High-Fidelity GAN Inversion for Image Attribute Editing

Tengfei Wang[1] Yong Zhang[2] ✉ Yanbo Fan[2] Jue Wang[2] Qifeng Chen[1] ✉
[1]The Hong Kong University of Science and Technology [2]Tencent AI Lab

Abstract

We present a novel high-fidelity generative adversarial network (GAN) inversion framework that enables attribute editing with image-specific details well-preserved (e.g., background, appearance, and illumination). We first analyze the challenges of high-fidelity GAN inversion from the perspective of lossy data compression. With a low bit-rate latent code, previous works have difficulties in preserving high-fidelity details in reconstructed and edited images. Increasing the size of a latent code can improve the accuracy of GAN inversion but at the cost of inferior editability. To improve image fidelity without compromising editability, we propose a distortion consultation approach that employs a distortion map as a reference for high-fidelity reconstruction. In the distortion consultation inversion (DCI), the distortion map is first projected to a high-rate latent map, which then complements the basic low-rate latent code with more details via consultation fusion. To achieve high-fidelity editing, we propose an adaptive distortion alignment (ADA) module with a self-supervised training scheme, which bridges the gap between the edited and inversion images. Extensive experiments in the face and car domains show a clear improvement in both inversion and editing quality. The project page is `https://tengfei-wang.github.io/HFGI/`.

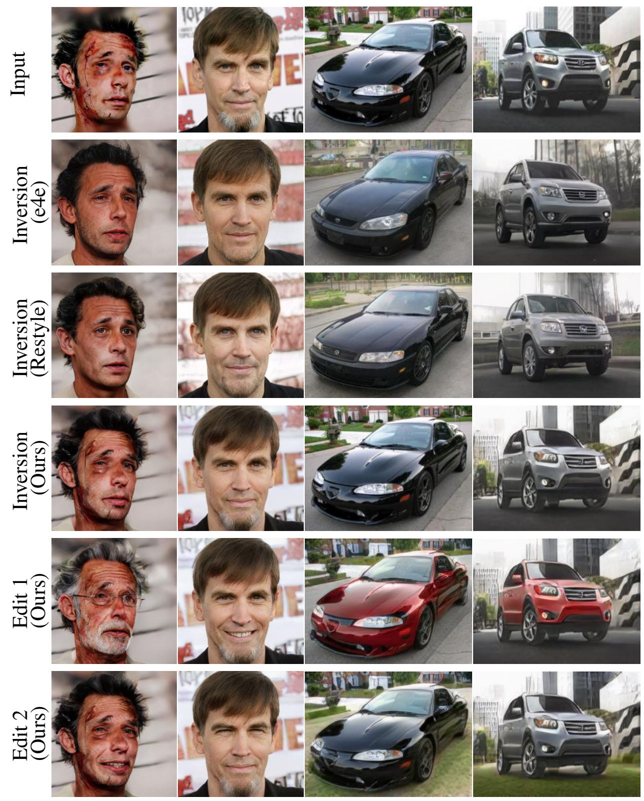


Figure 1. High-fidelity image inversion and editing (age, smile, eyes, color, grass). Our method performs well on details preservation in both inverted and edited results such as background, makeup, beard/hair style, reflection and shadow.

1. Introduction

Image attribute editing is the task of modifying desired attributes of a given image while preserving other details. With the rapid advancement of generative adversarial networks (GANs) [9], a promising direction is to manipulate images with the strong control capacity of StyleGAN [19,20]. To enable real-world image editing, GAN inversion techniques [40] have been recently explored, which aim at projecting images to the latent space of a pre-trained GAN generator.

Existing GAN inversion approaches either perform per-image optimization [1, 17, 45] or learn a data-driven encoder [26, 34]. Optimization approaches achieve higher reconstruction accuracy by over-fitting on a single image, but the latent code may get out of GAN manifold, leading to inferior editing quality. In contrast, encoder-based GAN inversion methods are faster and show better editing performance due to knowledge learned from numerous training images. Nevertheless, their reconstruction results are usually inaccurate and of low fidelity: these methods can reconstruct a coarse layout (low-frequency patterns), but the image-specific details (high-frequency patterns) are often ignored. For example, the reconstructed face images typically possess averaged patterns that agree with the majority of training images (e.g., normal pose/expression, occlusion/shadow-free), and the details that present minority patterns (e.g., background, illumination, accessory) in training data are subject to distortion. It is highly desirable to preserve these image-specific details in reconstruction and editing with high fidelity.

Though some works tried to improve the reconstruction accuracy of encoder-based methods, their editing performance usually decreases [34]. To analyze the limitation of existing approaches, we consider the GAN inversion problem as a lossy data compression system with a frozen decoder. According to Rate-Distortion theory [29], reversing a real-world image to a low-dimensional latent code would inevitably lead to information loss. As conjectured by information bottleneck theory [33], the lost information is primarily image-specific details as the deep compression model tends to retain common information of a domain. Based on these analyses and experimental observations, we present the Rate-Distortion-Edit trade-off for GAN inversion, which further inspires our framework.

