LSAP: RETHINKING INVERSION FIDELITY, PERCEPTION AND EDITABILITY IN GAN LATENT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

As the methods evolve, inversion is mainly divided into two steps. The first step is *Image Embedding*, in which an encoder or optimization process embeds images to get the corresponding latent codes. Afterward, the second step aims to refine the inversion and editing results, which we named *Result Refinement*. Although the second step significantly improves fidelity, perception and editability are almost unchanged, deeply dependent on inverse latent codes attained in the first step. Therefore, a crucial problem is gaining the latent codes with better perception and editability while retaining the reconstruction fidelity. In this work, we first point out that these two characteristics are related to the degree of alignment (or disalignment) of the inverse codes with the synthetic distribution. Then, we propose **L**atent **S**pace **A**lignment Inversion **P**aradigm (LSAP), which consists of evaluation metric and solution for this problem. Specifically, we introduce Normalized Style Space ($\mathcal{S}^{\mathcal{N}}$ space) and $\mathcal{S}^{\mathcal{N}}$ **C**osine **D**istance (SNCD) to measure disalignment of inversion methods. Since our proposed SNCD is differentiable, it can be optimized in both encoder-based and optimization-based embedding methods to conduct a uniform solution. Extensive experiments in various domains demonstrate that SNCD effectively reflects perception and editability, and our alignment paradigm archives the state-of-the-art in both two steps.

1 INTRODUCTION

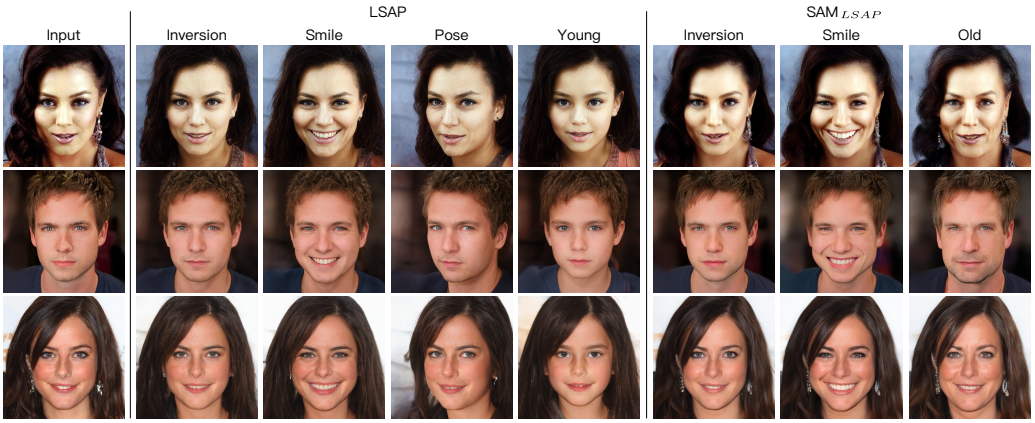


Figure 1: **Inversion and editing results from LSAP and SAM$_{LSAP}$(Parmar et al., 2022)**. Our method improves image quality and editability while retaining fidelity. It is compatible with the hybrid method and achieves better performance.

In recent years, Generative Adversarial Networks (GANs)(Goodfellow et al., 2014) are used in various tasks(Ledig et al., 2017; Yang et al., 2021b) and have dramatically improved image synthesis ability. Style-based generative models(Karras et al., 2019; 2020; 2021) further enhance the realism and resolution of image generation, achieving state-of-the-art. The intermediate latent space $\mathcal{W}$ space in StyleGAN encodes high-semantic information. As a strong prior, well-trained generator

has demonstrated powerful capabilities and improved multiple tasks from traditional approaches, e.g., neural talking head(Prajwal et al., 2020; Yin et al., 2022), face parsing(Yang et al., 2021a; Zhang et al., 2021), and style transfer(Li et al., 2020; Yang et al., 2022b).

These applications require latent codes, which are inherently available for synthetic images but cannot be applied directly to real images. To this end, inversion methods are designed to embed images into GAN's latent space via various approaches. Existing works can be mainly divided into two steps. The first step aims to attain latent codes, usually achieved by training an encoder or optimizing the reconstruction error, which we named *Image Embedding*. In the second step, researchers employ diversiform strategies to improve inversion and editing results, e.g., predicting generator weights(Alaluf et al., 2022; Dinh et al., 2022), and finetuning the generator(Roich et al., 2021; Feng et al., 2022), which we named *Result Refinement*. Previous worksTov et al. (2021) illustrate that fidelity, perception, and editability are three essential characteristics of inversion. However, in *Result Refinement*, more attention has been paid to improve fidelity, maintaining visual details like the background, hat, and eyeglasses while inheriting editability and perception from the inverse codes in the first step. Hence, in order to achieve superior performance in fidelity, perception, and editability, a robust latent code embedding technique which compatible with refinement mechanisms is still necessary.

The latent space from random sampling and transformation possesses a particular distribution, which we named synthetic distribution. Intuitively, latent codes from this distribution have better performance. Supervision from the discriminator constrains the sampled latent codes to generate photo-realistic images. Moreover, editing directions are gained by sampling(Shen et al., 2020) and analyzing(Härkönen et al., 2020) in synthetic latent space. Hence, the key point of perception and editability is the alignment between inverse codes and synthetic distribution. An existing method (Tov et al., 2021) solves this problem by latent code discriminator and achieves more reasonable perception and editability. However, there are two significant shortcomings. Firstly, it limits the reconstruction performance since introducing a discriminator makes training unstable. Secondly, this approach cannot be applied to the optimization-based inversion methods. Therefore, our key motivation is the idea of constructing an alignment paradigm between embedding latent space and synthetic latent space which can be applied to both encoder-based and optimization-based inversion methods and retains reconstruction ability.

In this work, we thoroughly analyze the disalignment in inversion and propose the **L**atent **S**pace **A**lignment Inversion **P**aradigm (LSAP). Specifically, we first introduce the Normalized Style Space ($\mathcal{S}^{\mathcal{N}}$ space). We prove that $\mathcal{S}^{\mathcal{N}}$ space is more suitable and efficient for measuring disalignment than $\mathcal{Z}/\mathcal{W}/\mathcal{S}$ space. Moreover, we introduce a metric $\mathcal{S}^{\mathcal{N}}$ Cosine Distance (SNCD) to evaluate the inversion methods at latent code level, which have shown experimentally reflecting perception and editability. Then, we conduct the alignment solution in encoder-based and optimization-based methods, employing an alignment loss based on SNCD. We present extensive experiments to demonstrate the effects and generality of our alignment paradigm. We achieve the best trade-offs in encoder-based methods and drastically improve the perception and editability in the optimization-based method. Besides, we reach the state-of-the-art with HFGI(Wang et al., 2022) and SAM(Parmar et al., 2022), which further demonstrates the potential of our method. As shown in Figure 1, our visual results are natural and faithful. The key contributions of this work are summarized as follows:

- We rethink the fidelity, perception and editability in inversion task. As dividing inversion process into *Image Embedding* and *Result Refinement*, we point out that fidelity is enhanced in the second step while perception and fidelity are related to alignment between inverse codes and synthetic distribution.

- We propose an effective and generalized Latent Space Alignment Inversion Paradigm (LSAP), including measurement (i.e., SNCD) and improvement solutions (i.e., LSAP_E and LSAP_O) of perception and editability.

- To demonstrate the effect of our aligning paradigm, we take extensive experiments in various domains. SNCD reflects the perception and editability in a numerical way. Our alignment paradigm reaches better trade-offs between fidelity-perception and fidelity-editability. Applying to hybrid methods, LSAP_E achieves state-of-the-art.

2 RELATED WORK

GAN Inversion. Generally, the inversion process can be divided into two steps. Firstly, an initial latent code is gained by optimization or encoder from a given image. Optimizing reconstruction error typically reach better fidelity, while it requires several minutes per image (Karras et al., 2020; Creswell & Bharath, 2018; Abdal et al., 2019; 2020). Training an encoder(Tov et al., 2021; Richardson et al., 2021; Wei et al., 2022; Guan et al., 2020; Creswell & Bharath, 2018) to invert images is efficient during inference but achieves inferior reconstruction results. The second step aims to refine the inversion and editing results, using various strategies. Some methods(Alaluf et al., 2022; Dinh et al., 2022) adjust the convolution weights of the generator by hypernetwork(Ha et al., 2016). ReStyle(Alaluf et al., 2021) introduces an iterative refinement mechanism, refining the latent code by a residual-based encoder. HFGI(Wang et al., 2022) proposes a distortion consultation approach for high-fidelity reconstruction. SAM(Parmar et al., 2022) inverses the different segments of image into different intermediate layer by predicting "invertibility". Generator tuning (Roich et al., 2021; Feng et al., 2022) can get the best inversion performance but is considerably time-consuming.

