

From Continuity to Editability: Inverting GANs with Consecutive Images

Yangyang Xu¹, Yong Du², Wenpeng Xiao¹, Xuemiao Xu^{1,3,4,5*} and Shengfeng He^{1,5*}

¹South China University of Technology ²Ocean University of China

³Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information

⁴State Key Laboratory of Subtropical Building Science

⁵Ministry of Education Key Laboratory of Big Data and Intelligent Robot

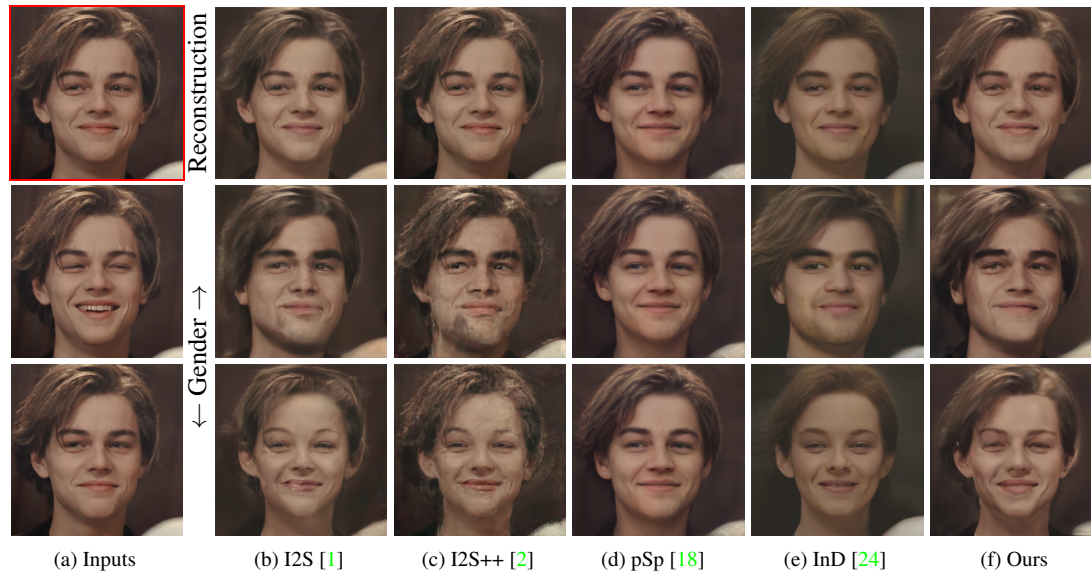


Figure 1: We propose an alternative GAN inversion method that jointly considers multiple consecutive images. We leverage the inborn continuity between inputs, to simultaneously maximize the reconstruction fidelity (top row) and editability (bottom row) of the inverted latent code. Our method jointly inverts three images from (a), we only show the result of the first one with the red box.

Abstract

Existing GAN inversion methods are stuck in a paradox that the inverted codes can either achieve high-fidelity reconstruction, or retain the editing capability. Having only one of them clearly cannot realize real image editing. In this paper, we resolve this paradox by introducing consecutive images (e.g., video frames or the same person with different poses) into the inversion process. The rationale behind our solution is that the continuity of consecutive images leads to inherent editable directions. This inborn property is used for two unique purposes: 1) regularizing the joint inversion process, such that each of the inverted codes is semantically accessible from one of the other and fastened in an editable domain; 2) enforcing inter-image coherence, such that the fidelity of each inverted code can be maximized with the complement of other images. Extensive experiments demonstrate that our alternative significantly outperforms state-of-the-art methods in terms of reconstruction fidelity and editability on both the real image dataset and synthesis dataset. Furthermore,

our method provides the first support of video-based GAN inversion and an interesting application of unsupervised semantic transfer from consecutive images. Source code can be found at: https://github.com/cnnlstm/InvertingGANs_with_ConsecutiveImgs.

1. Introduction

Generative Adversarial Networks (GANs) [5, 9, 10] has demonstrated versatile image editing capability, especially by discovering the spontaneously learned interpretable directions that can manipulate corresponding image attributes [17, 8, 19, 4, 22]. Concretely, given a random latent code w , image editing can be achieved by pushing the latent vector along a specific semantic direction (e.g., age, gender):

$$I' = G(w + \alpha \times \vec{n}), \quad (1)$$

where I' is the edited image, G is the generator, α is a scaling factor, and \vec{n} is the interpretable direction.

As a consequence, recent attempts [1, 2, 18, 24] aim to migrate this power to real image editing by inverting an image to the latent code w . There are two keys for this task,

*Corresponding authors ({xuemx,hesfe}@scut.edu.cn).

i) whether the inverted code can faithfully reconstruct the original input, and, ii) whether the pre-acquired semantic directions can be successfully applied. However, existing methods seem to stuck in a paradox, as achieving one end will inevitably sacrifice the other. As shown in Fig. 1, I2S, I2S++, and pSp [1, 2, 18] concentrate only on obtaining faithful reconstruction, but the inverted codes show limited editability. In contrast, latent codes obtained from In-domain inversion [24] (Fig. 1e) are regularized in the semantically meaningful domain at the expense of fidelity. We argue that balancing these two factors solely based on a single image is extremely challenging, as there is no indicator to shed light on the editable domain in the latent space, preventing the optimization from obtaining a perfect balance between two factors.