According to this trade-off, the low-rate latent codes are insufficient for high-fidelity GAN inversion. However, it is non-trivial to improve the reconstruction accuracy by directly increasing the rate. A higher-rate latent codes can easily achieve a low distortion by overfiting on the reconstruction process, but would suffer a dramatic editing performance drop. To achieve both accuracy and editability (high-fidelity editing), we propose a novel framework that equips low-rate encoder models with **distortion consultation**. The consultation branch serves as a 'cheat sheet' for generation that only conveys the ignored image-specific information. Specifically, we leverage the distortion map between source and low-fidelity reconstructed image as a reference and project it to higher-rate latent maps. Compared with high-rate latent codes inferred from a full image, the distortion map only conveys image-specific details and can thus alleviate the aforementioned overfitting issue. The high-rate latent map and low-rate latent code are further embedded and fused in the generator via **consultation fusion**. Our scheme shows a clear improvement in reconstruction quality, and no test-time optimization is involved.

For attribute editing, following previous works, we perform vector arithmetic [25] on the low-rate latent code, while the consultation is desired to bring back lost details. While the distortion consultation substantially contributes to the inversion quality, it cannot directly apply the distortion map observed on the inversion image for editing due to the misalignment between inverted and edited images. To this end, we additionally design an **adaptive distortion alignment (ADA)** network to adjust the distortion map with the edited images. To disentangle the alignment from the consultation encoder and stabilize the training, we impose intermediate supervision on ADA by proposing an alignment regularization with a self-supervised training scheme.

Extensive experiments show that our method significantly outperforms current approaches in terms of details preservation in both reconstructed and edited results. On account of the high-fidelity inversion capacity, our approach is robust to viewpoint and illumination fluctuation and can thus perform temporally consistent editing on videos. Our primary contributions can be summarized as follows.

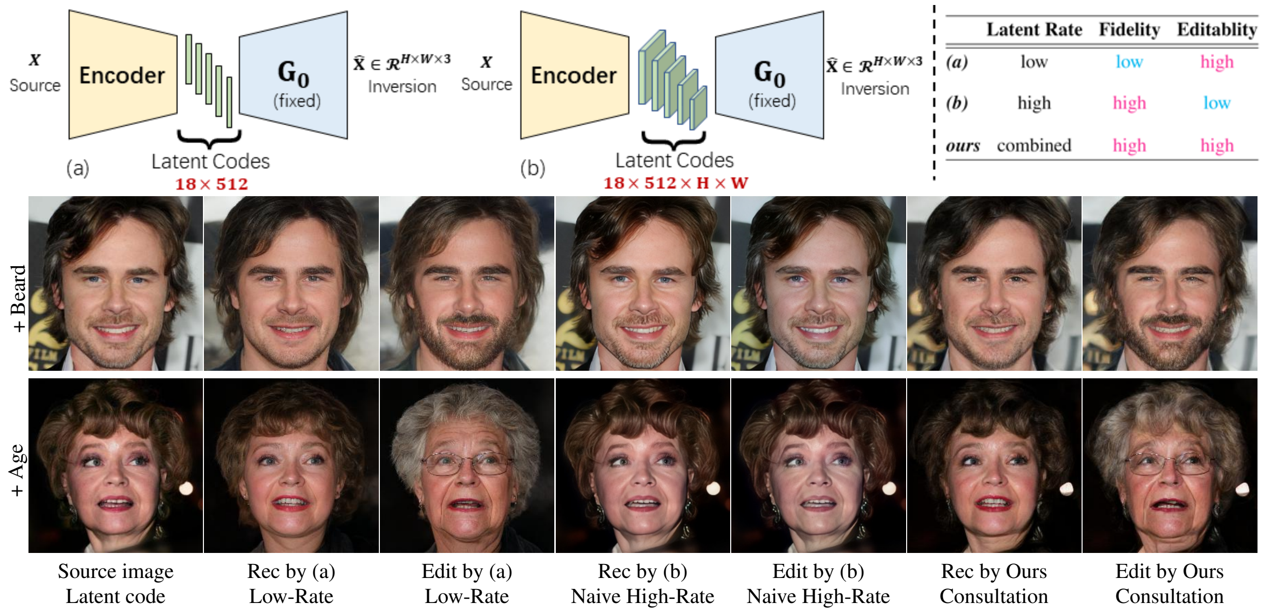
- We propose a distortion consultation inversion scheme that combines both high reconstruction quality and compelling editability with consultation fusion.
- For high-fidelity editing, we propose the adaptive distortion alignment module with a self-supervised learning scheme. By alignment, the distortion information can be propagated well to the edited images.
- Our method outperforms state-of-the-art approaches qualitatively and quantitatively on diverse image domains and videos. The framework is simple, fast and can be easily applied to GAN models.

2. Related Work

2.1. GAN Inversion

Existing GAN inversion approaches can be categorized into optimization-based, encoder-based, and hybrid methods. Optimization approaches can achieve high reconstruction quality but are slow for inference. [45] used L-BFGS, and I2S [1] adopted ADAM for solving the optimization. [15] adopted Covariance Matrix Adaptation for gradient-free optimization. Instead of per-image optimization, [45] learned an encoder to project images. [44] proposed an in-domain method on real images. pSp [26] and GHFeat [41] proposed to embed latent codes in a hierarchical manner. Further, e4e [34] analyzed the trade-offs between reconstruction and editing ability. [38] improved the inversion efficiency by a shallow network with efficient heads. ReStyle [4] projected the latent codes with iterative refinements. These methods are more efficient but fail to achieve high-fidelity reconstruction. Hybrid approaches make a compromise. [45] initialize the optimization with the encoder output for acceleration. [11] designed a collaborative learning scheme for encoder and optimization iterator. [27] fine-tuned StyleGAN parameters for each image after predicting an initial latent code, which takes a few minutes for an image. Compared with previous methods, our method considerably improves the reconstruction quality of encoder models without inference-time optimization.