GAN-based Manipulation. Thanks to the rich semantic information of GAN's(Karras et al., 2019; 2020; 2021) latent space, many works have proposed various methods to control generated results by manipulating latent representation. Some methods(Denton et al., 2019; Goetschalckx et al., 2019; Spingarn-Eliezer et al., 2020; Shen et al., 2020) find editing directions of attributes (e.g., smile, gender, age, and pose) by semantic labels. Others find meaningful directions in an unsupervised (Härkönen et al., 2020; Shen & Zhou, 2021; Voynov & Babenko, 2020; Wang & Ponce, 2021) or self-supervised(Jahanian et al., 2019; Plumerault et al., 2020) way. Moreover, language-image models are explored to edit images by back-propagating the gradient of objective text(Patashnik et al., 2021). Some works(Sun et al., 2021; Kim et al., 2022) further introduce segmentation information to gain better performance, which may be extended to the human body by body GANs(Frühstück et al., 2022; Fu et al., 2022) and human parsing techniques(Yang et al., 2019; 2020; 2021a; 2022a) in the future. Since those manipulation approaches are almost built on latent codes, editability is also a crucial characteristic of inversion.

3 DISALIGNMENT IN LATENT SPACE

In this section, we first rethink the source of fidelity, perception and editability, and we point out that the latter two are deeply related to the alignment between inverse codes and synthetic distribution. To illustrate the alignment (or disalignment) in existing inversion methods, we further formulate it and introduce a new latent space Normalized Style Space ($\mathcal{S}^{\mathcal{N}}$ space) to better measure the discrepancy between latent codes.

3.1 FIDELITY, PERCEPTION AND EDITABILITY

As first mentioned in Tov et al. (2021), fidelity[1], perception and editability are three characteristics of inversion methods. Fidelity measures the reconstruction ability, requiring methods embedding image into latent space which can reconstruct images similar to given images. Perception evaluates the reconstructed images' perceptual quality, which consists of sharpness and naturalness in practice, which we illustrate in Appendix E. Besides, editability represents editing capability of inverse codes, which is a comprehensive measurement, including editing effects, attributes disentanglement, etc.

Inversion process can be divided into two stages: *Image Embedding* and *Result Refinement*. The latent codes are first attained by an encoder or optimized by minimizing image distortion. In this phase, reconstruction error is slightly large. In *Result Refinement* step, methods focus on recovering visual details (e.g., background, cloths) by adjusting weights(Alaluf et al., 2022) or intermediate features(Wang et al., 2022) of generator. This stage further improves the fidelity and even can invert the out-of-distribution images(Roich et al., 2021; Feng et al., 2022). However, perception and editability are inherited from the first step, by inverse codes specifically. In other words, these two properties are not enhanced in this step and are even damaged. In practice, if the inverse codes cannot be edited or generate images with good perceptual quality, the refined results still show the

[1]Tov et al. (2021) originally uses the words *image distortion*. To represents the ability of inversion methods, we use *fidelity* instead of it.

same effect. Hence, a essential problem is obtaining latent codes with better performance. In this work, we focus on the first step to study fidelity, perception, and editability of latent codes.

Let us start by tracing these characteristics. Minimizing image distortion is a significant objective function applied in all inversion methods. It gives the algorithm the ability to reconstruct given images faithfully. Perception is gained by the powerful generating capability of GANs, since generator is trained to generate the photorealistic results with high resolution by a discriminator. Editability benefits from the highly semantic latent space of GANs. Given the editing direction, we can modify the latent codes to edit corresponding attribute. However, the ability of perception and editability is conditional. Specifically, under the constraint from the discriminator, latent space in GANs is required to fit dataset distribution, from which latent code can generate high-quality images. The generator may not generate good results from out-of-distribution latent code. That is also verified by latent code truncation, that codes near to mean code can generate high-quality results. Moreover, editing directions are obtained by sampling latent codes(Shen et al., 2020) or analyzing generator weights(Shen & Zhou, 2021). That is also built on a specific latent space in GAN.

We name the latent space distribution in GAN as synthetic distribution, which is converted by pretrained networks from multivariant normal standard distribution. As we analyzed, latent codes aligning with synthetic distribution have better perception and editability. In practice, Tov et al. (2021) introduces a latent code discriminator to solve this problem. However, it is worth mentioning that there are trade-offs between fidelity and perception/editability. Improving the latter characteristics implies the sacrifice of fidelity. It is significant to design a meticulous solution to improve perception and editability while retaining fidelity. Moreover, the metric which can be used to evaluate perception and editability in current study is missing. Although many perceptual metrics are widely used in visual tasks (e.g., FID(Heusel et al., 2017) and SWD(Rabin et al., 2011)), they are deeply influenced by image distortion in inversion and cannot faithfully evaluate the inversion results(Tov et al., 2021).

Since perception and editability are positively related to the alignment degree of inverse latent codes with synthetic distribution, solving the disalignment of inversion methods is a straightforward solution for improving them. An intuition idea is that construct a numerical measurement of disalignment and optimize it in inversion. We next analyze how to formulate it first and build the Latent Space Alignment Inversion Paradigm (LSAP), including the evaluation metric and solutions.

3.2 Disalignment Formulation

To illustrate disalignment between synthetic and inverse latent space, we firstly define $\mathcal{P}$ space as a reference space, denoting $\mathcal{P}_{inv}$ and $\mathcal{P}_{syn}$ as inverse and synthetic latent space, respectively. $G_{\mathcal{P}}$ is defined as the generator from $\mathcal{P}$ space to image space. Suppose that $\mathcal{Z}$ is the multivariate standard normal distribution and $\mathcal{X}$ is the real image distribution. We establish two mapping functions $F : \mathcal{Z} \to \mathcal{P}_{syn}$ and $I : \mathcal{X} \to \mathcal{P}_{inv}$. In practice, I serves as an embedding method, used to convert images into $\mathcal{P}$ space latent code. Moreover, F is a mapping function consisting of multiple parts in the front of generator. If we find a distance function L to measure latent codes, we can define disalignment $\mathcal{D}$ between these two spaces as follow:

$$\begin{aligned}\mathcal{D} &= \mathbb{E}_{p_{syn}\sim\mathcal{P}_{syn},p_{inv}\sim\mathcal{P}_{inv}}[L(p_{syn}, p_{inv})] \\ &= \mathbb{E}_{z\sim\mathcal{Z},x\sim\mathcal{X}}[L(F(z), I(x))].\end{aligned} \tag{1}$$

Hence, two vital parts of disalignment measurement (Equation 1) are:

- **Latent Space**: which latent space is effective to measure disalignment;
- **Distance Function**: how to measure the distance between latent codes.

We respectively answer these two questions in the following parts.

3.3 Normalized Style Space ($\mathcal{S}^{\mathcal{N}}$)

Although $\mathcal{Z}/\mathcal{W}/\mathcal{W}^+$ spaces are primarily popular in previous research, in this work, we propose a new latent space, **Normalized Style Space ($\mathcal{S}^{\mathcal{N}}$)**, and we will prove that it is better to measure disalignment.

Let us revisit the existing latent spaces first. Given the sampled random variable z from $\mathcal{Z}$ space, the mapping network first converts it into w in $\mathcal{W}$ space. Then affine modules are applied on w at each resolution level, which output is $s = \{s_1, s_2, \ldots, s_k\}$, where $s_i = A_i w + b_i$ and the output space is named Style Space ($\mathcal{S}$ space).

