ReGANIE: Rectifying GAN Inversion Errors for Accurate Real Image Editing

Bingchuan Li, Tianxiang Ma, Peng Zhang, Miao Hua, Wei Liu, Qian He, Zili Yi*

ByteDance Ltd, Beijing, China
{libingchuan, matianxiang.724, liuwei.jikun, zhangpeng.ucas, heqian, huamiao, yizili}@bytedance.com

Abstract

The StyleGAN family succeed in high-fidelity image generation and allow for flexible and plausible editing of generated images by manipulating the semantic-rich latent style space. However, projecting a real image into its latent space encounters an inherent trade-off between inversion quality and editability. Existing encoder-based or optimization-based StyleGAN inversion methods attempt to mitigate the trade-off but see limited performance. To fundamentally resolve this problem, we propose a novel two-phase framework by designating two separate networks to tackle editing and reconstruction respectively, instead of balancing the two. Specifically, in Phase I, a $\mathcal{W}$-space-oriented StyleGAN inversion network is trained and used to perform image inversion and editing, which assures the editability but sacrifices reconstruction quality. In Phase II, a carefully designed rectifying network is utilized to rectify the inversion errors and perform ideal reconstruction. Experimental results show that our approach yields near-perfect reconstructions without sacrificing the editability, thus allowing accurate manipulation of real images. Further, we evaluate the performance of our rectifying network, and see great generalizability towards unseen manipulation types and out-of-domain images.

Introduction

As one of the flagship unconditional GANs, StyleGAN (Karras, Laine, and Aila 2019) and its advanced versions (Karras et al. 2020b,a, 2021) are able to achieve high-fidelity image generation, but also facilitate the semantic editing in latent space. For example, some researchers (Shen et al. 2020; Wu, Lischinski, and Shechtman 2021; Patashnik et al. 2021; et al. 2021a; Tewari et al. 2020) seek to control the attributes of generated images by manipulating the $\mathcal{W}$ or $\mathcal{S}$ latent space of StyleGAN. These methods attempt to enhance the semantic interpretability of the latent space, thus allowing the editing of the generated images. However, the real photograph needs to be embedded to StyleGAN's latent space to facilitate the editing (Abdal, Qin, and Wonka 2020; et al. 2021b), which is prone to noticeable GAN inversion errors and long-tail information loss (Abdal, Qin, and Wonka 2020; et al. 2021b).

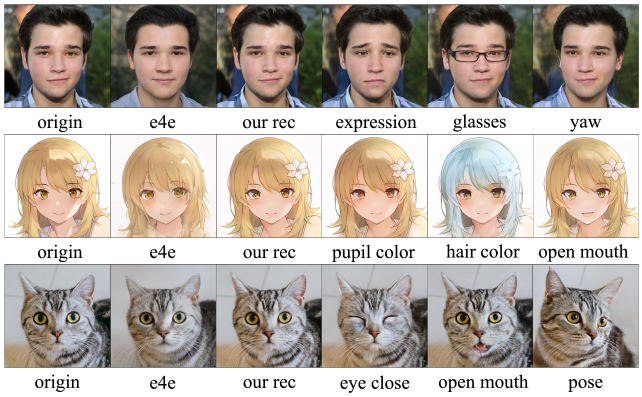

Figure 1: The image reconstruction and attribute editing results with our approach, compared with the baseline results (Tov et al. 2021) (second column).

Recent investigations have attempted to improve reconstruction quality while maintaining the editability. These methods mainly fall into three streams. The first stream is encoder-based (Tov et al. 2021; Alaluf, Patashnik, and Cohen-Or 2021b; Dinh et al. 2022; Alaluf et al. 2022; Wang et al. 2022) that optimizes the design of image encoders for improved reconstruction while keeping the StyleGAN generator fixed. The pioneering work is e4e (Tov et al. 2021), whose encoder is carefully designed to embed an image to the $\mathcal{W}$ space. Following this work, an iterative strategy (Alaluf, Patashnik, and Cohen-Or 2021b; Alaluf et al. 2022) to search the $\mathcal{W}$ code for better reconstruction is proposed, with the cost of increased inference time though. However, the embedding accuracy of these methods is far from being perfect. The second stream (Roich et al. 2022) is to fix the latent code and fine-tune the generator for better reconstruction. As the initial latent space is changed during optimization, the editability is weakened. In addition, the optimization process is time-consuming, typically two or three orders of magnitude more inference time than encoder-based methods. The third stream seeks to achieve near-perfect reconstructions by introducing spatial dimensions to the latent space (Kim et al. 2021), which makes the embedding more accurate but limits the editability to "in-place" and semantic-implicit manipulation. That means, global editing

*Corresponding author

(e.g., change the pose of an object) or attribute-conditioned editing (non-exemplar-based) of an image is intractable for these methods.