In this paper, we resolve the above problem by introducing consecutive images, which can be either a video segment or the same person with different poses, to form a joint optimization process. The rationale behind our alternative solution is that the *continuity* brought by consecutive images can be used as an indicator to constrain the *editability*. In particular, to ensure each of the inverted latent codes is editable, we jointly optimize multiple latent codes by enforcing each of them is semantically accessible from one of the other code with a simple linear combination. In addition, we further explore the sequential continuity for *fidelity*, by injecting multi-source supervision that common regions of the reconstructed images should be consistent in all the consecutive images. We establish dense correspondences between input images, and then apply the obtained correspondences to warp each of the reconstructions to the neighbors for a consistent and coherence measurement.

To evaluate the proposed method, we construct a real video-based dataset RAVDESS-12, and another consecutive images dataset synthesized by manipulating attributes from the generated images of StyleGAN [10]. Extensive experiments demonstrate the superior performance of our solution over existing methods in terms of editability and reconstruction fidelity. Furthermore, our method is capable to perform various applications, *e.g.*, video-based GAN inversion, unsupervised semantic transfer, and image morphing.

In summary, our contributions are three-fold:

- We propose an alternative GAN inversion method for consecutive images, and delve deep into the continuity of consecutive images for GAN inversion.
- We tailor two novel constraints, one is the mutually accessible constraint that formulates consecutive images inversion as a linear combination process in the latent space to ensure editability, and the inversion consistency constraint that works in the RGB space to guarantee reconstruction fidelity, by measuring reconstruction consistency across images.
- We demonstrate optimal performances in terms of editability and reconstruction fidelity, and we support various new applications like video-based GAN inversion and unsupervised semantic transfer.

2. Related Work

Image Editing via Latent Space Exploration. Generative models show great potential in synthesizing versatile images by taking random latent codes as inputs. Recent works show that the latent space of a pre-trained GAN encodes rich semantic directions. Varying the latent codes with a specific direction can edit the image with the target attribute. In particular, Radford *et al.* [17] observe that there are directions in the latent space corresponding to adding smiles or glasses on the faces. Ganalyze *et al.* [4] explore the memorability direction in the latent space by a fixed assessor. Jahanian *et al.* [8] study the steerability of GANs to fit some image transformations. Shen *et al.* [19] explore the semantic boundary in the latent space of the binary attributes. Voynov *et al.* [22] discover the semantic directions hidden in the latent space by an unsupervised model-agnostic procedure. Varying the latent codes with such directions can manipulate the corresponding attributes of the output images. It is natural to transferring these techniques on real image editing. Before that, it is required to invert a real image back to the latent code.

GAN Inversion. To realize real image editing, GAN inversion methods are proposed to inference a latent code of an input image based on the pre-trained GAN [16, 25, 1, 3, 15]. These methods can be categorized into two classes, optimization-based and encoder-based. The former individually optimizes the latent code for a specific image, with a concentration on pixel-wise reconstruction [1, 2, 3, 13]. However, ensuring reconstruction fidelity cannot guarantee the output latent code is editable by the learned directions. On the other hand, encoder-based methods train a general encoder that maps real images to latent codes [24, 18]. Especially, In-Domain GAN inversion [24] combines the learned encoder with an optimization process to align the encoder with the semantic knowledge of the generator. However, existing methods do not resolve the problem of editable domain in the latent space, thus they cannot achieve a perfect balance between editability and fidelity. We aim to solve this problem from a novel view of considering multiple consecutive images.

3. Method

3.1. Overview

The editability of the latent code and the fidelity of the reconstruction are the two vital factors that affect the performance of GAN inversion. To satisfy both sides, we exploit the continuity brought by consecutive images which depict a same subject with different variations. The pipeline of our