GAN inversion approaches can also be classified by the used latent space. $\mathbf{Z}$ space [19] is straightforward but suffers from feature entanglement. $\mathbf{W}$ [19] and $\mathbf{W}^+$ [1, 2] space in StyleGAN are more disentangled, where $\mathbf{W}^+$ space extends $\mathbf{W}$ space by using different W across layers. $\mathbf{S}$ space [39] is proposed by transforming $\mathbf{W}^+$ through the affine layers. $\mathbf{P}$ space [46] inverts images to the last ac-



	Latent Rate	Fidelity	Editablity
(a)	low	low	high
(b)	high	high	low
ours	combined	high	high

Figure 2. Rate-Distortion-Edit trade-off. "Rec" and "Edit" represent the reconstruction and editing results, respectively. (a) is a typical low-rate framework for GAN inversion but suffers detail loss and distortion. (b) is a naive high-rate GAN inversion framework with nearly perfect reconstruction but suffers inferior interpretability and editability. The proposed method (Fig. 3) combines both high details fidelity and compelling editing performance with a fast inference speed.

tivation layer in the non-linear mapping network. Besides StyleGAN, some works [10] also adopts multi-scale latent codes for ProgressGAN [18]. Nevertheless, these latent spaces would inevitably lose details in reconstructed images due to limited bit-rate (Sec. 3.1). To perform a high-fidelity inversion, we propose a distortion consultation branch to convey high-frequency image-specific information.

2.2. Latent Space Editing

A number of supervised and unsupervised approaches explored GAN latent space for semantic directions under the vector arithmetic. The supervised methods need off-the-shelf attribute classifiers or annotated images for specific attributes. InterfaceGAN [30] trained SVM to learn the boundary hyperplane for each binary attribute. StyleFlow [3] learned reversible mapping by normalizing flow and off-the-shelf classifiers. Others [16,24] explored simple geometric transformation via self-supervised learning. Unsupervised approaches do not need pre-trained classifiers. GANspace [13] performed PCA on early feature layers. Similarly, SeFa [30] performed eigenvector decomposition of the affine layers. Some [22,35,47] found distinguishable directions based on mutual information. LatentCLR [42] explored directions by contrastive learning.

3. Approach

Given a source image X and a well-trained generator G_0, GAN inversion infers the latent code W via an encoder E_0, which is expected to faithfully reconstruct X. In this section, we first analyze the bottleneck of previous inversion methods and describe our proposed distortion consultation inversion strategy. To handle the features misalignment, we present the adaptive distortion alignment modules with a self-supervised training scheme. The whole framework is illustrated in Figure 3.

3.1. Overview

Motivation. Currently, GAN inversion frameworks lie in three categories, which are optimization-based, encoder-based and hybrid methods. Despite more accurate, optimization-based and hybrid approaches are time-consuming and thus intolerable in real-time applications. Existing encoder-based methods can be illustrated by Fig. 2 (a), where the decoder is a frozen well-trained generator (e.g., StyleGAN) while the encoder learns a mapping from the source image to the latent codes. As observed in many existing works (e.g., results (a) in Fig. 2), the encoder approaches fail to faithfully reconstruct the input images, and the inversion (and editing) results are of low fidelity in terms of details. Noted the fact that the latent codes in previous methods are of (relatively) low dimension, we conjecture that the low-rate latent codes are insufficient for high-fidelity reconstruction. This conjecture is also supported by the Rate-Distortion theory [5,7,29], which will be reviewed in the **Supplement**.

To further analyze the effect of the latent rate in high-fidelity GAN inversion, we formulate the encoder-based GAN inversion as a problem of lossy data compression. In