Generally, although previous schemes have made enormous efforts to improve the inversion accuracy of StyleGAN, the inherent trade-off between inversion quality and editability is not solved. Unlike previous solutions, our cleverly designed framework separates editability preservation and reconstruction improvement into two phases, thus avoiding the trade-off at all. As shown in Figure 2, we visually demonstrate the concepts and editability-reconstruction trade-offs of a representative StyleGAN inversion methods. Similar to existing StyleGAN inversion and latent-based image editing models (et al. 2021b; Tov et al. 2021; Alaluf, Patashnik, and Cohen-Or 2021b; Alaluf et al. 2022), the first phase is targeted on plausible editability and fair inversion quality. We design to generate 4 sets of paired images in the first phase for later learning to fully repair the loss of pre-inversion. The rectifying network in the second phase is designated to rectify the StyleGAN inversion errors caused in the first phase. The rectifying network takes the difference of the original image against its inversion and the initial editing result as inputs, and aims to reconstruct an ideal editing result. The difference image indicates the long-tail information loss (e.g., occlusion, accessories or other out-of-domain features) and contains sufficient supplementary information for ideal rectifying.

Generally speaking, our contributions include:

- For the first time, we propose REGANIE, a scheme that fundamentally addresses the editability-reconstruction trade-off for the task of latent-based real image editing.
- The unique two-phase training framework achieves near-perfect reconstruction without compromising editability, especially for the reproduction of long-tail information loss.
- We elaborate a novel architecture of the rectifying network equipped with the novel SG module, significantly leveraging the adaptability or generalizability of the rectifying network towards unseen edit types and out-of-domain images.
- Qualitative and quantitative comparisons with extensive GAN inversion methods validate the superiority of our method in terms of the reconstruction quality and editing accuracy without significantly compromising the inference time.

Related Work

Attribute Editing

Attribute control of images by manipulating the latent space of StyleGAN (Karras, Laine, and Aila 2019) is widely used. Early unsupervised methods (Shen and Zhou 2021; Härkönen et al. 2020) apply Principal Component Analysis (PCA) on latent space or model weights, and interpretable control can be performed by layer-wise perturbation along the principal directions. Due to limited control accuracy of unsupervised methods, a number of supervised methods are investigated. For example, InterfaceGAN (Shen et al. 2020) trains an SVM to discover the separation plane and editing direction for more explicit attribute control. Nonlinear methods such as (Wang, Yu, and Fritz 2021; Li et al. 2021; et al. 2021a) train neural networks to disentangle the GAN latent space by attributes. Further, the proposal of StyleSpace (Wu, Lischinski, and Shechtman 2021) and StyleCLIP (Radford et al. 2021; Patashnik et al. 2021) expands the scope of semantic control. Besides, regional semantic control methods (Ling et al. 2021; Shi et al. 2022; Chong, Lee, and Forsyth 2021; Hou et al. 2022) enables precise regional editing of images without unwanted global variation. Although existing StyleGAN-based image manipulation methods perform well for GAN-generated images, manipulating real images relies on accurate GAN inversion methods, which makes it challenging. Our method targets on resolving the GAN inversion errors for more accurate real image editing.

GAN Inversion

GAN inversion is to project the image into the GAN latent space and then reconstruct it by the generator, which is widely used in image analysis, image manipulation and compression. A straightforward approach is to iteratively adjust the latent codes to minimize the difference between the generated image and the target through gradient descent. More advanced research is the encoder-based inversion methods (Tov et al. 2021; Alaluf, Patashnik, and Cohen-Or 2021b; Dinh et al. 2022; Alaluf et al. 2022) that elaborate the design of image encoders for improved reconstruction accuracy. Limited reconstruction accuracy and the trade-off between reconstruction and editability are the major limitations of this stream. Some recent works (Wang et al. 2022; Yao et al. 2022; Kim et al. 2021) have attempted to incorporate spatial features directly into the intermediate layers of StyleGAN generators. Although it helps improve the reconstruction, it is prone to introducing artifacts when performing the edits that involves geometric deformation or pose variation. We are motivated by the joint spatial and global features of these methods when designing our rectifying network.

Methodology

We exploit a two-phase framework to fundamentally resolve the reconstruction-editability trade-off. In the first phase, a network consisting of a StyleGAN generator, an embedding encoder and a style editor is designated to embed the input real image and perform editing in the latent space: see more details in Section Phase I. The second phase involves a rectifying network that learns to rectify the inversion error and restore missing information produced in the first phase.

Phase I: Editing & Pre-inversion

As shown in Figure 3, our framework relies on a pre-trained StyleGAN2. A high-fidelity image can be generated from a randomly-sampled latent vector z_{init} from a normal distribution. The pre-trained mapping network M maps the z_{init} into $w_{init} \in \mathcal{W}$ space, and then generates an image through the generator G. That is