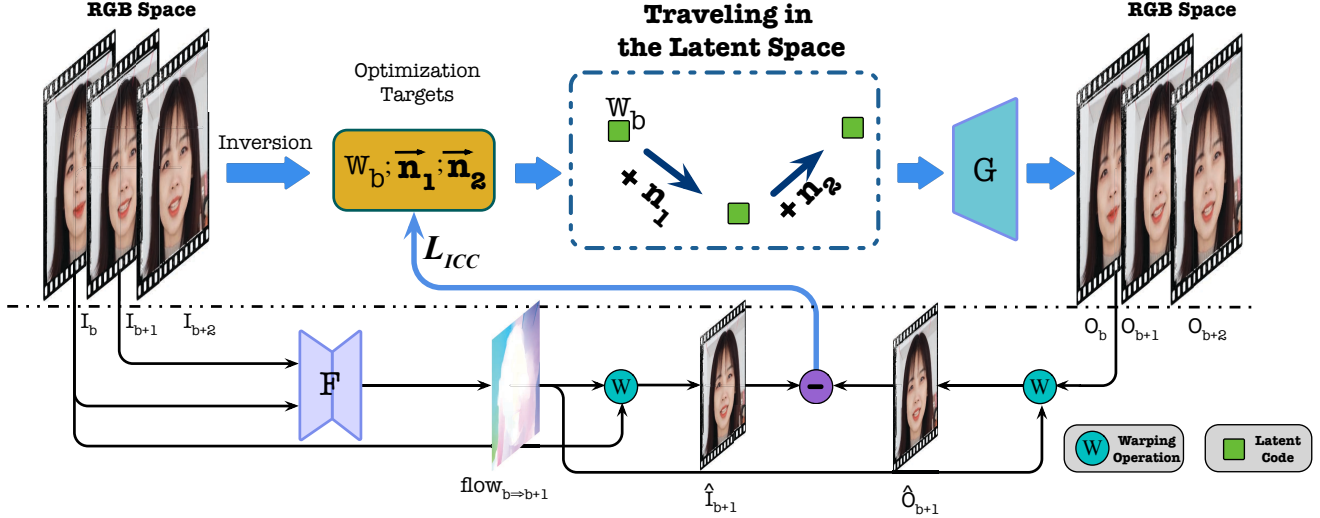


Figure 2: The pipeline of the proposed consecutive images based GAN inversion. G is a pretrained generator of StyleGAN and F is a pretrained FlowNet for calculating the bidirectional optical flow. The upper part shows the Mutually Accessible Constraint. Given consecutive images as input, we enforce each of them is semantically accessible from one of the other codes with a simple linear combination. Both w_b and \vec{n}_k are the optimization targets. The bottom shows the Inversion Consistency Constraint in the RGB space. Note here we only show the forward flow among I_b and I_{b+1} when calculating the \mathcal{L}_{ICC} for simplicity, meanwhile, \mathcal{L}_C and \mathcal{L}_P are also omitted.

approach is illustrated in Fig. 2. Given a sequence of consecutive images as input, the proposed method aims to seek their optimal latent codes in the latent space, and then feed them into a pretrained and fixed generator for reconstruction. In particular, i) we define a linear combination mechanism among consecutive images, which would facilitate the editability of the latent codes via a joint optimization with semantic directions, and ii) we establish a consistency constraint in the RGB space between the warped results of the reconstructed images and their corresponding originals, promoting the fidelity of the reconstruction. Note that we choose the generator of StyleGAN [9] as the pretrained one in our model.

3.2. Consecutive Images Based GAN Inversion

Mutually Accessible Constraint. For each image in an input set, it may gradually change into the others, just as shown in Fig. 2, the mouth opens progressively. Or it can vary to the others in a drastic way, such as the same person at very different poses. In either case, given the base image I_b , it can be intuitively assumed that the latent codes of the other images are linear combinations of that of the base image along with a specific semantic direction \vec{n} (e.g., expression, pose). Then, the latent code w_{b+K} of the image I_{b+K} can be formulated as follows:

$$w_{b+K} = w_b + \alpha_K \times \vec{n}, \quad K = 1, 2, \dots, T-1, \quad (2)$$

where T denotes the total number of images. However, the assumption is too strong since consecutive images are likely to vary from one to the others in different attributes. On the

other hand, the specific semantic direction \vec{n} can be predefined [19, 22, 21], but the semantically equivalent input images are required. To cope with images with arbitrary semantic variations, we regard the direction as one of our optimization targets. Note that the scaling factors α_K -s are also learnable, and therefore, we absorb them into the direction and reformulate Eq.(2) as the proposed mutually accessible constraint, which is as follows:

$$w_{b+K} = w_b + \sum_{k=1}^K \vec{n}_k. \quad (3)$$

In this way, we can figure out the latent codes of all the images via a joint optimization of w_b and \vec{n}_k -s. Such a simple linear combination mechanism can promote the editability of the latent codes. The main reasons are that, i) each of the inverted latent codes can be regarded as an edited code with respect to one of the others, and ii) if the images vary in a specific semantic direction, the scales of variations could be learned adaptively, and more importantly, iii) it is able to deal with the variations of different attributes among consecutive images. Moreover, the learned \vec{n}_k -s have the potential to be transferred to other latent codes as predefined semantic directions.