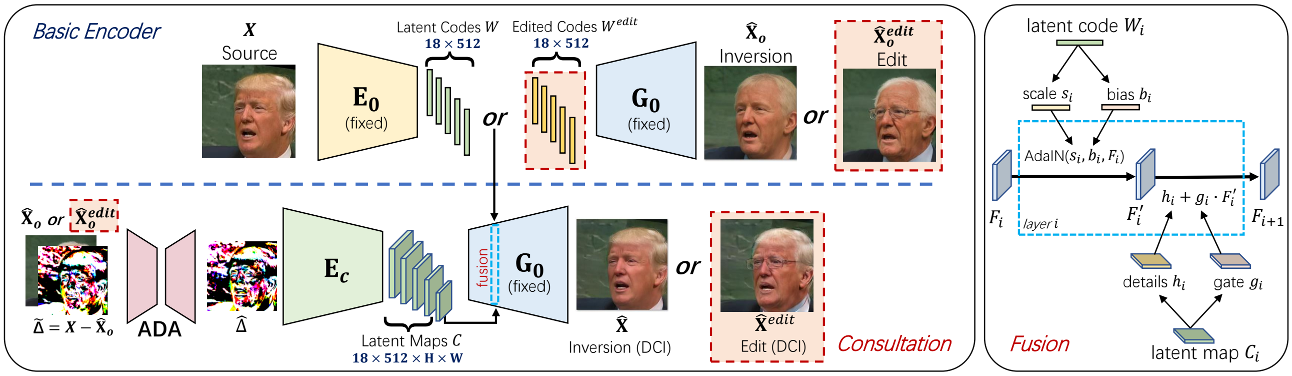


Figure 3. Overview of our high-fidelity image inversion and editing framework. The basic encoder E_0 infers a low-rate latent code W corresponding to a low-fidelity reconstruction image $\hat{X}_o$. The distortion map $\tilde{\Delta}$ between $\hat{X}_o$ and the source image X contains the lost high-frequency image-specific details to improve the reconstruction fidelity. The red dotted boxes indicate the editing behaviour with certain semantic direction N^{edit}, where $W^{edit} = W + \alpha N^{edit}$ corresponds to a low-fidelity editing image $\hat{X}_o^{edit}$. To achieve high-fidelity image editing, we propose the distortion consultation branch to facilitate the generation. In the distortion consultation, $\tilde{\Delta}$ is first aligned with the low-fidelity edited image $\hat{X}_o^{edit}$ by ADA and then embedded to a high-rate latent map C via the consultation encoder E_c. Latent code W and latent map C are combined via the consultation fusion (see details in the right part) across layers of G_0 to generate the final edited image $\hat{X}^{edit}$.

this formulation, *rate* can be interpreted as the dimension of latent codes (e.g., 18×512), and *distortion* indicates the reconstruction quality (fidelity). A compelling inversion method is desired to produce high-fidelity images for both inversion and editing (*low distortion*). Nevertheless, the current dimension of latent codes is much smaller than that of images (*low rate*). This implies a contradiction with [29, 32], which shows the low-rate latent codes are insufficient for faithful reconstruction and some information is inevitably lost. Therefore, we are motivated to design a large-rate GAN inversion system.

Challenge. However, it is non-trivial to reduce the distortion by simply increasing the latent rate. A naive idea for faithful reconstruction is to adopt a higher-rate latent code like Fig. 2 (b). This Unet-like structure is adopted by some recent image restoration works [6, 37] that conveys latent maps (e.g., $18 \times 512 \times H \times W$) to decoder. Benefited from the higher bit rate, the restoration quality is gratifying (e.g., results (b) in Fig. 2). However, we cannot apply this structure in our case since the high-dimensional latent codes are difficult to interpret and manipulate for attribute editing (e.g., results (b) in Fig. 2). Similarly, prior work [34] also observed tradeoffs between the reconstruction and editability brought by over-fitting. The high-rate latent code is easy to overfit on the reconstruction, thereby compromising the edit performance. As the inversion is just an intermediate step to achieve the goal of editing, it is essential to balance the rate, reconstruction, and editing quality, which we call the *Rate-Distortion-Edit* trade-offs (Fig. 2). To this end, a delicate system design is needed.

Design. As analyzed above, with a (relatively) low-rate latent code, the GAN inversion system is subject to inevitable information loss. By analyzing the visual results of previous GAN inversion approaches (Fig. 1, Fig. 2, Fig. 4), we found that these reconstruction results can successfully preserve frequent patterns and principle attributes of the source images. In contrast, the lost information is mostly the image-specific details such as background, make-up and illumination. This observation is consistent with the Information Bottleneck theory [31–33], which hypothesizes the deep models primarily learn common patterns in the dataset while forgetting infrequent details for reconstruction.

Considering the *Rate-Distortion-Edit* trade-offs, now that we have prioritized the editability (with a low-rate latent code), the main concern is how to convey the lost information to improve the fidelity (lower the distortion) without compromising the edit performance. To this end, we propose a distortion consultation branch that only conveys image-specific details to enhance the reconstruction quality, which avoids the trivial solution of a simple overfitting. For editing, we still perform vector arithmetic on the low-rate latent code for its high editability. By combining the best of both worlds, the proposed approach achieves a high fidelity in both reconstruction and editing (Fig. 2).

3.2. Distortion Consultation Inversion (DCI)

Basic Encoder. With a basic encoder E_0, we can obtain a low-rate latent code $W = E_0(X)$ and initial inversion image $\hat{X}_o = G_0(W)$. In this case, the generator G_0 takes W as the input in each layer to obtain the feature map:

$$F_{i+1} = AdaIN(F_i, f_i^s(W_i), f_i^b(W_i)), \tag{1}$$

where $f_i^s(W_i), f_i^b(W_i)$ are affine layers for scale and bias in $AdaIN$ [14]. $\hat{X}_o$ is low-fidelity due to the information loss of low-rate latent codes, and the subscript o denotes an (unsatisfactory) observation of source image X.

Consultation Encoder. To enhance E_0 with higher fidelity, we propose a distortion consultation branch to convey the lost image-specific details. We refer it to ***Consultation***, since the network explicitly consults the image-specific information as a reference for generation. Specifically, we see the distortion map $\tilde{\Delta} = X - \hat{X}_o$ between source X and initial reconstruction $\hat{X}_o$ as the lost details [36]. The distortion map is projected to a high-rate latent map $C = E_c(\tilde{\Delta})$ via the consultation encoder E_c. Compared with prior methods relying on W only, G_0 additionally consults C for lost details to achieve high-fidelity reconstruction as $\hat{X} = G_0(W, C)$.

