tensor containing complementary information. Additionally, γ_i and θ_i represent the i-th scaling and shifting parameters.

However, the affine transform, being a linear transform for each channel, limits the effectiveness of distortion information fusion. To address this limitation, we introduce a CNN with activation functions after the affine transform, introducing nonlinearity into the fusion process. This expands the representation space further and facilitates the fusion of different distortion information and image features.Additionally, the normalization operation converting the feature map to a normal distribution contradicts the affine transform's objective, which aims to increase the distance between different samples. As a result, the normalization process is removed in CDIF, as it does not contribute to the generation process.The final fused w and F can be represented as,

$$\begin{aligned} w_{fused} = CNN_w(AFF(w|D)), \\ F_{fused} = CNN_F(AFF(F|D)). \end{aligned} \tag{5}$$

where $(AFF(w|D)$ and $AFF(F|D)$ denote w and F, respectively, after performing affine transformations on each channel, and CNN denotes convolutional neural network. For w, we design a convolutional layer and an activation function; for F, we design two convolutional layers and an upsampling layer.

Loss

During the training phase, the generator remains frozen, so we focus on training the encoding process. We design the loss functions to address both reconstruction quality and editability. For the reconstruction quality, we define the loss functions for the original image I and the reconstructed image R_f as,

$$\mathcal{L}_{rec} = \mathcal{L}_2 + \lambda_{LPIPS}\mathcal{L}_{LPIPS} + \lambda_{id}\mathcal{L}_{id}, \tag{6}$$

where $\mathcal{L}_2$ loss is utilized to evaluate structural similarity, LPIPS (Zhang et al. 2018) loss is employed to assess perceptual similarity, and $\mathcal{L}_{id} = 1 - < F(I), F(R_f) >$ is utilized to measure identity consistency. Specifically, for the face domain, F represents the pre-trained AceFace (Deng et al. 2019) model, while for other domains, F refers to the pre-trained ResNet-50 (Tov et al. 2021) model. To ensure editability, we also incorporate the $\mathcal{L}_edit$ loss,

$$\mathcal{L}_{edit} = \mathcal{L}_1(w, w_{fused}) + \mathcal{L}_1(F, F_fused), \tag{7}$$

this loss is utilized to regulate the distance between w, F and the w_{fused}, F_{fused} . Incorporating additional information while keeping them close in the latent space helps maintain editability, as suggested in (Tov et al. 2021).Ultimately, the total loss is expressed as,

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{edit}\mathcal{L}_{edit}. \tag{8}$$

Experimentation

Experimental Settings

Datasets We applied CDIC in the face domain and achieved excellent results. We used FFHQ (Karras, Laine, and Aila 2019) for training and tested on CelebA-HQ (Karras et al. 2017; Liu et al. 2015). Besides, we use InterfaceGAN (Shen et al. 2020) and GANSpace (Härkönen et al. 2020) for editing.

Baseline We compare our method with various GAN Inversion methods, including the optimization-based methods I2S (Abdal, Qin, and Wonka 2019) and PTI (Roich et al. 2022) and encoder-based methods. pSp (Chang and Chen 2018), e4e (Tov et al. 2021) and Restyle (Alaluf, Patashnik, and Cohen-Or 2021).

Implementation Details Our experiments use the pre-trained StyleGAN generator and the e4e encoder. The size of the input and output images of the network are both 1024 × 1024. $\lambda_{LPIPS}, \lambda_{ID} and \lambda_{edit}$ in Eq. (6) and Eq. (8) are set to 0.8, 0.2 and 0.5, respectively. We used the ranger optimizer (Yong et al. 2020) with the learning rate set to 0.001 and the batch size set to 2 and trained 100000 steps on a 3080 GPU.

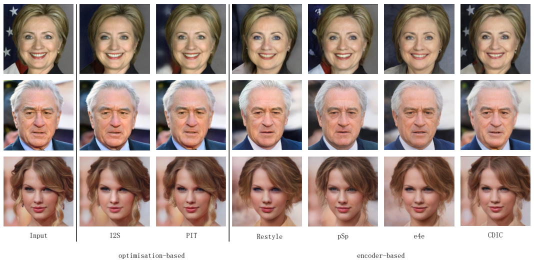

Figure 3: Comparison of reconstruction quality. Our method is comparable to optimisation-based methods and retains more detail.

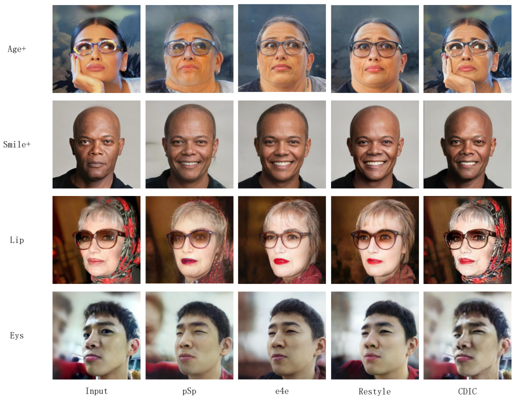

Figure 4: Editing quality comparison. Optimisation-based methods are less editable, so we mainly compare them with encoder-based methods. It can be seen that our method has higher fidelity while retaining editability.

Reconstruction Quality

Quantitative Evaluation We compared our method with the mainstream encoder-based methods. The results of the quantitative comparison of the quality of the inverted images are shown in Table 1. We offer the L_2 distance, the $LPIPS$ (Zhang et al. 2018) distance, and the ID (Richardson et al. 2021) score between each reconstruction and source. Additionally, we perform and compare the network inference time. These metrics are computed on the first 1000 images of CelebA-HQ. As can be seen from the data, the method significantly outperforms both the encoder-based baseline and the optimization-based baseline in terms of reconstruction quality. It is also significantly faster than the optimization-based method in inference.