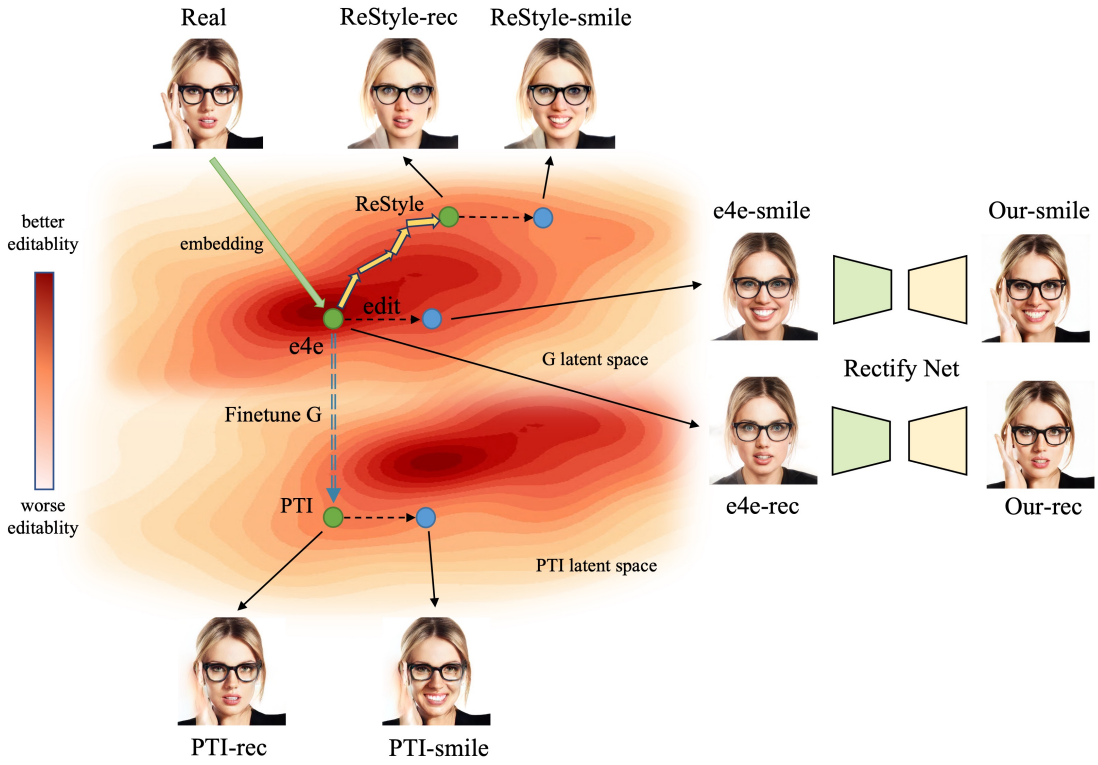

Figure 2: Conceptual visualization of the representative StyleGAN inversion approaches and the editability-reconstruction trade-off. The heatmap highlights the editability of StyleGAN latent space ($\mathcal{W}$ or $\mathcal{S}$) (Karras et al. 2020b; Wu, Lischinski, and Shechtman 2021), where darker color indicates better editability. The encoder-based approaches (e4e, ReStyle) employ either one-pass or multi-pass embedding strategy without tuning the generator. As shown, the multi-pass embedding strategy (ReStyle) produces better reconstruction but compromises editability. Optimization-based methods (PTI) seek to tune the StyleGAN generator for better reconstruction but inevitably sacrifice the editability (the latent space is shifted and distorted). Our method adopts the stable embedding and editing scheme from e4e (Tov et al. 2021) while introducing an additional rectifying network to achieve near-perfect reconstruction (hand occlusion) without compromising the editability.

$$X = G(w_{init}) \tag{1}$$

In general, the $\mathcal{W}$ space and its hierarchical expansion $\mathcal{W}+$ demonstrate good semantic interpretability (Wu, Lischinski, and Shechtman 2021). w_{init} can be semantically manipulated by a pre-trained style editing models (Shen and Zhou 2021; Härkönen et al. 2020; Shen et al. 2020; Patashnik et al. 2021; Wang, Yu, and Fritz 2021; Li et al. 2021; et al. 2021a; Ling et al. 2021; Shi et al. 2022; Chong, Lee, and Forsyth 2021; Hou et al. 2022). Without losing generality, we choose the state-of-the-art non-linear multi-attribute style editor DyStyle (Li et al. 2021) to perform the editing. The corresponding editing result can be obtained by

$$Y = G(Editor(w_{init})) \tag{2}$$

The next task is to produce paired data for the second-stage network training, so pre-inversion of X, Y is required. We train the encoder E (Tov et al. 2021) to embed images into the latent $\mathcal{W}$ space of the original StyleGAN generator G, which guarantees consistency of editability in training and testing phases. Then, the generator G is employed to invert the embedding of $\mathcal{W}$ space back to the image. So we can invert Image X or Y to its inversion as

$$\hat{X},\ \hat{Y} = G(E(X)),\ G(E(Y)) \tag{3}$$

Phase II: Rectifying Network

The rectifying network in Phase II targets on rectifying the GAN inversion errors and restoring the missing information produced in Phase I. The rectifying network is conditioned on the difference image between the original image X and its inversion $\hat{X}$, and the inversion of the editing result $\hat{Y}$, with the expect to reconstruct the ideal editing result Y.