Consultation Fusion. To combine the consultation branch with the basic encoder for image generation, we adopt a layer-wise consultation fusion for latent codes W and latent maps C, as shown in Fig. 3. As artifacts and inaccurate details introduced by W can degrade the generation quality, we design a gated fusion scheme to adaptively filter out undesired features. In layer i of G_0, C_i is embedded to a gate map g_i and a high-frequency details map h_i:

$$g_i = f_i^{gate}(C_i),\ h_i = f_i^{hf}(C_i), \tag{2}$$

where mapping functions f^{gate} and f^{hf} are convolution layers. h_i contains the image-specific details, and facilitates the low-fidelity features obtained from W_i (Eq. (2)) to produce high-fidelity feature maps F_{i+1} in StyleGAN[1]:

$$F_{i+1} = g_i \cdot AdaIN(F_i, f_i^s(W_i), f_i^b(W_i)) + h_i. \tag{3}$$

To avoid overfitting on the inversion result, we only perform the consultation fusion in early layers of G_0.

3.3. Adaptive Distortion Alignment (ADA)

For attribute editing, the low-rate latent code W would be moved along certain semantic direction N^{edit} as $W^{edit} = W + \alpha N^{edit}$ [30]. The initial edited image by the basis encoder is denoted as $\hat{X}_o^{edit} = G_0(W^{edit})$, which suffers details distortion. So far, we have improved the fidelity of the inversion image $\hat{X}_o$ with the proposed DCI, where the distortion map $\tilde{\Delta} = X - \hat{X}_o$ is calculated for $\hat{X}_o$. However, $\hat{X}_o^{edit}$ would be deformed from $\hat{X}_o$ when editing attributes such as age, pose and expression. This means the observed $\tilde{\Delta}$ may not align with the edited image $\hat{X}_o^{edit}$. Applying DCI directly to $\hat{X}_o^{edit}$ leads to obvious artifacts by consulting misaligned details $\tilde{\Delta}$ (see Sec. 4.3). To advance DCI from inversion to editing, the observed distortion map $\tilde{\Delta}$ is supposed to be adaptively aligned with the edited image $\hat{X}_o^{edit}$. We thus propose the ADA module, which is an encoder-decoder-like structure for distortion alignment.

[1]For StyleGAN2 [20], the fusion layer would be $F_{i+1} = g_i \cdot ModulatedConv(F_i, f_i^s(W_i), f_i^b(W_i)) + h_i$.

Considering a misaligned pair of $\{I, \tilde{\Delta}\}$ where I is $\hat{X}_o$ for inversion and $\hat{X}_o^{edit}$ for editing, ADA is to align the distortion map $\tilde{\Delta}$ with a target image I. For inversion, ADA is ideally an identity mapping. For editing, the distortion map is desired to be adaptively transformed as $\hat{\Delta}^{edit} = ADA(\hat{X}_o^{edit}, \tilde{\Delta})$ that aligns with the initial editing result $\hat{X}_o^{edit}$. With $C^{edit} = E_c(\hat{\Delta}^{edit})$ as a reference, $\hat{X}^{edit} = G_0(W^{edit}, C^{edit})$ can preserve more details.

Self-supervised Training. To alleviate the entanglement between distortion alignment and distortion consultation, involving intermediate supervisions on ADA outputs in preferred. To this end, we need numerous misaligned pairs of $\{I, \tilde{\Delta}\}$ and their ground-truth aligned maps Δ for training, but the data collection is labor-intensive. To conduct a self-supervised training, we take X as the source image, and the low-fidelity inversion $\hat{X}_o$ as the target image I for alignment, and the ground-truth aligned distortion is thus $\Delta = X - \hat{X}_o$. During the training, we augment Δ with random perspective transformation to simulate misaligned distortion maps $\tilde{\Delta}$. We empirically observe these simulated pairs work well on training and expect a better simulation scheme in future works. The ADA module is encouraged to produce the aligned distortion $\hat{\Delta} = ADA(\hat{X}_o, \tilde{\Delta})$ that approximates Δ. The alignment loss is defined as: $\mathcal{L}_{align} = \|\hat{\Delta} - \Delta\|_1$. See the **Supplement** for more details.

3.4. Losses

During the training, the generator G_0 and basic encoder E_0 are frozen. For faithful reconstruction, we calculate L_2 loss and LPIPS [43] between $\hat{X}$ and X. We also calculate the identity loss $\mathcal{L}_{\text{id}} = 1 - \langle F(X), F(\hat{X})\rangle$, where F is pre-trained ArcFace [8] or a ResNet-50 model for different domains [34]. The reconstruction loss is

$$\mathcal{L}_{rec} = \mathcal{L}_2 + \lambda_{per}\mathcal{L}_{LPIPS} + \lambda_{id}\mathcal{L}_{id}. \tag{4}$$

We also impose adversarial loss to improve image quality:

$$\mathcal{L}_D = \mathbb{E}[\log D(\hat{X})] + \mathbb{E}[\log(1 - D(X))], \tag{5}$$

$$\mathcal{L}_{adv} = -\mathbb{E}[\log(D(\hat{X}))], \tag{6}$$

where D is initialized with the well-trained discriminator. In summary, the overall loss is a weighted summation of $\mathcal{L}_{rec}$, $\mathcal{L}_{adv}$, and $\mathcal{L}_{align}$. Note that the training process only involves the inversion images, and no editing direction N^{edit} is needed. After training, the model can generalize to diverse attribute editing explored by different methods.