Property 3.1. *Suppose that $s = \{s_1, s_2, \ldots, s_k\}$ is a set of $\mathcal{S}$ space latent codes and corresponding to image $x = G_{\mathcal{S}}(s)$. For $\forall a \in \mathbb{R}$ and $\forall l \in [1, k]$, if $\hat{s}$ follows:*

$$\hat{s}_n = \begin{cases} s_n, & n \neq l \\ a \times s_n, & n = l \end{cases}$$

we have $x = G_{\mathcal{S}}(s) = G_{\mathcal{S}}(\hat{s})$.[2]

Property 3.1 illustrates that $\mathcal{S}$ space latent codes are scaled-independent in every component. Those with the same angles will generate the same results. Hence, converting codes in the unit hypersphere, we construct a new latent space, Normalized Style Space ($\mathcal{S}^{\mathcal{N}}$), in which codes are normalized from $\mathcal{S}$ space by the euclidean norm. It follows:

$$s_i^N = \frac{s_i}{\|s_i\|_2} = \frac{A_i w + b_i}{\|A_i w + b_i\|_2} \tag{2}$$

To demonstrate differences between each latent space in measuring disalignment, we take extensive analyses:

Property 3.2. *Given a sets of $\mathcal{S}$ space latent codes $s = \{s_1, \ldots, s_k\} \neq \mathbf{0}$, $\exists s' = \{s'_1, \ldots, s'_k\} \neq s$ such that $G_{\mathcal{S}}(s) = G_{\mathcal{S}}(s')$.*

Proof. According to Property 3.1, for $\forall l \in [1, k]$ when $s'_l = a \times s (a \in \mathbb{R})$ and $s'_i = s_i (i \neq l)$, we have $G_{\mathcal{S}}(s) = G_{\mathcal{S}}(s')$. Since $s_l \neq \mathbf{0}$, $s'_l \neq s_l$. □

Property 3.3. *For lth layer ($\forall l \in [1, k]$), define $F_l : \mathcal{Z}/\mathcal{W} \to \mathcal{S}$ as the mapping function between $\mathcal{S}$ and $\mathcal{Z}/\mathcal{W}$ space. For all $p_l \in \mathcal{Z}/\mathcal{W}$ ($F_l(p) \neq \mathbf{0}$), exist $p'_l \neq p_l$ such that the corresponding $\mathcal{S}$ space latent codes satisfy: $s'_l = a \times s_l$ ($a \in \mathbb{R}$), where $s_l = F_l(p_l)$ and $s'_l = F_l(p'_l)$.*[3]

Corollary 3.1. *Given a sets of latent codes $p = \{p_1, \ldots, p_k\}$ in $\mathcal{Z}/\mathcal{W}/\mathcal{S}$ space and $p \neq \mathbf{0}$, $\exists p' = \{p'_1, \ldots, p'_k\} \neq p$ such that $G_{\mathcal{P}}(p) = G_{\mathcal{P}}(p')$.*

According to Corollary 3.1, different latent codes in $\mathcal{Z}/\mathcal{W}/\mathcal{S}$ space can generate the same images, which implies the distance of these latent codes can not reflect discrepancies in generated results. Hence, we choose $\mathcal{S}^{\mathcal{N}}$ as reference space to measure disalignment in inversion.

3.4 Cosine Distance

Illustrated in Equation 1, the distance function is required to measure the discrepancy between pairwise latent codes. When we denote $\mathcal{S}^{\mathcal{N}}$ space as reference space, an intuition distance function is the cosine distance:

$$L = 1 - \cos(s_{syn}^N, s_{inv}^N) \tag{3}$$

Since s_{syn}^N and s_{inv}^N have unit norm, we have:

$$L = 1 - s_{syn}^N \cdot (s_{inv}^N)^T \tag{4}$$

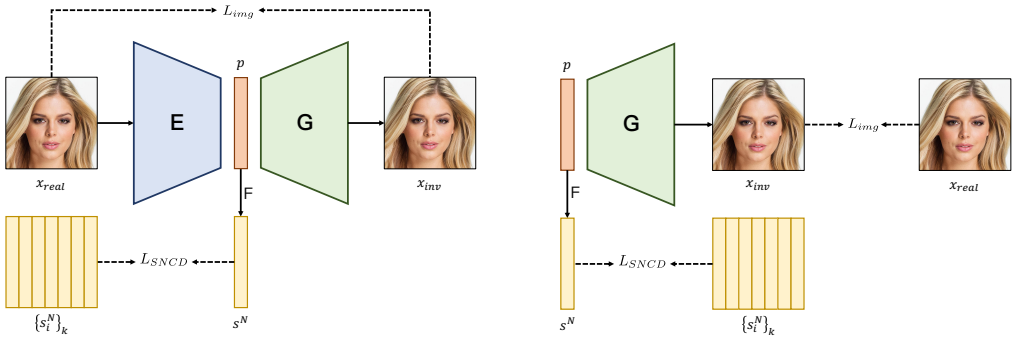
Notably, as s_{inv}^N is the inversion result, cosine distance is differentiable for it to minimize in inversion process.

4 Latent Space Alignment Inversion Paradigm

In this section, we construct the **L**atent **S**pace **A**lignment Inversion **P**aradigm (LSAP) to measure and improve perception and editability of inversion methods. Specifically, we propose the $\mathcal{S}^{\mathcal{N}}$ Cosine Distance (SNCD) and generalized alignment solutions in *Image Embedding* phase, including LSAP_E and LSAP_O for encoder-based and optimization-based methods, respectively.

[2] Proof can be found in Appendix A.1.

[3] Proof can be found in Appendix A.2.



(a) LSAP$_E$ (b) LSAP$_O$

Figure 2: **Alignment Inversion Solutions of LSAP.** We show the details of encoder-based and optimization-based inversion methods in our alignment paradigm. The pivotal part is the L_{SNCD}, which represents the disalignment degree of inverse latent codes.

4.1 $\mathcal{S}^{\mathcal{N}}$ Cosine Distance

In § 3, we analyze how to formulate disalignment and determine the reference latent space and distance function. Based on them, we propose a new evaluation metric $\mathcal{S}^{\mathcal{N}}$ Cosine Distance (SNCD) as:

$$SNCD = 1 - \mathbb{E}_{s^N_{syn} \sim \mathcal{S}^{\mathcal{N}}_{syn}, s^N_{inv} \sim \mathcal{S}^{\mathcal{N}}_{inv}} [s^N_{syn} \cdot (s^N_{inv})^T] \tag{5}$$

Since s^N_{syn} and s^N_{inv} are independent, we have:

$$SNCD = 1 - \mathbb{E}_{s^N_{syn} \sim \mathcal{S}^{\mathcal{N}}_{syn}} [s^N_{syn}] \cdot \mathbb{E}_{s^N_{inv} \sim \mathcal{S}^{\mathcal{N}}_{inv}} [s^N_{inv}]^T \tag{6}$$

The small value of SNCD means that $\mathcal{S}^{\mathcal{N}}_{inv}$ space aligns with $\mathcal{S}^{\mathcal{N}}_{syn}$ space. SNCD can reflect the perception and editability in image level, which we will show in qualitative and quantitative experiments.

4.2 Alignment Inversion

Inversion methods in *Image Embedding* phase aim to embed images into latent space in encoder-based or optimization-based way, and the process is as follows:

$$p^* = \arg\min_p [\mathcal{L}(x, G_{\mathcal{P}}(p)] \tag{7}$$

$$E^* = \arg\min_E [\mathbb{E}_{x \sim \mathcal{X}} (\mathcal{L}(x, G_{\mathcal{P}}(E(x))))] \tag{8}$$

where x is given image, $\mathcal{X}$ is image dataset, $\mathcal{L}$ is image level loss function (e.g., MSE, LPIPS(Zhang et al., 2018), identity loss(Deng et al., 2019)) and E is an encoder. Since inversion methods are mainly supervised at the image level, there is a lack of limitation of inverse latent space distribution. To construct a uniform solution to train encoder or optimize the latent codes, we can supervise the disalignment degree and add an alignment term in $\mathcal{L}$. According to Equation 6, disalignment in inversion is modeled by:

$$SNCD = 1 - \mathbb{E}_{s^N_{syn} \sim \mathcal{S}^{\mathcal{N}}_{syn}} [s^N_{syn}] \cdot \mathbb{E}_{x \sim \mathcal{X}} [F(I(x))]^T \tag{9}$$

As $\mathbb{E}_{s^N_{syn} \sim \mathcal{S}^{\mathcal{N}}_{syn}} [s^N_{syn}]$ is independent to inversion, solving disalignment is exactly aligning the output space of I with synthetic latent space. Thanks to SNCD's differentiable property, we can apply it to inversion methods to construct a direct and efficient alignment solution, shown in Figure 2. According to Equation 9, we first sample k latent codes from the multivariant normal distribution and convert them into $\mathcal{S}^{\mathcal{N}}$ by pre-trained generator to simulate the synthetic distribution $\mathcal{S}^{\mathcal{N}}_{syn}$, which

Table 1: **Quantity results on face domain.** We show the quantity results of encoder-based, optimization-based, and hybrid methods, respectively. LSAP-based methods achieve better distortion-perception/editability trade-offs, for our method slightly damages the reconstruction performance and attains superior perception and editability ability. - means evaluation is unavailable for unstable results or time-consuming. $\times$ means the generated result is highly unnatural.