Inversion Consistency Constraint. Once we have the latent codes, we feed them into the generator $G(\cdot)$ to reconstruct consecutive images. For a certain base image I_b , its reconstruction O_b with regard to the latent code w_b is calculated by

$$O_b = G(w_b), \quad b = 1, 2, \dots, T. \quad (4)$$

In order to ensure a fidelity of the reconstructions, we particularly consider an inversion consistency between common regions of the reconstructed images and the input consecutive images. Specifically, we tailor an inversion consistency constraint loss based on bidirectional flows in the RGB space. As shown in the bottom of Fig. 2, first, the forward flows $f_{b \Rightarrow b+k}$ between base image I_b and the other images I_{b+k} can be calculated by a pretrained FlowNet2 [7] $F(\cdot)$, which is formulated as follows:

$$f_{b \Rightarrow b+k} = F(I_b, I_{b+k}), \quad (5)$$

where $k \in \mathbb{Z} \cap [1 - b, T - b] - \{0\}$. Then we warp I_b with this flow to form the warped images \hat{I}_{b+k} , which can be presented as follows:

$$\hat{I}_{b+k} = \text{warp}(I_b, f_{b \Rightarrow b+k}). \quad (6)$$

Also, we have the warped results \hat{O}_{b+k} of the recovered images O_b , which is presented as follows:

$$\hat{O}_{b+k} = \text{warp}(O_b, f_{b \Rightarrow b+k}). \quad (7)$$

Since consecutive images describe a same subject, an inherent relationship should be existed among the generated warpings $\{\hat{I}_{b+k}\}$ for each base image I_b . And this relationship should be transferred to the other warpings $\{\hat{O}_{b+k}\}$. In the same way, we can calculate the backward flow $f_{b+k \Rightarrow b}$ between I_{b+k} and I_b , and the corresponding warpings $\{\hat{I}_b^{b+k}\}, \{\hat{O}_b^{b+k}\}$. By iterating over all the values of b and k , we inject multi-source supervision from the warpings of input images to confine the reconstructions, and the proposed inversion consistency constraint loss \mathcal{L}_{ICC} is given by

$$\mathcal{L}_{ICC} = \sum_b \sum_k (\|\hat{I}_b^{b+k} - \hat{O}_b^{b+k}\|_2 + \|\hat{I}_{b+k} - \hat{O}_{b+k}\|_2), \quad (8)$$

where $\|\cdot\|_2$ denotes a pixel-wise L_2 distance.

Moreover, we consider to maintain a pixel-wise consistency between the input images and its corresponding recoveries. A pixel-wise consistency loss \mathcal{L}_C is therefore introduced in our objective, that is

$$\mathcal{L}_C = \sum_{m \in q \cup \{0\}} \|O_{1+m} - I_{1+m}\|_2, \quad (9)$$

To guarantee a fine visual perception of the reconstructions, we also utilize a perceptual loss \mathcal{L}_P which is formulated as follows:

$$\mathcal{L}_P = \frac{1}{4} \sum_{j=1}^4 \sum_{m \in q \cup \{0\}} \|\Phi_V^j(I_{1+m}) - \Phi_V^j(O_{1+m})\|_2, \quad (10)$$

where $\Phi_V^j(\cdot)$ denotes the j th layer of a pretrained VGG-16 network, and we follow Abdal *et al.* [1] to select the

features produced by the `conv1_1`, `conv1_2`, `conv3_2` and `conv4_2` layers of VGG-16 for modeling the loss.

Finally, the whole objective function \mathcal{L} is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ICC} + \lambda_2 \mathcal{L}_C + \lambda_3 \mathcal{L}_P, \quad (11)$$

where λ_s denote the balance factors. Then the latent code w_b and the directions $\vec{\mathbf{n}}_k$ can be optimized by

$$\{w_b^*, \vec{\mathbf{n}}_k^*\} = \arg \min_{\{w_b, \vec{\mathbf{n}}_k\}} \lambda_1 \mathcal{L}_{ICC} + \lambda_2 \mathcal{L}_C + \lambda_3 \mathcal{L}_P. \quad (12)$$

Note that we follow [1] that initialize w_b with the mean latent vector of \mathcal{W}_+ space. And the direction $\vec{\mathbf{n}}_k$ -s are initially set to zero and updated during optimization.

4. Experiments

4.1. Implementation Details

We implement the proposed method in Pytorch on a PC with an Nvidia GeForce RTX 3090. We utilize the generator of StyleGAN [9] pre-trained on the FFHQ dataset [9] with the resolution of 1024×1024 . The latent codes and semantic directions are optimized using Adam optimizer [11]. We follow [1] that use 5000 gradient descent steps with a learning rate of 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-8}$. We empirically set the balancing weights in Eq. (12) as $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 1$. And we set $T = 5$ in Eq. (4), which indicates 5 consecutive images are contained in each input sequence.

4.2. Experimental Settings

Datasets. We first conduct our experiments on the RAVDESS dataset [12] with real videos. The original RAVDESS dataset contains 2,452 videos with 24 subjects speaking and singing with various semantic expressions. We select 12 videos of them for evaluation, resulting in 1,454 frames, we name this dataset as RAVDESS-12 Dataset. Since there are no ground truth latent codes for real images, we cannot subjectively evaluate the inverted code and its editability in the latent space. On the other hand, it is demonstrated that the learned semantic directions work very well in generated images. As a result, we construct a synthesized dataset, containing 1000 samples that were randomly generated by StyleGAN. For each sample, we vary its latent code with 5 random combinations of the α value (ranging from -3 to 3) and semantic direction (acquired from InterfaceGAN [19]), producing 5000 images. We record the latent codes of the original samples, the corresponding editing specifications, and the edited latent codes for evaluating the editability. GAN inversion methods will invert the original samples and edit them to target attributes for comparisons.