Qualitative Evaluation Fig. 3 demonstrates the qualitative comparison of CDIC with PTI, Restyle, pSp, and e4e. While optimisation-based techniques often enable accurate reconstruction, they are computationally expensive. CDIC provides visually comparable results, but the inference time is orders of magnitude faster. Compared to one-shot encoders (pSp and e4e), CDIC better captures the input identity (third row). Compared to recent ReStyle encoders, CDIC still better reconstructs finer details such as complex hairstyles (hair on the right ear part of the third row of images) and backgrounds (star-spangled banner in the background of the first row of images).

Method	$L_2 \downarrow$	$LPIPS \downarrow$	$ID \uparrow$	$Time(s) \downarrow$
I2S	0.020	0.09	0.78	156
PTI	0.015	0.09	0.85	283
pSp	0.034	0.17	0.56	0.11
Restyle	0.041	0.19	0.52	0.46
e4e	0.052	0.20	0.050	0.11
CDIC	0.010	0.09	0.87	0.24

Table 1: Quantitative comparison for inversion quality on faces. The horizontal lines in the table delineate the optimisation-based methods, the encoder-based methods and our method

Editing Performance

Qualitative Evaluation Fig. 4 demonstrate a qualitative comparison of the editing performance of SDIC with the baseline methods. The first row gives an image of a face occluded by a hand and the last row shows an image of a face with a large angular rotation. Existing methods cannot faithfully reconstruct these challenging images. They produce distorted results and artefacts in both inversion and editing. In contrast, with the proposed contextual information complementation, our method is more robust and produce high fidelity results. In addition to the improved robustness, CDIC successfully preserves more details such as the background (fourth row), shadows (second row), attachments (third row), and expressions (fourth row).

Conclusion

In this paper, we propose a method to solve the task of text-driven image attribute editing, which is carried out in the context of Generative Adversarial Network inversion. The innovative approach we propose leverages contextual information completion to overcome the Distortion-Edit trade-off dilemma. By introducing the Contextual Distortion Information Prediction Network and a Context Information Fusion module, our method significantly improves image fidelity while maintaining editability. Through empirical verification, our method shows significant improvements in image fidelity and perceptual quality while maintaining editability. Compared with existing GAN inversion methods, our method can better retain specific image details while providing higher reconstruction quality and editing performance. Our method has potential practical applications in various applications, including high-resolution face synthesis, image super-resolution, restoration of oil painting effect images, style transfer, image-to-image translation, and representation learning.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4432–4441.

Abdal, R.; Zhu, P.; Mitra, N. J.; and Wonka, P. 2021. Styleflow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3): 1–21.

Alaluf, Y.; Patashnik, O.; and Cohen-Or, D. 2021. ReStyle: A residual-based StyleGAN encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6711–6720.

Alaluf, Y.; Tov, O.; Mokady, R.; Gal, R.; and Bermano, A. 2022. HyperStyle: StyleGAN inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18511–18521.

Chang, J.-R.; and Chen, Y.-S. 2018. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5410–5418.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4690–4699.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Conference on Neural Information Processing Systems (NeurIPS)*, 27.

Gu, J.; Shen, Y.; and Zhou, B. 2020. Image processing using multi-code GAN prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3012–3021.

Härkönen, E.; Hertzmann, A.; Lehtinen, J.; and Paris, S. 2020. GANspace: Discovering interpretable GAN controls. *Conference on Neural Information Processing Systems (NeurIPS)*, 33: 9841–9850.

Jahanian, A.; Chai, L.; and Isola, P. 2019. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*.

Kang, K.; Kim, S.; and Cho, S. 2021. GAN inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 13941–13949.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4401–4410.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3730–3738.

Lu, Y.-D.; Lee, H.-Y.; Tseng, H.-Y.; and Yang, M.-H. 2020. Unsupervised discovery of disentangled manifolds in gans. *arXiv preprint arXiv:2011.11842*.

Plumerault, A.; Borgne, H. L.; and Hudelot, C. 2020. Controlling generative models with continuous factors of variations. *arXiv preprint arXiv:2001.10238*.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a StyleGAN encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2287–2296.

Roich, D.; Mokady, R.; Bermano, A. H.; and Cohen-Or, D. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1): 1–13.

Shen, Y.; Gu, J.; Tang, X.; and Zhou, B. 2020. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9243–9252.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14.

Voynov, A.; and Babenko, A. 2020. Unsupervised discovery of interpretable directions in the GAN latent space. In *arXiv preprint arXiv:2002.03754*.

Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Zhang, W.; Yuan, L.; Hua, G.; and Yu, N. 2021. A simple baseline for stylegan inversion. *arXiv preprint arXiv:2104.07661*, 9: 10–12.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, Z.; Lischinski, D.; and Shechtman, E. 2021. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12863–12872.

Xu, Y.; Shen, Y.; Zhu, J.; Yang, C.; and Zhou, B. 2021. Generative Hierarchical Features from Synthesizing Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yong, H.; Huang, J.; Hua, X.; and Zhang, L. 2020. Gradient centralization: A new optimization technique for deep neural networks. In *European Conference on Computer Vision (ECCV)*, 635–652. Springer.

Yüksel, O. K.; Simsar, E.; Er, E. G.; and Yanardag, P. 2021. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14263–14272.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595.

Zhu, P.; Abdal, R.; Qin, Y.; Femiani, J.; and Wonka, P. 2020. Improved StyleGAN embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*.