	Fidelity			Perception & editability						
Metric	MSE ↓	LPIPS ↓	Similarity ↑	SNCD ↓	LEC_{pose} ↓	LEC_{smile} ↓	LEC_{age} ↓	Similarity_{pose} ↑	$\text{Similarity}_{smile}$ ↑	Similarity_{age} ↑
pSp	**0.0351**	**0.1628**	**0.5591**	0.1013	89.3529	55.8694	64.6177	0.4374	**0.4785**	0.3077
e4e	0.0475	0.1991	0.4966	0.0495	26.6551	22.3202	23.2861	0.4178	0.4173	0.3441
LSAP_E	0.0397	0.1766	0.5305	**0.0385**	**19.0211**	**14.0360**	**14.6715**	**0.4597**	0.4519	**0.3944**
StyleGAN2-$\mathcal{W}$	0.0696	0.1987	0.3066	0.0656	-	-	-	0.2389	0.2470	$\times$
LSAP_O-$\mathcal{W}$	0.0690	0.1986	0.2989	0.0492	-	-	-	0.2182	0.2362	$\times$
StyleGAN2-$\mathcal{W}^+$	**0.0279**	**0.1179**	**0.7463**	0.1063	-	-	-	$\times$	$\times$	$\times$
LSAP_O-$\mathcal{W}^+$	0.0359	0.1376	0.6587	**0.0407**	-	-	-	**0.4846**	**0.5161**	$\times$
HFGI_{e4e}	0.0210	0.1172	0.6816	-	-	-	-	$\times$	0.5409	0.4763
HFGI_{LSAP}	0.0210	0.0945	**0.7405**	-	-	-	-	$\times$	**0.5766**	**0.5704**
SAM_{e4e}	0.0143	0.1104	0.5568	-	-	-	-	-	-	-
SAM_{LSAP}	**0.0117**	**0.0939**	0.6184	-	-	-	-	-	-	-

are denoted as $\{s_i^N\}_k$. Then, we define an alignment loss as follows:

$$L_{SNCD}(x) = 1 - \frac{1}{k}\sum_{i}^{k}(s_i^N) \cdot (F(I(x))^T \tag{10}$$

$$= 1 - \mu_{s^N} \cdot (F(I(x))^T \tag{11}$$

where μ_{s^N} represents the mean of $\{s_i^N\}_k$. L_{SNCD} is calculated by given images x (i.e., a batch of images in encoder training or one image in optimization) in each iteration. Moreover, we present the details of our encoder and optimization methods separately.

Encoder. The pipeline of encoder-based alignment inversion method is shown in Figure 2a. Given real images, encoder is optimized by minimizing multiple losses at image level and latent code level. Following Richardson et al. (2021) and Tov et al. (2021), L_{img} consists of distortion loss, perception loss, and identity loss. Besides, delta-regulation loss is also applied to inverse codes. The whole training object is defined by:

$$\mathcal{L} = \mathcal{L}_2 + \lambda_1\mathcal{L}_{lpips} + \lambda_2\mathcal{L}_{sim} + \lambda_3\mathcal{L}_{d-reg} + \lambda\mathcal{L}_{SNCD} \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda$ are hyper-parameters to adjust the weight of each component in loss function. In encoder-based method, L_{SNCD} aims to align the encoder's output space with synthetic latent space.

Optimization. Optimization-based inversion method updates latent code iteratively. Compared to encoder-based approach, L_{SNCD} is used to minimize the distance between a certain latent code with synthetic latent space. Following Karras et al. (2020), we apply two losses in image level and latent code level, respectively:

$$\mathcal{L} = \mathcal{L}_{lpips} + \lambda\mathcal{L}_{SNCD} \tag{13}$$

The encoder-based and optimization-based method are denoted as LSAP_E and LSAP_O respectively.

5 EXPERIMENTS

In this section, we conduct extensive experiments to demonstrate the effects of our LSAP in the face domain. Results on other domains can be found in Appendix F.1. We evaluate inversion methods from three perspectives. We use MSE, LPIPS, and identity similarity to evaluate fidelity and use SNCD and LEC to evaluate perception and editability. Furthermore, we calculate the identity similarity between origin images and edited images under the same editing effects by each inversion method. For image perception, we discuss in Appendix E to illustrate the discrepancy.

Implementation Details. See Appendix C.

Quantitative Results. We provide the inversion results in Table 1. For encoder-based methods, LSAP_E's reconstruction performance is slightly decreased from pSp(Richardson et al., 2021) and obtains a significant improvement on perception and editability metrics. Compared to e4e, LSAP_E

Figure 3: **Inversion and editing results of different inversion methods.** We show the comparison of encoder-based, optimization-based and hybrid method respectively. LSAP$_E$ improves the perception and editability while retaining the fidelity, and HFGI$_{LSAP}$ and SAM$_{LSAP}$ further reduce image distortion.

is superior in all perspectives. LSAP$_E$ achieves the best LEC in three editing attributes and identity preservation in two attributes. It is worth mentioning that although e4e has decent editability, it gets worse similarity between origin images and edited images than pSp in "pose" and "smile", which is caused by their reconstruction gap. Nevertheless, LSAP$_E$ reaches a higher similarity in "pose" and "age", which indicates that it can preserve portrait identity well during manipulation.

For optimization-based methods, LSAP$_O$ obtains comparable reconstruction performance in both $\mathcal{W}$ and $\mathcal{W}^+$ spaces. $\mathcal{W}$ space inversions both get weak results. Directly optimizing latent codes in $\mathcal{W}^+$ achieves the best fidelity; however, which gets the worst SNCD and cannot be used to edit. The latent codes may locate in a low-density place of synthetic distribution under only minimizing image distortion per image. LSAP$_O$-$\mathcal{W}^+$ employs additional alignment supervision and makes inverse codes editable. Meanwhile, thanks to its terrific reconstruction ability, the edited images preserve identity best.

To demonstrate our approach's generality and potential, we evaluate the performance with *Result Refinement* methods (i.e., HFGI(Wang et al., 2022) and SAM(Parmar et al., 2022)) employed with e4e and LSAP$_E$. Since HFGI and SAM aim to refine the generated and edited results and do not produce new latent codes, SNCD and LEC are unavailable to measure. Therefore, we focus on reconstruction ability and identity preservation in editing. As shown in the last four rows of Table 1, our method achieves the best performance in every perspective. It attains lower image distortion even than $\mathcal{W}^+$ space optimization. HFGI with e4e and LSAP$_E$ reach the same MSE results while LSAP$_E$ gains better LPIPS and identity similarity. With SAM, LSAP$_E$ achieves the best MSE and LPIPS in all experiments. For image editing, HFGI with e4e and LSAP$_E$ both reach higher identity preservation, while LSAP$_E$ has a 0.03 improvement in "smile" and a 0.10 improvement in "age".

Qualitative Results. We perform the qualitative comparison of our methods and previous inversion methods in Figure 3. Our alignment paradigm attains comparable reconstruction quality with pSp and StyleGAN2 optimization (both in $\mathcal{W}$ and $\mathcal{W}^+$ spaces) in encoder-based and optimization-based methods. Meanwhile, LSAP improves image perception and editability a lot. In comparison with e4e, which is the most popular encoder in recent research, LSAP$_E$ achieves better fidelity and editability. For example, editing results of man in the first image of Figure 3 from e4e have redundant glasses (smile) and change eyes gaze (pose). In optimization-based methods, $\mathcal{W}$ space inversion can not get acceptable results. LSAP$_O$ makes $\mathcal{W}^+$ codes editable and preserves identity much during editing. It demonstrates that our approach provides a concise and straightforward solution even in optimization-based methods, while previous methods(Roich et al., 2021) inverse images into $\mathcal{W}$ to attain editability. In hybrid methods, HFGI and SAM improve the reconstruction ability from e4e and LSAP$_E$. Inversion and editing results from those methods are similar to the corresponding results from encoders while retaining more image details. SAM$_{LSAP}$ achieves state-of-the-art for its high fidelity perception, and editability performance. We show more qualitative results in Appendix F.2.

***Image Embedding* and *Result Refinement*.** In § 3.1, we point out that fidelity is improved in the *Result Refinement* step, while perception and editability are inherited from *Image Embedding*. As can be seen in Figure 3, if editing results from encoder are weak, such as attributes entanglement, the corresponding results from hybrid methods are basically similar. For example, editing the third image with "smile" and "age", results from e4e appear with additional glasses, then the results from HFGI$_{e4e}$ and SAM$_{e4e}$ also have the same impacts. Hence, although *Result Refinement* can largely improve the fidelity, the *Image Embedding* still plays an important role in inversion.

SNCD and Perception/editability. In § 3 we find these two characteristics are related to alignment between inverse codes and synthetic distribution. Then we introduce that SNCD can numerically reflect them, which is validated in our experiments. Those with smaller SNCD have better image quality and editing results (e.g., e4e and LSAP$_E$), while the large SNCD means weak generating and manipulation results (e.g., pSp and StyleGAN2-$\mathcal{W}^+$).