Competitors. We mainly compare with four GAN inversion methods: Image2StyleGAN (I2S) [1], Image2StyleGAN++ (I2S++) [2], In-domain Inversion

Table 1: Comparisons with existing GAN inversion methods on image reconstruction with four metrics on two datasets. \downarrow denotes the lower the better and the best results are marked in **bold**.

Methods	Metrics	RAVDESS-12 Dataset				Synthesized Dataset			
		NIQE \downarrow	FID \downarrow	LPIPS \downarrow	MSE $\downarrow(\times e-3)$	NIQE \downarrow	FID \downarrow	LPIPS \downarrow	MSE $\downarrow(\times e-3)$
I2S [1]		3.770	16.284	0.162	8.791	3.374	48.909	0.271	35.011
pSp [18]		3.668	29.701	0.202	22.337	3.910	84.355	0.391	46.244
InD [24]		3.765	18.135	0.193	9.963	3.152	42.773	0.352	44.645
Ours		3.596	13.136	0.148	5.972	2.807	37.225	0.250	24.395
I2S++ [2]		3.358	0.320	0.003	0.174	2.644	2.967	0.014	1.458
Ours++		3.352	0.311	0.003	0.165	2.597	2.897	0.014	1.432



(a) Original (b) I2S (c) pSp (d) InD (e) Ours (f) I2S++(g) Ours++
 Figure 3: Qualitative comparison on image reconstruction. Compared with the works that optimized in the $\mathcal{W}+$ space (left part), our method can reconstruct the most faithful appearances. Involving the \mathcal{N} space largely improves reconstruction (right part), but Ours++ show better color preservation than I2S++ (second row).

(InD) [24], and pSp network [18]. All the methods are inverted to the same $\mathcal{W}+$ latent space of StyleGAN, applying the same directions for editing. It is worth noting that I2S++ introduces the additional noise space \mathcal{N} for small details recovery. A main problem is that, the inverted two latent codes in the $\mathcal{W}+$ and \mathcal{N} spaces are highly coupled, but the learned semantic directions are optimized in $\mathcal{W}+$ only. Applying them changes the $\mathcal{W}+$ space latent code but leaves the noise vector unchanged, these unpaired vectors yield “ghosting” artifacts after editing (see Fig. 4). As a result, we mainly use it for reconstruction comparisons, and we also extend our method to include the noise space, named as Ours++.

Evaluation Metrics. For the quantitative comparisons, we use four metrics, Naturalness Image Quality Evaluator (NIQE) [14], Fréchet Inception Distance (FID) [6], Learned Perceptual Image Patch Similarity (LPIPS) [23], and pixel-wise Mean Square Error (MSE), for evaluating the reconstruction fidelity. Especially, FID computes the Wasserstein-2 distance between the distribution of input and output images. NIQE evaluates the quality perceived by a human, which is a completely blind assessment with no re-

Table 2: Quantitative evaluation on real image manipulation with two blind metrics on the RAVDESS-12 Dataset. \downarrow denotes the lower the better and the best results are marked in **bold**.

Metrics	I2S [1]	pSp [18]	InD [24]	Ours
NIQE \downarrow	3.776	5.242	3.693	3.254
FID \downarrow	21.609	30.128	19.271	15.482

quest for the GT image. Since there is no GT for the semantic editing task on the real RAVDESS-12 Dataset, we use FID and NIQE to evaluate real image editing results.

4.3. Evaluation on Image Reconstruction

Quantitative Evaluation. We first evaluate the reconstruction fidelity of the inverted codes. Quantitative comparisons are shown in Tab. 1. We can see that our method significantly outperforms three editable GAN inversion methods (upper part) on both the real dataset and the synthesized one. Especially for the pixel-wise difference metric MSE, we largely improve the state-of-the-art by 31%. This indicates that the proposed joint optimization successfully incorporates complementary information from neighboring images. Besides, by involving the noise space \mathcal{N} , I2S++ and Ours++ achieve the most faithful reconstruction compared with other methods. Thanks to the inter-image coherence, we further push the reconstruction record a bit.

Qualitative Evaluation. Qualitative comparisons are shown in Fig. 3. We can see that the Image2StyleGAN cannot recover the image color correctly. Meanwhile, the pSp and InD cannot recover the finest facial details of the original images (see teeth in the first row). Compared with the above three methods, our method can reconstruct faithful appearance details. Unsurprisingly, I2S++ recovers the finest details among all competitors. That mainly because of their optimization performed in the \mathcal{N} space encodes high-frequency information. We also depict our results optimized in \mathcal{N} space, and we preserve the original color better than I2S++ (see the second row).