4. Experiments

4.1. Settings

Datasets. For the human face domain, we use the FFHQ [19] dataset for training and the CelebA-HQ [18] dataset for evaluation. For the car domain, we use Stanford Cars [21] for training and evaluation. For attribute

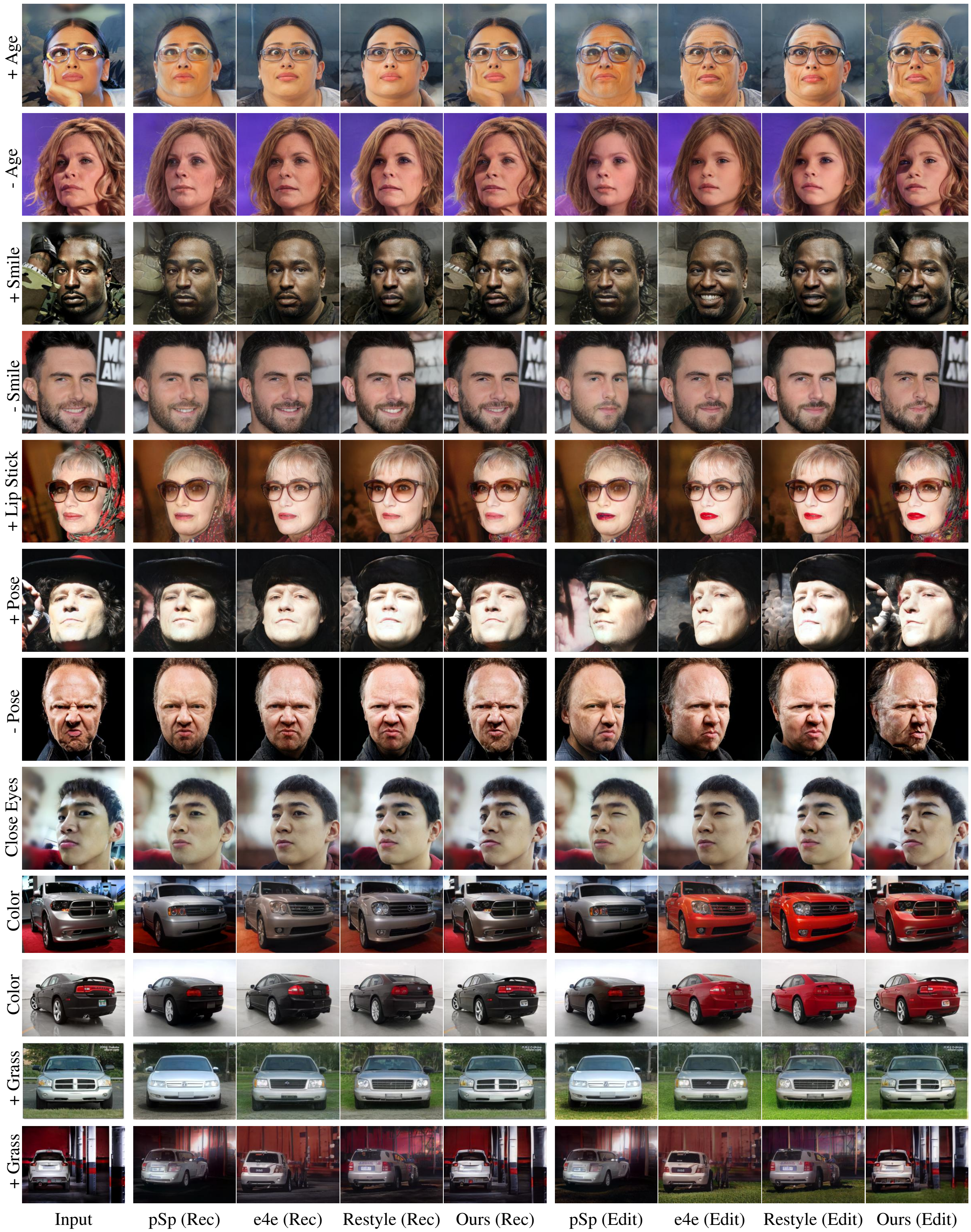


Figure 4. Visual comparisons on Face inversion and editing. More results are shown in **Appendix**.

Method	MAE ↓	SSIM ↑	LPIPS ↓	Time ↓
I2S [1]	.0636±.0010	.872±.005	.134±.006	156s
PTI [27]	.0622±.0004	.877±.003	.132±.003	283s
pSp [26]	.0789±.0006	.793±.006	.169±.002	0.11s
Restyle$_{pSp}$ [4]	.0729±.0005	.823±.004	.145±.002	0.46s
e4e [34]	.0919±.0008	.742±.007	.221±.003	0.11s
Restyle$_{e4e}$ [4]	.0887±.0008	.758±.007	.202±.003	0.46s
Ours$_{e4e}$	**.0617±.0004**	**.877±.002**	**.127±.001**	0.24s

Table 1. Quantitative comparison for inversion quality on faces.

editing, we adopt InterfaceGAN [30] for face images and GANSpace [13] for car images.

Implementation details. See the **Appendix**.

4.2. Evaluation

4.2.1 Quantitative Evaluation

We compare our method (with e4e as the basic encoder) with state-of-the-art encoder-based GAN inversion approaches, pSp [26], e4e [34] and Restyle [4] (with pSp and e4e as backbones, respectively). We report quantitative comparisons of the inversion performance in Table 1. The metrics are calculated on the first 1,500 images from CelebA-HQ. We also compare the proposed method with two optimization-based approaches [1, 27]. Our approach substantially outperforms encoder-based baselines in terms of reconstruction quality and is considerably faster than optimization-based methods when inference.