6 CONCLUSION

Fidelity, perception and editability are three critical characteristics of inversion methods. While dividing inversion methods into two steps *Image Embedding* and *Result Refinement*, although the second step improves fidelity a lot, it only inherits the perception and editability from the first step. Hence, designing a image embedding mechanism with excellent perception and ediatbility and retaining fidelity is a critical problem of gaining a better inversion method. To this end, we start by tracing the source of perception and editability in inversion process and find that it is significant to embed images aligning with synthetic distribution in *Image Embedding* step. Hence, we propose a Latent Space Alignment Inversion Paradigm (LSAP), containing the measurement and solution of latent space disalignment. Specifically, to illustrate the disalignment straightforwardly and numerically, we propose $\mathcal{S}^{\mathcal{N}}$ Cosine Similarity (SNCD) as metric with Normalized Style Space ($\mathcal{S}^{\mathcal{N}}$) latent space. Thanks to the differentiable characteristic of SNCD, we conduct a uniform solution in encoder-based and optimization-based approaches. Through extensive experiments, LSAP shows promising results in all three features, and hybrid methods with LSAP achieve state-of-the-art.

REFERENCES

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*, pp. 4432–4441, 2019.

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*, pp. 8296–8305, 2020.

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6711–6720, 2021.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18511–18521, 2022.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018.

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*, pp. 4690–4699, 2019.

Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. 2019.

Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11389–11398, 2022.

Qianli Feng, Viraj Shah, Raghudeep Gadde, Pietro Perona, and Aleix Martinez. Near perfect gan inversion. *arXiv preprint arXiv:2202.11833*, 2022.

Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. Insetgan for full-body image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7723–7732, 2022.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *arXiv preprint arXiv:2204.11823*, 2022.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5744–5753, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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. *Advances in neural information processing systems*, 30, 2017.

Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

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*, pp. 4401–4410, 2019.

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*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

Taewoo Kim, Chaeyeon Chung, Yoonseo Kim, Sunghyun Park, Kangyeol Kim, and Jaegul Choo. Style your hair: Latent optimization for pose-invariant hairstyle transfer via local-style-aware hair alignment. *arXiv preprint arXiv:2208.07765*, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Zuoxin Li, Fuqiang Zhou, Lu Yang, Xiaojie Li, and Juan Li. Accelerate neural style transfer with super-resolution. *Multimedia Tools and Applications*, 79(7):4347–4364, 2020.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Gaurav Parmar, Yijun Li, Jingwan Lu, Richard Zhang, Jun-Yan Zhu, and Krishna Kumar Singh. Spatially-adaptive multilayer selection for gan inversion and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11399–11409, 2022.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. *arXiv preprint arXiv:2001.10238*, 2020.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 484–492, 2020.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2011.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2287–2296, 2021.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021.

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

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Nurit Spingarn-Eliezer, Ron Banner, and Tomer Michaeli. Gan" steerability" without optimization. *arXiv preprint arXiv:2012.05328*, 2020.

Jiaze Sun, Binod Bhattarai, Zhixiang Chen, and Tae-Kyun Kim. Secgan: Parallel conditional generative adversarial networks for face editing via semantic consistency. *arXiv preprint arXiv:2111.09298*, 2021.

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.

Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pp. 9786–9796. PMLR, 2020.

Binxu Wang and Carlos R Ponce. The geometry of deep generative image models and its applications. *arXiv preprint arXiv:2101.06006*, 2021.

Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11379–11388, 2022.

Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. E2style: Improve the efficiency and effectiveness of stylegan inversion. *IEEE Transactions on Image Processing*, 31:3267–3280, 2022.

Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 364–373, 2019.

Lu Yang, Qing Song, Zhihui Wang, Mengjie Hu, Chun Liu, Xueshi Xin, Wenhe Jia, and Songcen Xu. Renovating parsing r-cnn for accurate multiple human parsing. In *European Conference on Computer Vision*, pp. 421–437. Springer, 2020.

Lu Yang, Qing Song, Zhihui Wang, Zhiwei Liu, Songcen Xu, and Zhihao Li. Quality-aware network for human parsing. *arXiv preprint arXiv:2103.05997*, 2021a.

Lu Yang, Qing Song, and Yingqi Wu. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. *Multimedia tools and applications*, 80(1):855–875, 2021b.

Lu Yang, Zhiwei Liu, Tianfei Zhou, and Qing Song. Part decomposition and refinement network for human parsing. *IEEE/CAA Journal of Automatica Sinica*, 9(6):1111–1114, 2022a.

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7693–7702, 2022b.

Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. *arXiv preprint arXiv:2203.04036*, 2022.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. *Advances in neural information processing systems*, 32, 2019.

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 conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10145–10155, 2021.

A PROPERTY PROOF

In this section, we provide detailed proof of Property 3.1 and Property 3.3.

A.1 PROOF OF PROPERTY 3.1

Property 3.1 Suppose that $s = \{s_1, s_2, \ldots, s_k\}$ is a set of style space latent codes and corresponded to image $x = G_{\mathcal{S}}(s)$. For $\forall a \in \mathbb{R}$ and $\forall l \in [1, k]$, $\hat{s}$ follows:

$$\hat{s_n} = \begin{cases} s_n, & if n \neq l \\ a \times s_n, & if n = l \end{cases}.$$

We have $x = G_{\mathcal{S}}(s) = G_{\mathcal{S}}(\hat{s})$.

Proof. According to StyleGAN2, style latent codes are used in weight demodulation way. For lth layer, convolution layer weights $W_{i,j,k}$ are scaled by lth style latent codes s_l firstly:

$$W'_{ijk} = s_l^i \times W_{ijk}, \tag{14}$$

where i, j, k enumerate input feature maps, output feature maps and spatial footprint, respectively.

To integrate instance normalization in convolution layer, StyleGAN2 demodulates each ouput feature map by $\sigma_j = \sqrt{\sum_{i,k} {W'_{ijk}}^2}$, assuming that input activations are i.i.d. random variables with unit standard deviation (ignore ϵ used for numerical stable):

$$W''_{ijk} = \frac{W'_{ijk}}{\sqrt{\sum_{i,k} {W'_{ijk}}^2}} \tag{15}$$

Substitue formula 14 into formula 15, then we can reach:

$$W''_{ijk} = \frac{s_l^i \times W_{ijk}}{\sqrt{\sum_{i,k}(s_l^i \times W_{ijk}^2)}} \tag{16}$$

Suppose that $\hat{s_l} = a \times s_l$,

$$\begin{aligned} \hat{W}''_{ijk} &= \frac{\hat{s}_l^i \times W_{ijk}}{\sqrt{\sum_{i,k}(\hat{s}_l \times W_{ijk}^2)}} \\ &= \frac{a \times s_l^i \times W_{ijk}}{\sqrt{\sum_{i,k}(a \times s_l^i \times W_{ijk}^2)}} \\ &= \frac{s_l^i \times W_{ijk}}{\sqrt{\sum_{i,k}(s_l^i \times W_{ijk}^2)}} = W''_{ijk} \end{aligned} \tag{17}$$

Thus, if scale s by $a \in \mathbb{R}$ in an arbitrary layer, convolution weights are identical, meaning generated images are the same. □

A.2 PROOF OF PROPERTY 3.3

Property 3.3 For lth layer ($\forall l \in [1, k]$), define $F_l : \mathcal{Z}/\mathcal{W} \to \mathcal{S}$ as the mapping function between $\mathcal{S}$ and $\mathcal{Z}/\mathcal{W}$ space. For $\forall p_l \in \mathcal{Z}/\mathcal{W}$ ($F_l(p) \neq \mathbf{0}$) and $a \in \mathbb{R}$, $\exists p'_l \neq p_l$ such that the corresponding S space latent codes satisfy: $s'_l = a \times s_l$, where $s_l = F_l(p_l)$ and $s'_l = F_l(p'_l)$.

Proof. We prove this property separately under $\mathcal{W}$ and $\mathcal{Z}$ spaces. Since cases under each layer level are the same without loss of generality, to express concisely, we consider the situation under an arbitrary layer and ignore l in the later formulation.