4.4. Evaluation on Image Editing

In this section, we evaluate our GAN inversion method on real image editing as well as synthesized images. We



Figure 4: Qualitative comparison on semantic editing with Pose and Age attributes on the real RAVDESS-12 Dataset. Images marked by red boxes are the reconstructed targets, and images in the middle row of each example are the inversion results. We can tell that our method can support more favorable semantic editing.

Table 3: Quantitative evaluation on image manipulation with four metrics on the Synthesized Dataset. \downarrow denotes the lower the better and the best results are marked in **bold**.

Metrics	I2S [1]	pSp [18]	InD [24]	Ours
NIQE \downarrow	3.390	3.917	3.193	3.163
FID \downarrow	35.894	58.342	48.867	33.872
LPIPS \downarrow	0.399	0.452	0.424	0.347
MSE \downarrow ($\times e-3$)	89.671	126.642	101.563	70.224

conduct two editing tasks based on the inverted latent codes, the first one is semantic manipulation and the second is image morphing.

4.4.1 Semantic Manipulation

Semantic manipulation aims to edit a real image by varying its inverted codes along with a specific semantic direction. We use five semantic directions (*i.e.*, gender, pose, smile, eyeglasses, and age) acquired by [19] in the experiment.

Qualitative Evaluation. The qualitative comparisons on real data are shown in Fig. 4. We can see that our manipulated faces have visually more plausible results than those of the competitors. In particular, the manipulated results gained by Image2StyleGAN [1] present noisy artifacts with pose changes and glasses are entangled with the age attributes, revealing that the edited latent codes are escaped from the editable domain. Similar situations can be found

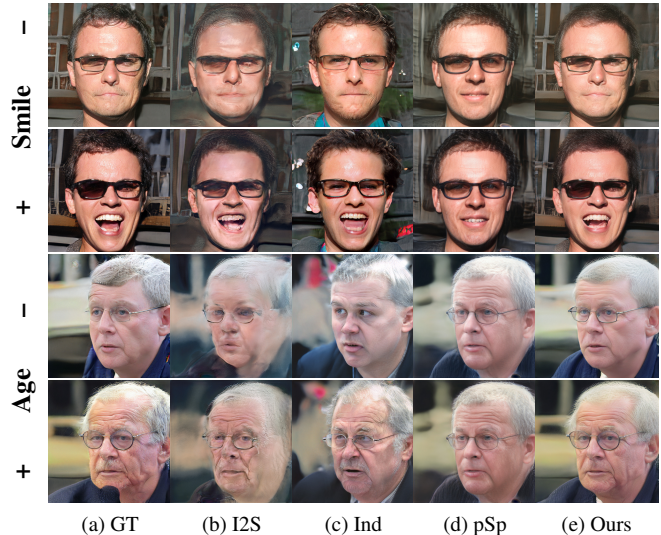


Figure 5: Qualitative comparison on semantic editing with Smile and Age attributes on the Synthesized Dataset. It’s noticed that we force the input sequence to contain different semantic changes from its corresponding editing direction. the first sequence contains the semantic changes with “gender” for optimization and “smile” for semantic edit, and the second is “pose” for optimization while “age” for the edit testing. And our edited results are more similar with the ground truths.

in I2S++ [2]. The manipulated faces based on pSp [18] are almost unchanged. This is because pSp focuses on learning a direct mapping from the input to latent code, ignoring the editability. This problem is addressed by In-domain inversion [24], but it also sacrifices the reconstruction quality. In contrast, thanks to the jointly considered inherent editability constraint between consecutive images, our inverted latent codes are more semantically editable, leading to more disentangled manipulations. On the other hand, the noise space optimization methods (right part of Fig. 4) show obvious noise artifacts than the others with pose changes, this is because the pre-optimized noise vector is not suitable for the edited latent vector w . However, Ours++ can disentangle gender from glasses better than the inverted vector from I2S++.

To evaluate whether the obtained inversion can be edited by arbitrary directions, we force the editing direction different from the semantic changes contained the input sequence on the synthesized dataset. The qualitative comparisons on the synthesized data are shown in Fig. 5. Similar to the evaluation on real data, I2S produces obvious artifacts, pSp fails to edit the results, and InD cannot preserve the original identity. Our manipulated results are more similar with the GTs, which indicates that our inverted codes are much closer with the GT latent codes and also inherit their ed-