Dual-Pathway Encoder As shown in figure 3, the rectifying network employs a dual-pathway encoder that processes the difference image $\Delta X = X - \hat{X}$ and the primary image $\hat{Y}$ in separate branches. Since the $Y - \hat{Y}$ pair and the $X - \hat{X}$ pair are thought to suffer similar loss of information caused by the same inversion model, we assume that the difference between X and $\hat{X}$ provides sufficient information for perfect reconstruction of Y from $\hat{Y}$. We build separate

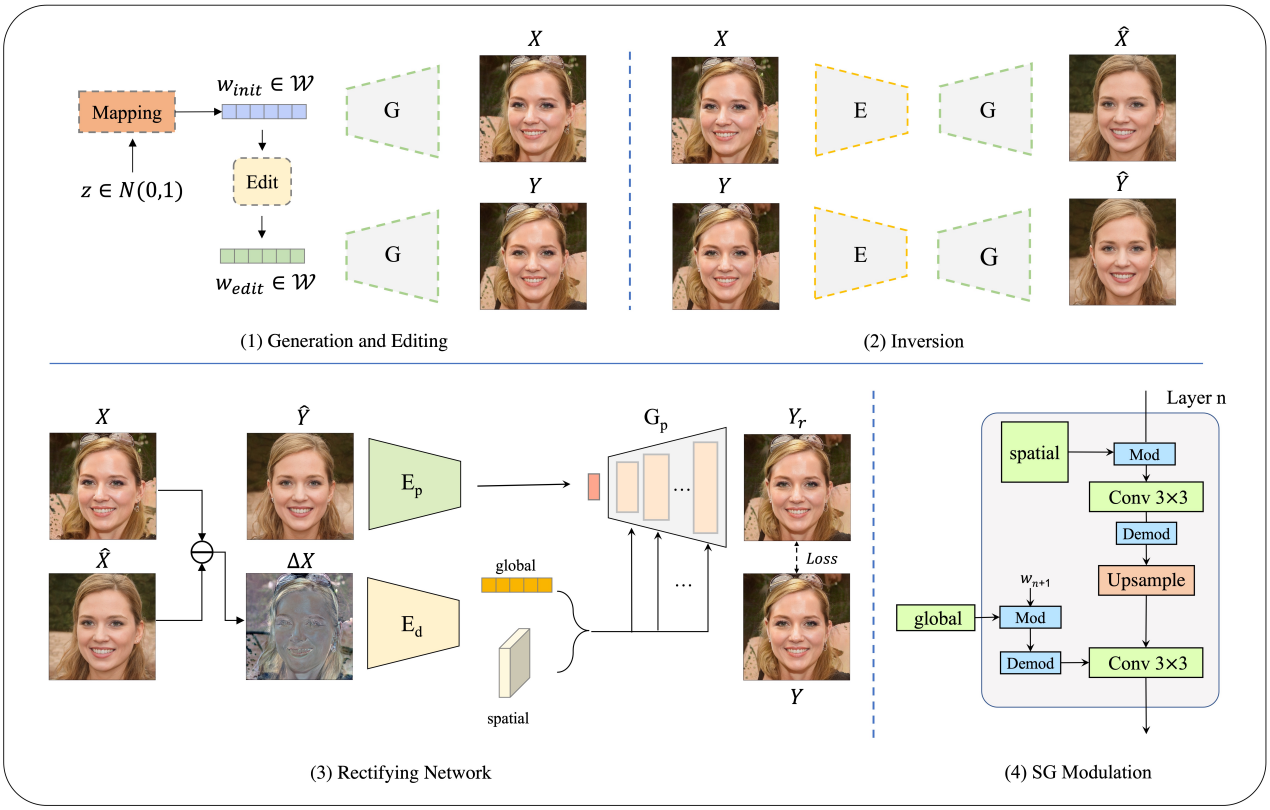

Figure 3: The overview of our two-phase framework as for training. The upper part demonstrate the inversion and editing (e.g. pose) process of the first phase. The bottom part depicts the architecture of the rectifying network and the SG module used in Phase II. Note that the inference follows a slightly different pipeline as detailed in Alg. 1.

encoders (E_p and E_d) to process the primary image and the difference image respectively: see Figure 3 (bottom).

Considering an image of resolution 512×512, Encoder E_p and Encoder E_d consist of 4 downsampling layers respectively. Specifically, E_p extracts a latent feature map f_x of size 32×32 from the primary image. E_d extracts a spatial style code sp of size 32×32 and a global style vector gl of dimension 2048, which is written as

$$sp,\ gl = E_d(\Delta X) \tag{4}$$

where sp is the style codes with spatial dimensions, and gl is the global style vector.

Generator A style-based Generator G_p is then used to fuse the extracted style codes (sp and gl) and latent features (f_x), and generates an image with the inversion errors rectified. The generator starts from the latent feature f_x and generates a 512×512 image by applying 4 upsampling layers. The style codes (sp and gl) are joined to the generator through style modulation as used in (Park et al. 2020). Note that X and Y are both generated images, they are actually exchangeable to each other, just as

$$X_r = G_p(E_p(\hat{X}), E_d(\Delta Y)) \tag{5}$$

$$Y_r = G_p(E_p(\hat{Y}), E_d(\Delta X)) \tag{6}$$

Spatial and Global Modulation The style codes sp and gl are designated to convey spatial and global information respectively. The original StyleGAN architecture projects an image from a 512-d vector that causes loss of spatial information and leads to an optimization upper bound for inversion accuracy. Therefore, we modify the generator structure to allow for joint spatial and global modulation.