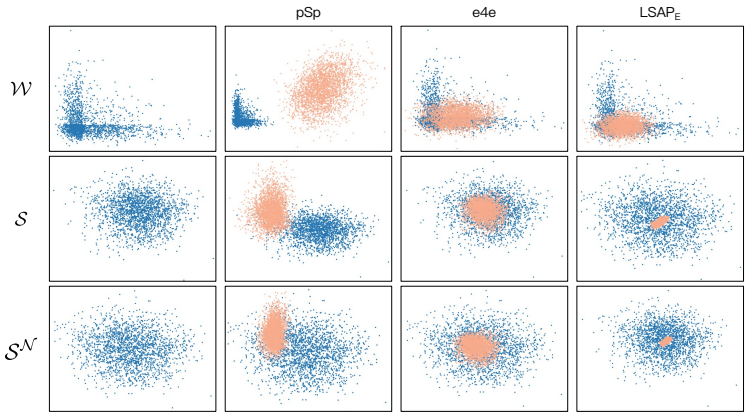


Figure 4: **Illustrate the Latent Space Distribution.** We show the latent code distribution of random sampling and inversion from real images in multiple latent spaces (i.e., $\mathcal{W}/\mathcal{S}/\mathcal{S}^{\mathcal{N}}$). For all spaces, we choose the first two dimensions of codes for visualization. Our method and e4e clearly bring the distributions closer.

$\mathcal{W}$ space The mapping function between $\mathcal{W}$ and $\mathcal{S}$ space is established by linear projection in generator, as follows:

$$s = F(w) = Aw + b \tag{18}$$

If $\exists y$, such that

$$Ay = (a-1)b \tag{19}$$

and let

$$w' = aw + y \tag{20}$$

we have

$$s' = Aw' + b = A(aw + y) + b = aAw + ab = as \tag{21}$$

In StyleGAN, $A \in \mathbb{R}^{m \times n}(m \leq n)$ may not be square matrix in some resolution levels and the rank of A is unstable. It indicates Equation 19 can not be solved by inverse of A directly. We can obtain y by solving the least squares problem:

$$\min_{y} \|Ay - (a-1)b\| \tag{22}$$

Hence, for $\forall w$, when $w' = aw + y$, $F(w) = F(w')$. In addition, we can prove $w' \neq w$ by the counterfactual method. If $w' = w$, we have $y = (1-a)w$ and $A(1-a)w = (a-1)b$, so $Aw = -b$ and $s = \mathbf{0}$. Due to $s \neq \mathbf{0}$, $w' \neq w$ and $w' = aw + y$, $F(w) = F(w')$, we prove that property holds in $\mathcal{W}$ space.

$\mathcal{Z}$ space Although we have proved in $\mathcal{W}$ space, the mapping function between $\mathcal{Z}$ and $\mathcal{W}$ or $\mathcal{Z}$ and $\mathcal{S}$ is represented by a multilayer perception, which is difficult to prove directly by formula. Fortunately, as the objective function is defined, we can obtain z' by optimization, satisfying $s = F(z) = a \times F_(z') = ks'$ and $z' \neq z$. □

B Latent Space Distribution

We illustrate the discrepancy between inverse distribution and synthetic distribution in each space from pSp, e4e, and LSAP$_E$ as shown in Figure 4. Blue points are first sampled in standard normal distribution and then fed into generator to convert into other spaces. We also scatter the inverse latent codes from CelebA-HQ test dataset.

In pSp, for only image level losses are used to get better fidelity, inverse distributions clearly disalign with synthetic distribution, especially in $\mathcal{W}$ space. As we can see, these two distributions do

not overlap almost at all. However, in $\mathcal{S}$ and $\mathcal{S}^{\mathcal{N}}$ spaces, this phenomenon is alleviated, although disalignment still occurs. This also validates our previous analysis of these spaces that $\mathcal{S}^{\mathcal{N}}$ is more suitable to measure distribution disalignment. LSAP$_E$ and e4e both employ latent code alignment techniques. A $\mathcal{W}$ space latent code discriminator is introduced in e4e, and we propose the $\mathcal{S}^{\mathcal{N}}$ space alignment loss to minimize the discrepancy between two distributions directly. Hence, disalignments are significantly decreased in these two methods, as can be seen in the last two columns. In $\mathcal{S}$ and $\mathcal{S}^{\mathcal{N}}$ spaces, latent codes inversed by e4e and LSAP$_E$ are gathered in the middle of the synthetic distribution while LSAP$_E$'s output space is much closer.

Table 2: **Latent Codes Distance in Synthetic Images Inversion**. We inverse randomly synthetic images to obtain pairwise latent codes in each latent space and calculate the average distance.

	$\mathcal{W}$	$\mathcal{S}$	$\mathcal{S}^{\mathcal{N}}$
pSp	63.50	859.02	0.10
e4e	**11.63**	**299.53**	**0.03**
LSAP$_E$	11.87	312.20	**0.03**

Furthermore, we conduct a quantitative experiment to study the disalignment in inversion methods. First, we randomly generate 5,000 images by StyleGAN2 and convert their latent codes into $\mathcal{W}/\mathcal{S}/\mathcal{S}^{\mathcal{N}}$ spaces. Then we inverse these images by each encoder and measure the average distance of latent codes from synthetic and inverse sources. For $\mathcal{W}$ and $\mathcal{S}$ spaces, we use MSE to measure distance and use SNCD in $\mathcal{S}^{\mathcal{N}}$ space. Since each value in the table represents the distance in latent space, the lower value indicates inverse codes are close to synthetic codes. Without any supervision, codes inversed by pSp are far away from their origin latent codes in all latent spaces. LSAP$_E$ and e4e have similar performance, while e4e has slight improvements in $\mathcal{W}$ and $\mathcal{S}$ spaces. It is reasonable that e4e is trained with $\mathcal{W}$ space supervision, and LSAP$_E$ is supervised in $\mathcal{S}^{\mathcal{N}}$ space.

C Implementation Details

Datasets. We conduct the whole experiment on four domains: face, cars, church, and wild animal, corresponding to human, object, scene, and animal, respectively. In all domains, we use the official StyleGAN2 generator. For face domain, We train the LSAP$_E$ on FFHQ(Karras et al., 2019) (70,000 images) and evaluate on CelebA-HQ(Liu et al., 2015; Karras et al., 2017) test dataset (2824 images). Editing directions are gained by Shen et al. (2020). For car domain, we use Stanford Cars(Krause et al., 2013) dataset with 8,144 images for training and randomly selected 1000 images for evaluation and edit images by Härkönen et al. (2020). For church domain, we use LSUN(Yu et al., 2015) Church dataset with 126,227 training images and 300 test images. For wild animal domain, we use AFHQ(Choi et al., 2020) Wild dataset.

LSAP$_E$. The input image resolution is 192×256 in car domain and 256×256 for the others. For data augmentation, we only employ random horizontal flips. We adopt the Ranger optimizer, combining the Rectified Adam(Liu et al., 2019) and the Lookahead technique(Zhang et al., 2019), with 0.001 learning rate. We take all experiments on a single GPU with batch size of 8. Besides, we follow the progressive training from e4e. In LSAP$_E$, perceptual loss weight λ_1 is 0.8, delta-regulation loss λ_3 is $2e-5$, and SNCD loss λ is 0.5 for all domains. Similarity loss weight λ_2 is 0.1 for face domain over pre-trained ArcFace(Deng et al., 2019) and 0.5 for others with MOCOv2(Chen et al., 2020) and ResNet-50(He et al., 2016).

Optimization-based Method. Following Karras et al. (2020), we adopt Adam(Kingma & Ba, 2014) optimizer to minimize perceptual loss and SNCD loss with noise regularization. λ is set to 20 for $\mathcal{W}^+$ space and 5 for $\mathcal{W}$ space.

Hybrid Method. We apply e4e and LSAP$_E$ to two hybrid methods, HFGIWang et al. (2022) and SAM(Parmar et al., 2022), to illustrate the effects of *Image Embedding* step. For HFGI, we use official weight to evaluate HFGI$_{e4e}$ and follow its training script to train HFGI$_{LSAP}$. In practice, we only replace the encoder weight from e4e to LSAP$_E$. Since SAM only releases the optimization

Figure 5: **Ablation study of image perception.** We show the inversion result from LSAP$_E$ and the same encoder without $\mathcal{L}_{SNCD}$ to illustrate the effect. LSAP$_E$ significantly improves the image quality and solves the unnatural generation.

codes, we first embed images into latent codes by encoder, and then optimize the latent codes with intermediate feature for 500 iterations, with threshold $\tau = 0.225$.

Evaluation. Since inversion and editing results are gained by multiple codebases, we conduct all image level evaluations on saved image files. MSE, LPIPS and ID similarity are calculated on 256×256 resolution by script from pSp(Richardson et al., 2021). To get SNCD, we choose $k = 50,000$ to random sample latent codes from generator and convert them into $\mathcal{S}^{\mathcal{N}}$ space. For LEC and identity similarity, we use different editing factor to ensure the same editing effect for all inversion methods, which can be found in quality results.

D Ablation Study

Table 3: **Ablation study on hyper-parameter of LSAP$_E$.** We set $\lambda = 0.5$ in our experiments by default.