4.2.2 Qualitative Evaluation

Encoder baselines. We show visual results of both inversion and editing in Fig. 4. Compared with previous approaches, our method is robust to images with occlusion and extreme viewpoints. For example, the first row in Fig. 4 gives a face image occluded by the hand, and the last row demonstrates a car image with an out-of-range viewpoint. Existing methods fail to reconstruct these challenging images faithfully. They generate distorted results and suffer artifacts for both inversion and editing. In contrast, with the proposed distortion consultation scheme, our method is more robust with high-fidelity results. Besides the robustness improvement, our approach also successfully preserves more details in backgrounds (4th row), shadow (2nd row), reflect (10th row), accessory (5th row), expressions (7th and 8th rows), and appearance (9th and 11th rows).

Optimization baselines. We also compare our method with optimization-based methods [1, 20, 27] in Fig. 5. Note that PTI [27] optimizes both latent codes and StyleGAN parameters, but we still report their results for better comparison. With $\sim 1000\times$ faster inference, our method achieves a comparable or even better reconstruction quality. Also, the editing results produced by the proposed scheme successfully preserve the image-specific details of source images without compromising the edit performance.

	Ours > pSp	Ours > e4e	Ours > Restyle
Preference Rate	81.2%	84.4%	79.7%

Table 2. The results of the user study. The reported value indicates the preference rate of Ours against a baseline.

4.2.3 User Study

To perceptually evaluate the editing performance, we conduct a user study in Table 2. We select the first 50 images from CelebA-HQ and perform editing on extensive attributes. We collect 1,500 votes from 30 participants. Each participant is given a triple of images (source, our editing, baseline editing) at once and asked to choose the higher-fidelity one with proper editing. The user study shows our method outperforms baselines by a large margin.

4.3. Ablation Study

4.3.1 Effect of Distortion Consultation

As discussed before, the distortion consultation inversion (DCI) scheme brings back ignored image details to complement the low-rate basic encoder, thereby achieving the high-fidelity reconstruction. To validate the effectiveness of DCI, we show our inversion results in Fig. 6. With the proposed distortion consultation branch, the model is more robust to occlusion and extreme poses and keeps more details in the reconstruction results.

4.3.2 Effect of Adaptive Distortion Alignment

To analyze the effect of ADA, we show the editing results with and without ADA in Fig. 7. Without the adaptive alignment, the distortion map fails to generalize to the edited image and degrades the generated image quality. In the proposed method, the aligned distortion map is embedded and integrated into the feature space via consultation encoding and consultation fusion. A naive alternative is to directly add the distortion map $\tilde{\Delta}$ to X_o^{edit} in the image space with warping estimated by face landmarks [12]. As shown in Fig. 7, performing warping and fusion in the image space also leads to obvious artifacts, where the warping is implemented by coordinates interpolation of facial landmarks.

4.4. Application on Video Editing

Compared with image inversion and editing, the key challenge for the video counterpart is the temporal consistency of details across frames. This puts a higher demand for reconstruction fidelity since the distortion of every single image would be magnified in a video in terms of consistency and quality [23]. We show inversion and editing results on a real video [28] in Fig. 8. Previous low-rate inversion approaches lack robustness to pose variation, and fail to preserve the identity of the original person and suffer notable distortion in editing results. When the pose and viewpoint change across video frames, their results show

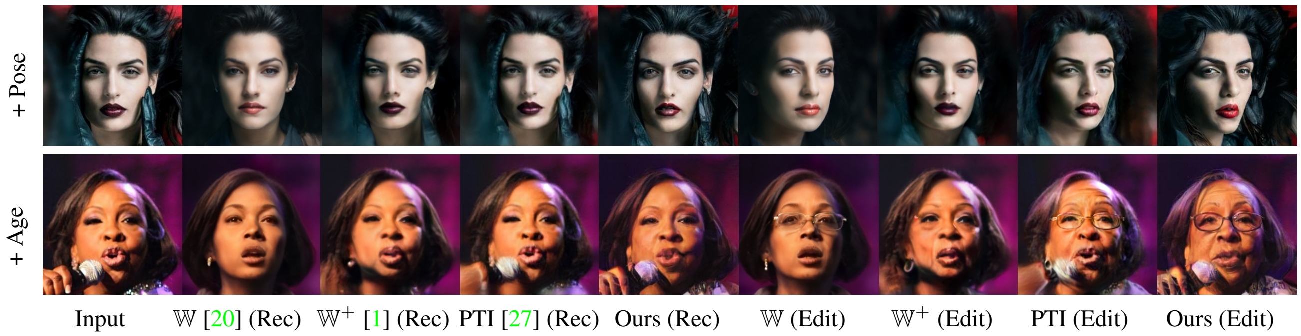


Figure 5. Visual comparison with optimization-based methods. Our method is much faster than these baselines.

Figure 6. Effects of distortion consultation inversion (DCI).

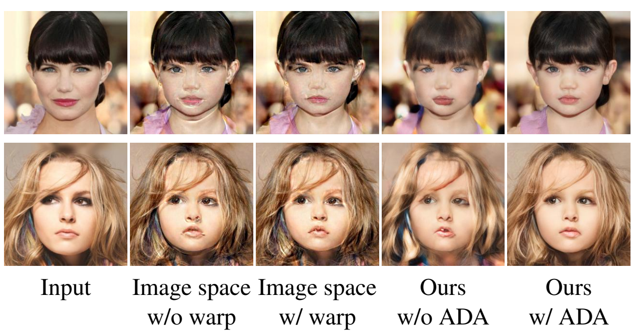


Figure 7. Effects of ADA. We integrate the distortion map in the image and feature space respectively and show the editing results.

inconsistent details and abrupt identity discrepancy. In contrast, the proposed method is more robust to cross-frame discrepancy (e.g., pose, viewpoint) and achieves higher fidelity for details preservation. More results in *mp4 format* are given in the **Supplement**.