The style modulation module slightly different from that used in StyleGAN2 (Karras et al. 2020b) to allow for alternative modulations of the spatial codes and the global vector: see Figure 3 (bottomright). The equations are written as

$$f^{l+1} = (f^l \cdot sp) * w^l + b^l \tag{7}$$

$$f^{l+2} = (f^{l+1} \cdot gl) * w^{l+1} + b^{l+1} \tag{8}$$

where f^l represents the feature map of the lth layer. w^l and b^l represent the weights and biases of the lth layer, respectively. "$\cdot$" denotes element-wise multiplication and "$*$" refers to convolution.

Spatial intermediate features are used in (Wang et al. 2022; Kim et al. 2021; Yao et al. 2022) for improved reconstruction. However, they are prone to artifacts when deformation occurs in the editing, as the spatial features are not well aligned with the edited image. To avoid the issue, we alternatively place the spatial modulation and global modu-

Algorithm 1: Training & inference

Training:
Pre-trained Models: M, G, E, $Editor$ in Phase I
Models to be optimized: G_p, E_p, E_d, D in Phase II
$iter = 0$;
while *iter* $\leq N$ **do**
  sample $z \in N(0,1)$; $w_{init} \leftarrow M(z)$;
  $X = G(w_{init})$; $Y = G(Editor(w_{init}))$;
  $\hat{X},\ \hat{Y} = G(E(X)),\ G(E(Y))$;
  $t \leftarrow random\_choice(\{1,\ 2,\ 3\})$;
  if $t = 1$ **then**
    $\Delta I \leftarrow Y - \hat{Y}$; $\hat{I} \leftarrow \hat{X}$; $I \leftarrow X$;
  else if $t = 2$ **then**
    $\Delta I \leftarrow X - \hat{X}$; $\hat{I} \leftarrow \hat{Y}$; $I \leftarrow Y$;
  else
    $\Delta I \leftarrow X - \hat{X}$; $\hat{I} \leftarrow \hat{X}$; $I \leftarrow X$;
  end
  $I_r \leftarrow G_p(E_p(\hat{I}), E_d(\Delta I))$;
  $\mathcal{L} \leftarrow \mathcal{L}_{rec}(I_r, I) + \mathcal{L}_{GAN}(D(I_r), D(I))$;
  optimize G_p, E_p, E_d based on loss $\mathcal{L}$;
  $iter \leftarrow iter + 1$;
end

Inference:
  $\hat{X},\ \hat{Y} \leftarrow G(E(X_0)),\ G(Editor(E(X_0)))$;
  $\Delta X \leftarrow X_0 - \hat{X}$;
  $X_r \leftarrow G_p(E_p(\hat{X}), E_d(\Delta X))$;
  $Y_r \leftarrow G_p(E_p(\hat{Y}), E_d(\Delta X))$;
end

lation in each layer of the generator. Please refer to supplement for detailed architecture. We observe that the encoder and generator will adaptively learn to re-align the spatial information based on global information.

Training & Inference The rectifying network is trained with an objective consisting of three loss terms, which is defined as:

$$\mathcal{L} = \lambda_{l1}\mathcal{L}_1(I, I_r) + \lambda_{lpips}\mathcal{L}_{lpips}(I, I_r) + \lambda_{GAN}\mathcal{L}_{GAN}(I, I_r) \tag{9}$$

where $\mathcal{L}_1$ loss is used to suppress the pixel-wise disparities between the reconstructed image I_r and the original image I. $\mathcal{L}_{lpips}$ (Zhang et al. 2018) loss measures the features similarities between the two based on a pre-trained VGG16 network (Simonyan and Zisserman 2014), which enforces reconstructions at the feature level. The image realism is encouraged using an adversarial loss $\mathcal{L}_{GAN}$ with R_1 regularization (Karras et al. 2020b), where the discriminator D is trained adversarially against the rectifying network based on Loss $\mathcal{L}_{\mathcal{D}}$. We set $\lambda_{GAN} = 1.0$, $\lambda_{l1} = 1.0$ and $\lambda_{lpips} = 1.0$ in our experiments for the best performance.

As in Algorithm 1, in the first phase, we use off-the-shelf pre-trained models provided in (Karras et al. 2020b; Li et al. 2021; Tov et al. 2021). We train the rectifying network with generated samples from Phase I. Specifically, the training triplets (ΔI, $\hat{I}$ and I) can be obtained in three ways to encourage reconstructions of both edit-free or edit-involved scenarios. First, $\Delta I = Y - \hat{Y}$, $\hat{I} = \hat{X}$ and $I = X$. Second, $\Delta I = X - \hat{X}$, $\hat{I} = \hat{Y}$ and $I = Y$. Third, $\Delta I = X - \hat{X}$, $\hat{I} = \hat{X}$ and $I = X$. The former two encourage reconstructing from its inversion and the inversion errors of its edited partner. The third way encourages reconstructing from its own inversion and inversion errors. During training, the attribute set intended for manipulation are those provided by DyStyle (Li et al. 2021) network, including pose, age, glasses, expression, etc.

During inference, the rectifying network is expected to reconstruct an ideal edited image Y_r from its inversion $\hat{I} = \hat{Y}$ and the inversion errors of the input image $\Delta I = X_o - \hat{X}$. It is also encouraged to reconstruct the input image itself X_r from its own inversion $\hat{I} = \hat{X}$ and inversion errors $\Delta I = X_o - \hat{X}$.