	Fidelity			Perception & editability			
λ	MSE ↓	LPIPS ↓	Similarity ↑	SNCD ↓	LEC$_{pose}$ ↓	LEC$_{smile}$ ↓	LEC$_{age}$ ↓
0	0.0369	0.1657	0.5512	0.0736	24.8245	22.5007	24.8069
0.1	0.0382	0.1703	0.5438	0.0416	19.1594	14.0133	15.2246
0.25	0.0391	0.1737	0.5410	0.0395	19.1345	14.1382	15.1599
0.5	0.0397	0.1766	0.5305	0.0385	19.0211	14.0360	14.6715
0.75	0.0406	0.1792	0.5222	0.0381	19.0949	14.0128	14.3198
1.0	0.0413	0.1809	0.5168	0.0378	15.8013	13.8433	14.6084

We study the hyper-parameter λ of $\mathcal{L}_{SNCD}$ on face domain with LSAP$_E$ as exmaple, and the quantitative results are shown in Table 3. A higher value of λ makes image distortion increase. This result is in line with our expectations since λ controls the contributions of alignment loss. Conversely, perception and editability are improved as λ increased. We visualize the inversion results in Figure 5 with $\lambda = 0$ and 0.5. In the first row, quality of teeth, eyes and lip's texture is weak. For example, in the left image in first row, the end of the left eyelid (right in the figure) is located too far from the

Figure 6: **Ablation study of image editability.** We show the manipulation result from LSAP$_E$ with different hyper-parameter λ.

left eye. Besides, teeth is misaligned with adhesions and lips are too smooth without normal texture. These problems are solved by LSAP, as we can see in the second row. To demonstrate change in editability, we further compare the manipulation results with $\lambda = 0, 0.5, 1.0$, which is shown in Figure 6. The first two images are edited with "smile" while the third is edited with "pose". When $\lambda = 0$, the edited images are unphotorealistic, and glasses occur with editing "smile". Results of $\lambda = 0.5$ and 1.0 show the similar results with excellent editability. The inversion and editing results show the superiority of our alignment paradigm.

E IMAGE PERCEPTION

We illustrate the discrepancy of image perception from each inversion method by high-resolution inverse images. As can be seen in Figure 7, the inverse results from each approach are marginally different in high resolution, especially in hair, teeth, lip, and skin area. This is not obvious in low resolution or thumbnails, as can be seen the first row. However, it makes image unnatural and fake in high resolution. We recommend comparing the visual quality at a higher resolution (e.g., 1024×1024). Our alignment paradigm improves the image quality well, as can be seen in Figure 7, our results have natural visual details.

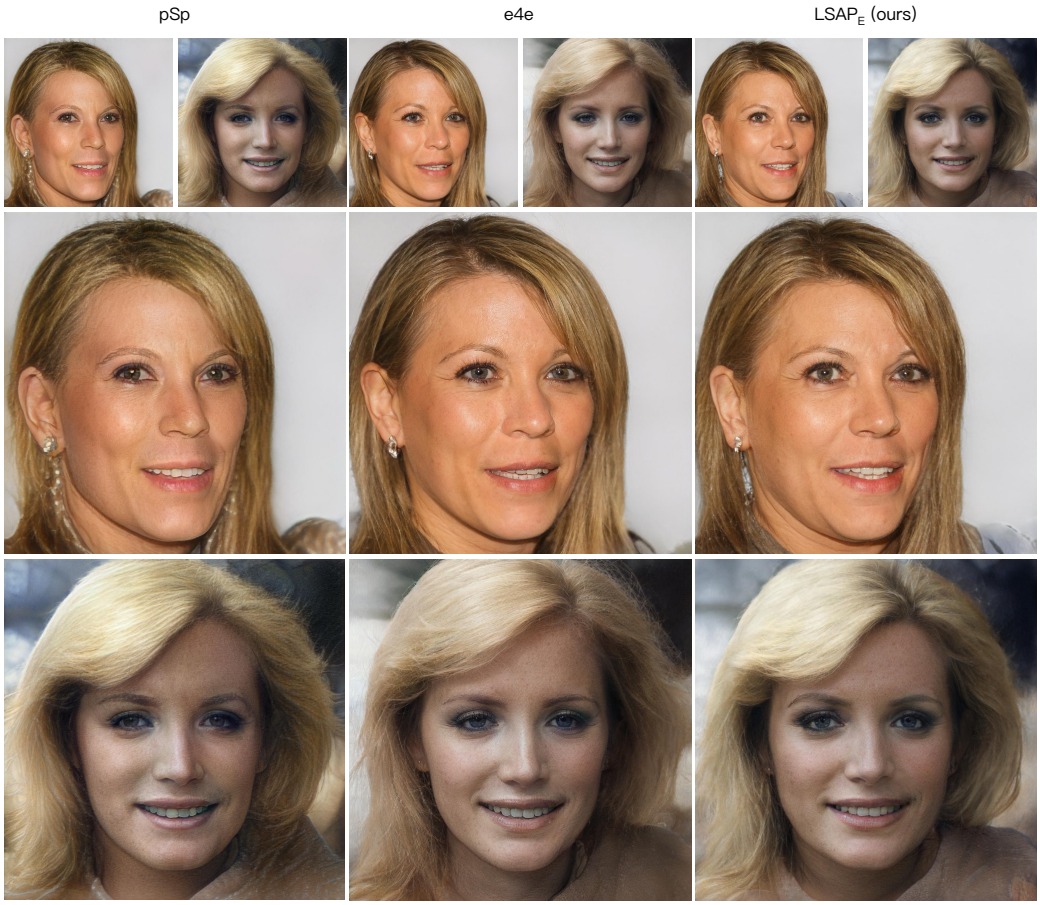


Figure 7: **Illustrate image perception in high resolution results**. We show the inversion results from pSp, e4e, and LSAP$_E$ in high resolution to demonstrate the details of images. We also provide the low-resolution results to compare in the first row. The difference of perception is not obvious in low-resolution images.

Domain	Car			Church			Wild Animal		
Metric	MSE↓	LPIPS↓	SNCD↓	MSE↓	LPIPS↓	SNCD↓	MSE↓	LPIPS↓	SNCD↓
e4e	0.1201	0.3252	0.0646	0.1505	0.4307	0.0761	0.0882†	0.2658†	0.0379†
LSAP$_E$ (ours)	**0.1049**	**0.3106**	**0.0492**	**0.1144**	**0.3426**	**0.0588**	**0.0785**	**0.2524**	**0.0224**
SAM$_{e4e}$	0.0289	0.1506	-	-	-	-	-	-	-
SAM$_{LSAP}$ (ours)	**0.0247**	**0.1361**	-	-	-	-	-	-	-

Table 4: **Quantity results on other domains.** † means the model is unavailable and we train the encoder by official code.

F Additional Results

F.1 Results on Other Domains

We conduct experiments on cars, churches, and wild animals to illustrate the universality. The quantity results are shown in Table 4. Since editing results are unstable and identity cannot be measured in these three domains, LEC and identity preservation are not applied. We can measure perception and editability by SNCD and visualization. Compared to e4e, LSAP$_E$ achieve better performance in all fields. We also employ SAM with these two encoders and evaluate the image