5. Conclusion

In this work, we propose a novel GAN inversion framework that enables high-fidelity image attribute editing. With an information consultation branch, we consult the observed distortion map as a high-rate reference for lost information. This scheme enhances the basic encoder for high-quality reconstruction without compromising editability.

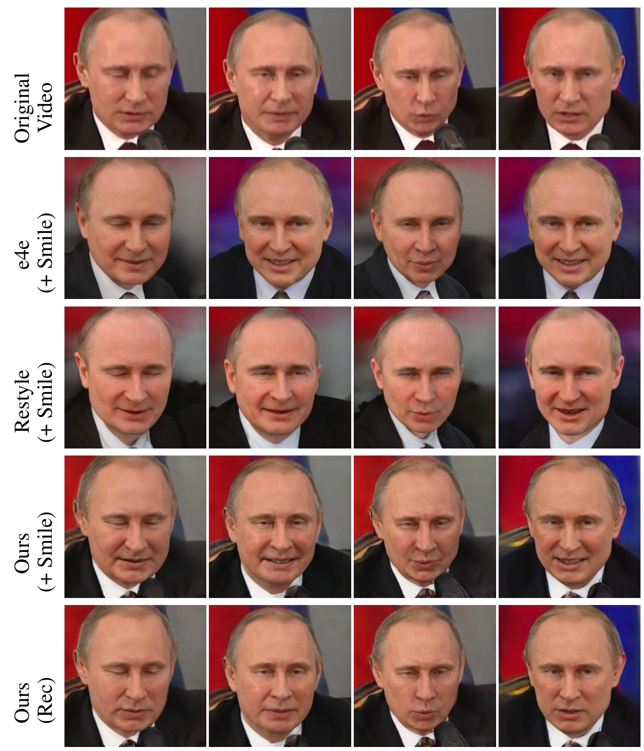


Figure 8. Inversion and editing results on a real video.

With the adaptive distortion alignment and distortion consultation technique, our method is more robust to challenging cases such as images with occlusion and extreme viewpoints. Benefiting from the additional information of the consultation branch, the proposed method shows clear improvements in terms of image-specific details preservation (e.g., background, appearance, and illumination) for both reconstruction and editing. The proposed framework is simple to apply, and we believe it can be easily generalized to other GAN models for future work.

Limitations. One limitation of the proposed method is the difficulty in handling large misalignment cases. As the augmented data used for ADA training in our experiments does not cover extreme misalignment, ADA is possibly insufficient when editing images with large viewpoint changes (see the **Supplement** for failure cases).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 7, 8

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[3] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. In *SIGGRAPH*, 2021. 3

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 7

[5] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning (ICML)*, 2019. 3

[6] Chaofeng Chen, Xiaoming Li, Yang Lingbo, Xianhui Lin, Lei Zhang, and Kwan-Yee K. Wong. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[7] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 3

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014. 1

[10] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[11] Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. Collaborative learning for faster stylegan embedding. *arXiv preprint arXiv:2007.01758*, 2020. 2

[12] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision (ECCV)*, 2020. 7

[13] Erik Harkonen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 3, 7

[14] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 4

[15] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images into class-conditional generative networks. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[16] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *The International Conference on Learning Representations (ICLR)*, 2020. 3

[17] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. Gan inversion for out-of-range images with geometric transformations. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *The International Conference on Learning Representations (ICLR)*, 2018. 3, 5

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 5, 7, 8

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International IEEE Workshop on 3D Representation and Recognition*, 2013. 5

[22] Yu-Ding Lu, Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Unsupervised discovery of disentangled manifolds in gans. *arXiv preprint arXiv:2011.11842*, 2020. 3

[23] Hao Ouyang, Tengfei Wang, and Qifeng Chen. Internal video inpainting by implicit long-range propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14579–14588, October 2021. 7

[24] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. *The International Conference on Learning Representations (ICLR)*, 2020. 3

[25] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *The International Conference on Learning Representations (ICLR)*, 2016. 2

[26] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 7

[27] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021. 2, 7, 8

[28] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 7

[29] Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959. 2, 3, 4

[30] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 5, 7

[31] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. 4

[32] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *Proceedings of Annual Allerton Conference on Communication, Control and Computing*, 1999. 4

[33] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015. 2, 4

[34] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 1, 2, 4, 5, 7

[35] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning (ICML)*, 2020. 3

[36] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2001–2010, October 2021. 5

[37] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[38] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. A simple baseline for stylegan inversion. *arXiv preprint arXiv:2104.07661*, 2021. 2

[39] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[40] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *arXiv preprint arXiv:2101.05278*, 2021. 1

[41] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[42] Oguz Kaan Yuksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. *arXiv preprint arXiv:2104.00820*, 2021. 3

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[44] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. Indomain gan inversion for real image editing. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[45] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[46] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 2

[47] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G Schwing. Enjoy your editing: Controllable gans for image editing via latent space navigation. *The International Conference on Learning Representations (ICLR)*, 2021. 3