Experiments

Experimental Setup

The pre-trained StyleGAN models that our method relies on are typically trained on facial datasets. To demonstrate the generalizability of our method on unseen realistic faces, we train our model on the FFHQ (Karras, Laine, and Aila 2019) dataset, and test on the CelebA-HQ (Karras et al. 2018) dataset. We choose the Dystyle (Li et al. 2021) and StyleCLIP as the latent editor, as the pre-trained models for multi-attribute-conditioned or text-conditioned editing are available. We also experiment our method on the animal portrait dataset (Choi et al. 2020) and the anime dataset (Branwen 2019).

Without losing generality, we adopt the GAN inversion model and training procedure of the embedding encoder from e4e (Tov et al. 2021). We re-trained the embedding encoder for the anime and animal portrait datasets. In addition, we also compare our method with PTI (Roich et al. 2022), a representative optimization-based method. The official pre-trained models provided in PTI is used as the baseline for evaluation.

Implementation Details

The proposed ReGANIE network is trained on a single Tesla V100 GPU with batch size of 8 on 512×512 resolution images. It is optimized using the Adam optimizer (Kingma and Ba 2014) with b1 = 0.0 and b2 = 0.99, and the learning rate is fixed at 0.002. In all experiments, the training is performed for 100,000 iterations. The inference costs 190ms on a single 512×512 image.

Evaluations of Image Reconstruction Quality

Qualitative Results To verify the reconstruction accuracy of the proposed method, we compare our approach with state-of-the-art GAN inversion methods including e4e (Tov et al. 2021), ReStyle (Alaluf, Patashnik, and Cohen-Or 2021b), HFGI (Wang et al. 2022), HyperStyle (Alaluf et al. 2022) and PTI (Roich et al. 2022): see Figure 4.

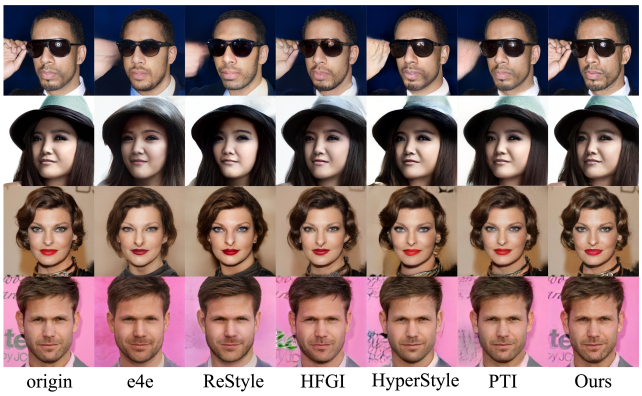

Figure 4: Comparison of reconstruction quality. Four different long-tail loss subjects are demonstrated, including hand and glasses occlusion (1st row), hats (2nd row), accessories and makeup (3rd row), and background (4th row).

Methods	L2(↓)	LPIPS(↓)	ID(↑)	MS-SSIM(↑)	Times(↓)
e4e	0.048	0.2	0.61	0.73	**0.05**
ReStyle	0.043	0.18	0.63	0.79	0.46
HFGI	0.039	0.16	0.78	0.83	0.24
HyperStyle	0.021	0.11	0.83	0.85	1.13
PTI	0.019	0.08	0.84	0.9	76
Ours	**0.016**	**0.07**	**0.85**	**0.9**	0.19

Table 1: Quantitative comparison of reconstruction quality on CelebA-HQ (Karras et al. 2018).

We can observe that the encoder-based methods e4e and ReStyle cannot restore difficult cases well (long-tail information loss). HFGI, HyperStyle and PTI are able to recover more details in the original image due to use of spatial information encoding or generator optimization, but the inversion accuracy is far from being perfect. Our approach achieves the best reconstruction accuracy and the reconstruction errors are mostly unnoticeable.

Quantitative Results We perform a quantitative comparison of related GAN inversion methods and ours in terms of the reconstruction accuracy and inference time: see Table 1. The similarity between the reconstructed image and the original was measured with L2 distance, LPIPS (Zhang et al. 2018) and MS-SSIM (Wang, Simoncelli, and Bovik 2003) scores. Additionally, the identity similarity is measured using a pre-trained face recognition model provided by (Huang et al. 2020). Note that, unlike other methods, we do not explicitly apply the identity preserving loss during training. Finally, the inference time is also tested. e4e (Tov et al. 2021), HFGI (Wang et al. 2022) and our method are single-pass methods, while ReStyle (Alaluf, Patashnik, and Cohen-Or 2021b) and HyperStyle (Alaluf et al. 2022) employ multi-pass forwarding (5 based on the official recommendation). Since PTI (Roich et al. 2022) is optimization-based, it is iterated for 450 times to search the initial code, followed by 350 iterations to update the generator weights. PTI costs 2-3 orders of magnitude more time than the encoder-based methods. As shown in Table 1, our method outperforms other methods in terms of the reconstruction accuracy by large margin, while the speed is among the best two.