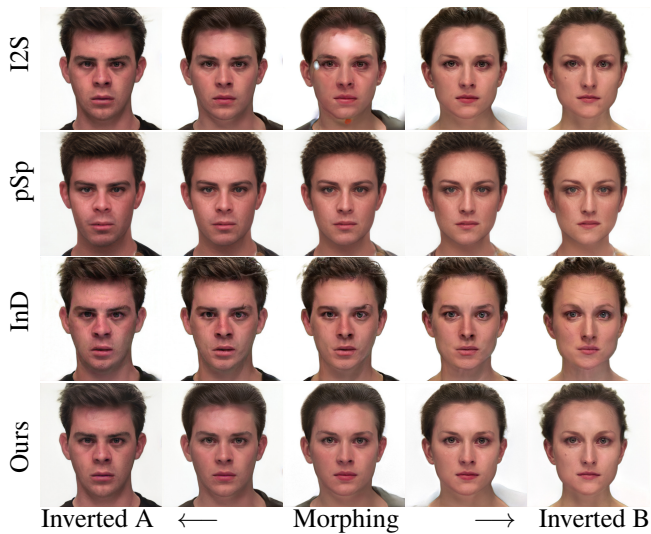


Figure 6: Qualitative comparison on image morphing task. We can see that our result present a continuous process and the morphing faces are realistic.

itability.

Quantitative Evaluation. We present the quantitative comparisons in Tab. 2 and Tab. 3. Our method achieves the best results on both the RAVDESS-12 Dataset and the Synthesized Dataset. In particular, for the blind metric NIQE, our edited results achieve 13.8% improvement over the state-of-the-art method, which indicates that our editing is more visually plausible. Quantitative results on the Synthesized Dataset can evaluate whether the inverted codes are close enough with the GT code such that we can reuse their semantic information. From two non-blind metrics LPIPS and MSE, we can see that our edited results are very similar to the GTs. Thanks to our semantically accessible regularization in the latent space, our inverted latent codes show a strong editability compared with the competitors.

4.4.2 Image Morphing

Image morphing aims to fuse two images semantically by interpolating their latent codes. It is another way to evaluate whether the inverted codes indeed lie in the latent space and reuse the semantic knowledge. For the high-quality inverted codes, their interpolated results should also stay in the editable domain and the semantic varies continuously. Qualitative comparisons are shown in Fig. 6. We can see that the morphing results produced by Image2StyleGAN [1] have noticeable artifacts. Meanwhile, the results produced by pSp [18] are unrealistic with the unnatural hairs. In contrast, our method presents high-quality results with a continuous morphing process. We also present the quantitative evaluation on the morphing task in Tab. 4 and Tab. 5, we can see that our inversion results outperform the other inversion methods on both the real dataset and the synthesized one.

Table 4: Quantitative evaluation on image morphing with two blind metrics on the RAVDESS-12 Dataset. \downarrow denotes the lower the better and the best results are marked in **bold**.

Metrics	I2S [1]	pSp [18]	InD [24]	Ours
NIQE \downarrow	4.255	5.350	4.051	3.688
FID \downarrow	40.627	38.474	38.925	37.695

Table 5: Quantitative evaluation on image morphing with four metrics on the Synthesized Dataset. \downarrow denotes the lower the better and the best results are marked in **bold**.

Metrics	I2S [1]	pSp [18]	InD [24]	Ours
NIQE \downarrow	3.389	3.800	3.212	3.115
FID \downarrow	31.776	30.192	21.901	18.621
LPIPS \downarrow	0.472	0.467	0.469	0.402
MSE $\downarrow(\times e-3)$	141.432	121.834	125.674	98.354

Table 6: Ablation study on image reconstruction with four metrics. \downarrow denotes the lower the better and the best results are marked in **bold**.

Variants	NIQE \downarrow	FID \downarrow	LPIPS \downarrow	MSE $\downarrow(\times e-3)$
Baseline	3.770	16.284	0.162	8.791
w/o MAC	3.685	13.375	0.151	8.065
w/o ICC	3.765	14.791	0.160	8.508
Ours	3.596	13.136	0.148	5.972

4.5. Semantic Transfer

As discussed in Sec. 3.2, both the latent code w and the semantic direction \vec{n} can be unsupervisedly obtained after inversion. Besides the latent code, our acquired direction \vec{n} represents the semantic changes of the input images. Given the input images as reference, we can transfer its semantic changes to the target faces.

The transfer results are shown in Fig. 7. We can see that the semantic attributes of target faces are modified following the reference image set. Note that there are more than one attribute has been changed in the reference. For example, in the right example, the mouth and pose varies simultaneously but we can still capture those changes. This shows that our acquired direction is disentangled with the referenced and can be applied on other faces. Other than existing supervised [19, 20] or unsupervised [21, 22] learning of interpretable directions, this sheds light on a new exemplar-based learning of semantic direction.

4.6. Ablation Study

In this section, we analyze the efficacy of our two components: mutually accessible constraint (MAC) and inversion consistency constraint (ICC). Note that without these two components, our method equals to the Image2StyleGAN inversion and we set it as our baseline. By unplugging one of these two constraints, we yield two variants of “w/o MAC” and “w/o ICC”. In this case, \mathcal{L}_{ICC} is removed and all the latent codes are optimized simultaneously.



Figure 7: Our unsupervisedly acquired direction \vec{n} from consecutive images can be used for transferring the semantics. The first row is the input set that is regarded as reference, and the images in red boxes are the target faces. We can transfer the semantic changes of the reference to the target faces, even with more than one attribute changed.