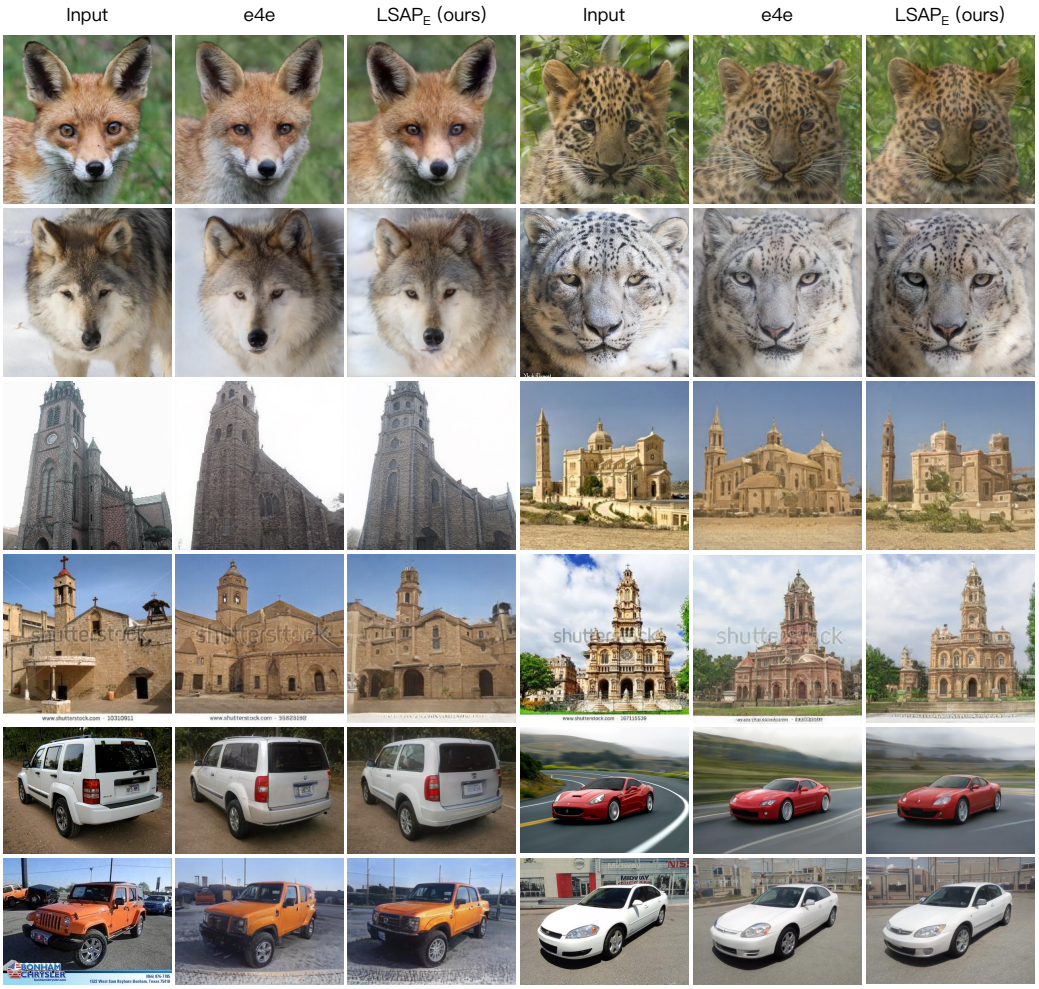


Figure 8: **Inversion results on other domains.** In car and church domains, the official e4e models are available and we train the encoder on AFHQ(Choi et al., 2020) Wild dataset.

distortion (i.e., MSE and LPIPS). The qualitative results can be found in Figure 8. Our results have much natural and high-quality performance.

We further edit images in car domain and compare the difference between two encoders: e4e and LSAP$_E$, which we show in Figure 9. For inversion, LSAP$_E$ achieves slight improvement in fidelity, since the color and reflection are reconstructed accurately. For example, in the second image, the reflection is represented in LSAP$_E$ result, while the result from e4e only shows the white color. During editing, LSAP$_E$ demonstrates the excellent ability to generate good editing results. With SAM technique, LSAP$_E$ also achieves a better result in both inversion and manipulation.

F.2 Additional Results on Face Domain

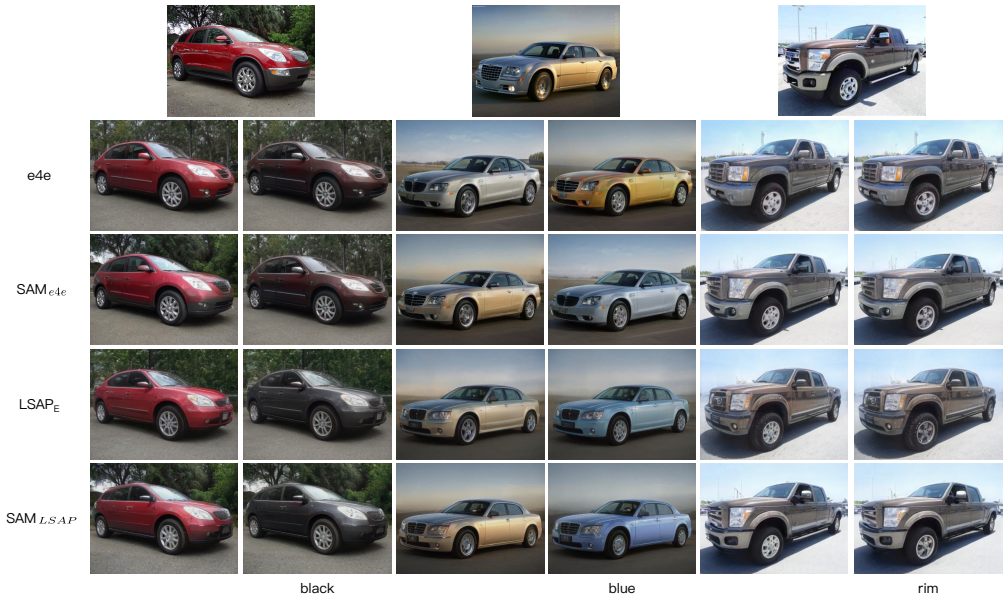


Figure 9: **Inversion and editing comparison between e4e and LSAP$_E$.** We illustrate these two encoders' inversion and editing results and the corresponding results with SAM. LSAP significantly improves editability and retains more visual details during inversion.

Figure 10: We show the additional inversion and editing results on face domain.

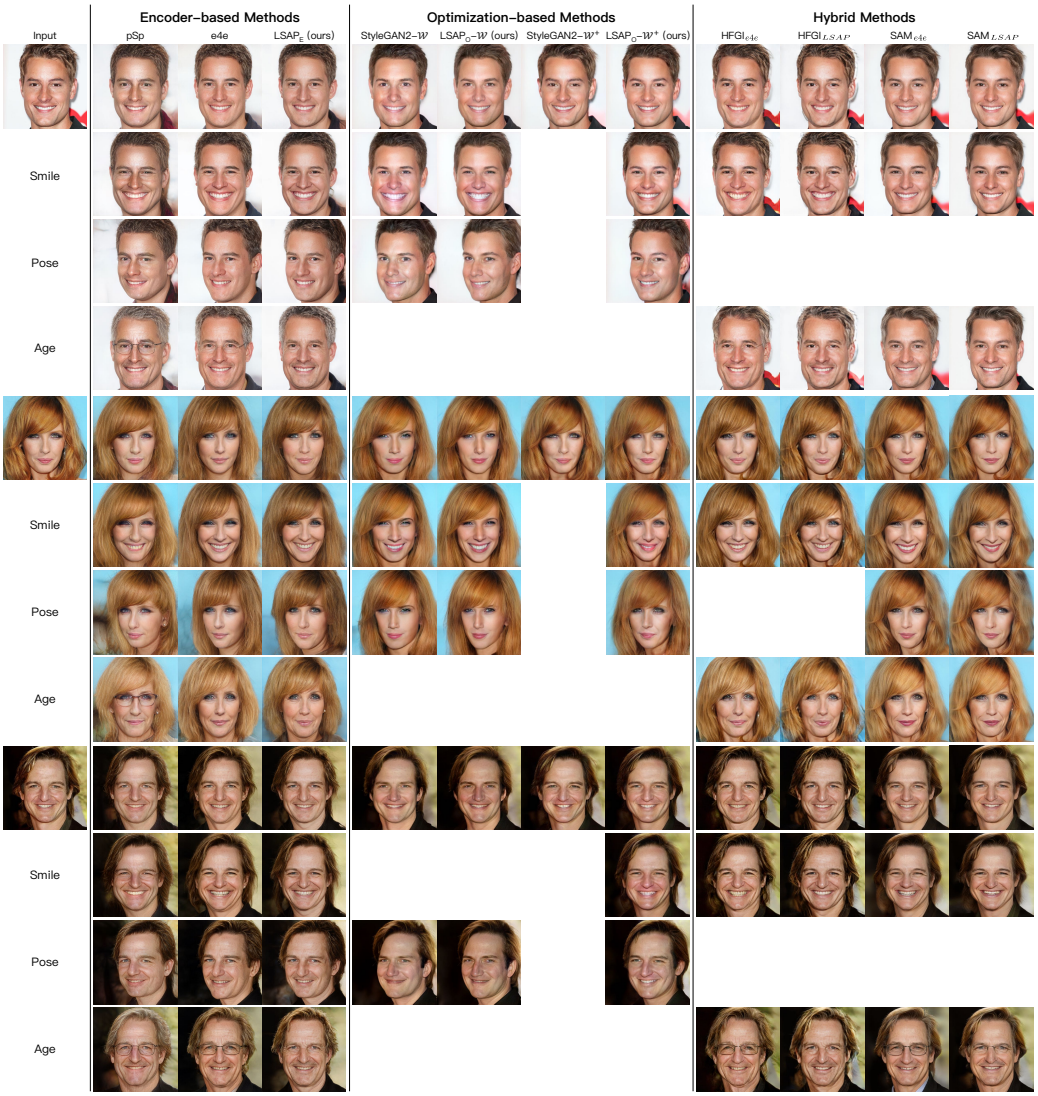


Figure 11: We show the additional inversion and editing results on face domain.