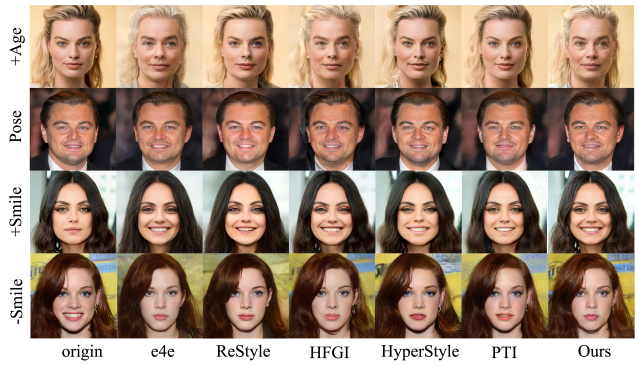

Figure 5: Comparisons of the real face editing results while manipulating the age, pose and expression.

Evaluations of Real Image Editing

Qualitative Results To visually compare different image editing methods, we demonstrate their results as for manipulating three types of facial attributes: age, pose and expression. The selected attributes are representatives of three typical editing scenarios. Specifically, age involves whole-image texture editing, pose involves global geometric deformation, and expression involves local deformation. As our rectifying network is trained with the editing results of DyStyle (Li et al. 2021), for fair comparisons, we use InterfaceGAN (Shen et al. 2020) as the style editor in the test phase for all methods.

It can be observed from Figure 5 that ReStyle (Alaluf, Patashnik, and Cohen-Or 2021b), HyperStyle (Alaluf et al. 2022), and PTI (Roich et al. 2022) see limited attribute control accuracy: see the failure to produce wrinkles while designated for aging. HFGI (Wang et al. 2022) produces silhouette artifacts, which can be caused by its naive injection of spatial information into the generator. The e4e (Tov et al. 2021) model obtains stable editing results, but suffer noticeable reconstruction errors and unwanted information loss. In general, our method achieves the best reconstruction quality while not compromising the attribute control accuracy. User study and more visualizations see the supplement.

Qualitative Results We quantitatively evaluate the editing quality of different methods, by measuring the attribute control accuracy and identity preservation. As InterfaceGAN edits image attributes by linearly manipulating the StyleGAN latent space $\mathcal{W}$. Specifically, $w_{edit} = w_{init} + \alpha * d$, where d is the normal vector of the separation plane (or edit direction) for a specific attribute, and α is the edit magnitude. The attribute of an image can be continuously manipulated by controlling the value of α.

We use two numeric attributes, age and yaw angle, for quantitative evaluation. We intend to see how the perceived attribute value responds to the value of α. It is believed

Attr.	α	e4e	ReStyle	HFGI	HyperStyle	PTI	Ours
Age	α=-2	-11.53	-8.37	-11.64	-9.73	-9.05	**-12.18**
	α=2	9.02	6.42	9.25	8.46	8.11	**10.59**
	Id	0.44	0.45	0.48	0.59	0.57	**0.62**
Yaw	α=-2	-9.12	-6.65	-7.56	-8.9	-8.02	**-9.13**
	α=2	9.52	6.75	7.83	9.06	8.37	**9.54**
	Id	0.52	0.54	0.61	0.72	0.69	**0.75**

Table 2: Quantitative comparisons of attribute editing accuracy on CelebA-HQ (Karras et al. 2018). The attribute control respondence is indicated by the perceived attribute variation given the fixed editing magnitude α (the greater absolute value the better).

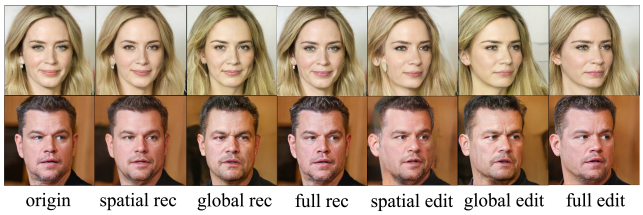

Figure 6: Ablation studies on the spatial and global modulation (SG) module for image reconstruction and pose editing.

that the greater the perceived attribute value varies given the same amount of α shift, the better the editability is. We employ the official pre-trained HopeNet (Ruiz, Chong, and Rehg 2018) model for pose estimation, and the age regressor is trained by ourselves based on the official implementation of Dex VGG (Alaluf, Patashnik, and Cohen-Or 2021a). In addition, we also calculated the average identity similarity (Huang et al. 2020) score as the indicator of identity preservation. As shown in Table 2, our method and e4e (Tov et al. 2021) achieve the best attribute control competency, while our method achieve the best identity preservation, implying that our method faithfully manipulates the target attribute without producing unwanted changes along other attributes.

Ablation Study

We compare the effects of spatial and global modulations mentioned on image reconstruction and editing. As shown in Figure 6, using the global modulation only (3rd column) sees missing details such as earrings and hairstyles. In contrast, using spatial modulation only improves the reconstruction quality, but produces noticeable artifacts for attribute editing where deformation happens (5th column). The rational design of the SG module that combines the spatial and global modulation performs the best in reconstruction (4th column) and editing (7th column). Furthermore, we conduct additional ablation studies on the input of the rectifying network. Please refer to supplementary materials for detailed experimental analysis.

Generalizability

Unseen Manipulations In this work, we use DyStyle (Li et al. 2021) to generate the editing results required for the training of the rectifying network. However, what is inspiring is that the performance of the rectifying network for manipulations of unseen attributes such as hair color, bangs, or face shape is also plausible: see Figure 7. This implies that the generator has learned to align the two inputs regardless of the edit types.