Table 7: Ablation study on semantic manipulation with two blind metrics. \downarrow denotes the lower the better and the best results are marked in **bold**.

Metrics	Baseline	<i>w/o</i> MAC	<i>w/o</i> ICC	Ours
NIQE \downarrow	3.776	3.659	3.398	3.254
FID \downarrow	21.609	16.121.	17.274	15.482

We perform ablation study experiment on image reconstruction and semantic manipulation tasks on the RAVDSS-12 Dataset. Quantitative comparisons of GAN inversion are shown in Tab. 6. We can see that every variant outperforms the baseline on all metrics. This indicates both two components contribute to the GAN inversion performance. Meanwhile, variant (*w/o* MAC) performs better than the variant (*w/o* ICC), this indicates that the inversion consistency brought by consecutive images contributes more for the GAN inversion task. In Tab. 7 of semantic editing, we observe a different situation. We can see the variant (*w/o* ICC) performs better than the variant (*w/o* MAC), this reveals that mutually accessible constraint confines the inverted latent codes to stay in the editable domain. The above two evaluations show that our two constraints work very well following our design principles.

We show the results of different variants in Fig. 8 by changing the “age” attribute. We can see that the baseline and variant (*w/o* MAC) entangles with glasses, showing that concentrating only on reconstruction fidelity lacks editability of the inverted codes. In contrast, variant (*w/o* ICC) and our final result can successfully modify the “age” attribute, revealing the strong regularization power of our designed mutually accessible constraint.

5. Conclusion

In this paper, we propose an alternative GAN inversion method for consecutive images, we formulate consecutive

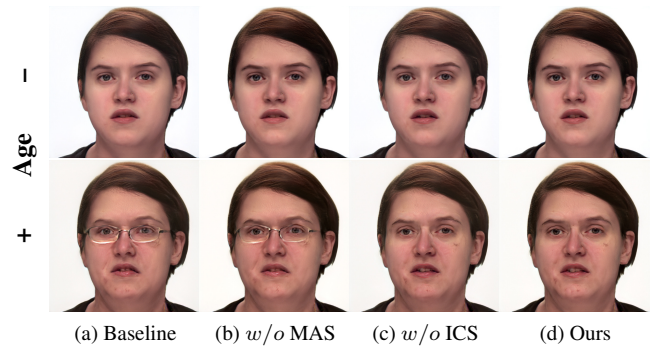


Figure 8: Ablation study on semantic editing with two variants and baseline by editing the “age” attribute.

images inversion as a linear combination process in the latent space that ensures editability, and transfer the reconstruction consistency across inputs in the RGB space to guarantee reconstruction fidelity. The experiment results demonstrate the effectiveness in terms of editability and reconstruction fidelity. Besides, we also support various new applications like video-based GAN inversion and unsupervised semantic transfer.

Acknowledgment: This project is supported by the Key-Area Research and Development Program of Guangdong Province, China (2020B010165004, 2020B010166003, 2018B010107003); National Natural Science Foundation of China (61972162, 61772206, U1611461, 61472145); Guangdong International Science and Technology Cooperation Project (2021A0505030009); Guangdong Natural Science Foundation (2021A1515012625); Guangzhou Basic and Applied Research Project (202102021074); CCF-Tencent Open Research fund (CCF-Tencent RAGR20190112, RAGR20210114); China Postdoctoral Science Foundation (2020M682240, 2021T140631) and Fundamental Research Funds for the Central Universities (202113035).

References

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, pages 4432–4441, 2019. 1, 2, 4, 5, 6, 7
- [2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, pages 8296–8305, 2020. 1, 2, 4, 5, 6
- [3] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE TNNLS*, 30(7):1967–1974, 2018. 2
- [4] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, pages 5744–5753, 2019. 1, 2
- [5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, volume 27, 2014. 1
- [6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 5
- [7] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017. 4
- [8] Ali Jahanian, Lucy Chai, and Phillip Isola. On the “steerability” of generative adversarial networks. 2019. 1, 2
- [9] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1, 3, 4
- [10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 1, 2
- [11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4
- [12] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 4
- [13] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. Invertibility of convolutional generative networks from partial measurements. In *NeurIPS*, volume 31, 2018. 2
- [14] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *Signal Process Letter*, 20(3):209–212, 2012. 5
- [15] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, pages 262–277, 2020. 2
- [16] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. 2016. 2
- [17] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2015. 1, 2
- [18] 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. In *CVPR*, pages 2287–2296, 2021. 1, 2, 5, 6, 7
- [19] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, pages 9243–9252, 2020. 1, 2, 3, 4, 6, 7
- [20] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE TPAMI*, 2020. 7
- [21] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, pages 1532–1540, 2021. 3, 7
- [22] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*, 2020. 1, 2, 3, 7
- [23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5
- [24] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020. 1, 2, 5, 6, 7
- [25] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, pages 597–613, 2016. 2