Figure 7: The performance of our rectifying network as used to reconstruct the results of unseen attribute manipulations by StyleCLIP (Patashnik et al. 2021).

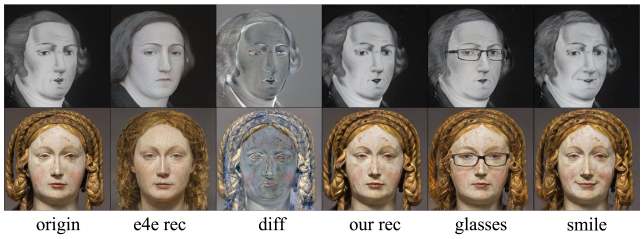

Figure 8: The GAN inversion error rectifying results on the Metface (Karras et al. 2020a) dataset.

Out-of-Domain Images All networks in Phase I and Phase II are trained with images from the same domain. However, we discover that our rectifying network works well on certain unseen images out of the target domain. As shown in Figure 8, the editing results on the artistic portrait dataset can be rectified with the rectifying network trained on FFHQ dataset without any fine-tuning. This implies that our rectifying network has learned to fuse and repair the inputs without overfitting into a specific domain.

Conclusions

We propose ReGANIE, a novel two-phase framework for accurate latent-based realistic image editing. Compared to previous encoder-based or optimization-based inversion-and-editing methods that perform reconstruction and editing with one generator, we successfully resolve the reconstruction-editability trade-off by designating separate networks to deal with the editing and reconstruction respectively. As a result, we achieve the most accurate real image editing results, without significantly increasing the inference time. Furthermore, ReGANIE exhibits great generalization towards unseen manipulation types (e.g., unseen attributes of editing and certain out-of-domain images). Finally, our framework is scalable and can be customized by choosing different GAN inversion methods or style editors for specific scenarios.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2020. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8296–8305.

Alaluf, Y.; Patashnik, O.; and Cohen-Or, D. 2021a. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 40(4): 1–12.

Alaluf, Y.; Patashnik, O.; and Cohen-Or, D. 2021b. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 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*, 18511–18521.

Branwen, G. 2019. Anonymous, and Danbooru Community. Danbooru2019 portraits: A large-scale anime head illustration dataset. https://www.gwern.net/Danbooru2019. Accessed: 2021-11-03.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8188–8197.

Chong, M. J.; Lee, H.-Y.; and Forsyth, D. 2021. StyleGAN of All Trades: Image Manipulation with Only Pretrained StyleGAN. *arXiv preprint arXiv:2111.01619*.

Dinh, T. M.; Tran, A. T.; Nguyen, R.; and Hua, B.-S. 2022. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11389–11398.

et al., A. 2021a. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *TOG*, 40(3): 1–21.

et al., R. 2021b. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2287–2296.

Härkönen, E.; Hertzmann, A.; Lehtinen, J.; and Paris, S. 2020. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33: 9841–9850.

Hou, X.; Shen, L.; Patashnik, O.; Cohen-Or, D.; and Huang, H. 2022. FEAT: Face Editing with Attention. *arXiv preprint arXiv:2202.02713*.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5901–5910.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020a. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33: 12104–12114.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34.

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*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020b. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kim, H.; Choi, Y.; Kim, J.; Yoo, S.; and Uh, Y. 2021. Exploiting spatial dimensions of latent in gan for real-time image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 852–861.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, B.; Cai, S.; Liu, W.; Zhang, P.; Hua, M.; He, Q.; and Yi, Z. 2021. DyStyle: Dynamic Neural Network for Multi-Attribute-Conditioned Style Editing. *arXiv preprint arXiv:2109.10737*.

Ling, H.; Kreis, K.; Li, D.; Kim, S. W.; Torralba, A.; and Fidler, S. 2021. EditGAN: High-Precision Semantic Image Editing. *Advances in Neural Information Processing Systems*, 34.

Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A.; and Zhang, R. 2020. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33: 7198–7211.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

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.

Ruiz, N.; Chong, E.; and Rehg, J. M. 2018. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2074–2083.

Shen, Y.; Yang, C.; Tang, X.; and Zhou, B. 2020. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*.

Shen, Y.; and Zhou, B. 2021. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1532–1540.

Shi, Y.; Yang, X.; Wan, Y.; and Shen, X. 2022. SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11254–11264.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tewari, A.; Elgharib, M.; Bharaj, G.; Bernard, F.; Seidel, H.-P.; Pérez, P.; Zollhofer, M.; and Theobalt, C. 2020. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6142–6151.

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.

Wang, H.-P.; Yu, N.; and Fritz, M. 2021. Hijack-gan: Unintended-use of pretrained, black-box gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7872–7881.

Wang, T.; Zhang, Y.; Fan, Y.; Wang, J.; and Chen, Q. 2022. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11379–11388.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, 1398–1402. Ieee.

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*, 12863–12872.

Yao, X.; Newson, A.; Gousseau, Y.; and Hellier, P. 2022. Feature-Style Encoder for Style-Based GAN Inversion. *arXiv e-prints*, arXiv–2202.

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 conference on computer vision and pattern recognition*, 586–595.