Revisiting Latent Space of GAN Inversion for Robust Real Image Editing

Kai Katsumata[1] Duc Minh Vo[1] Bei Liu[2] Hideki Nakayama[1]
[1]The University of Tokyo, [2]Microsoft Research
{katsumata,vmduc,nakayama}@nlab.ci.i.u-tokyo.ac.jp, bei.liu@microsoft.com

Abstract

We present a generative adversarial network (GAN) inversion with high reconstruction and editing quality. GAN inversion algorithms with expressive latent spaces produce near-perfect inversion but are not robust to editing operations in a latent space, leading to undesirable edited images, a phenomenon known as the trade-off between reconstruction and editing quality. To cope with the trade-off, we revisit the hyperspherical prior of StyleGANs $\mathcal{Z}$ and propose to combine an extended space of $\mathcal{Z}$ with highly capable inversion algorithms. Our approach maintains the reconstruction quality of seminal GAN inversion methods while improving their editing quality owing to the constrained nature of $\mathcal{Z}$. Through comprehensive experiments with several GAN inversion algorithms, we demonstrate that our approach enhances the image editing quality in 2D/3D GANs.[1]

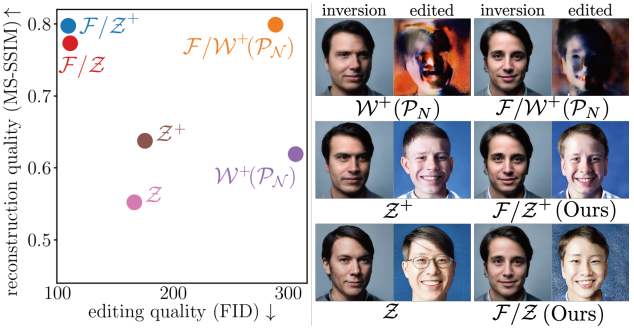


Figure 1. (left) Quantitative analysis of reconstruction and editing quality. We report MS-SSIM between target and inverted images and FID between target images and edited results obtained by GANSpace with the editing intensity of 20. The scores are calculated on the CelebA-HQ test set. $\mathcal{F}/\mathcal{W}(\mathcal{P}_\mathcal{N})$ [14], $\mathcal{Z}$, and $\mathcal{Z}^+$ show the trade-off between reconstruction quality and editing quality. $\mathcal{F}/\mathcal{Z}^+$ improves the editing quality of $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$ without losing reconstruction quality. (right) Examples of inversion and edited results for compared latent spaces. Our $\mathcal{F}/\mathcal{Z}^+$ achieves GAN inversion with high reconstruction and editing quality by overcoming the flaw of existing inversion methods.

1. Introduction

The combination of generative adversarial network (GAN) inversion [2, 3, 5, 11, 14, 29, 31, 47, 54, 55] and latent space editing [12, 34, 36] enables us to edit a wide range of real image attributes, such as aging, expression, and light condition, by applying editing operations in the latent space of GANs [12, 27, 34, 36] to the inverted latent codes obtained by GAN inversion methods. To this end, many methods [2, 3, 11, 14] aimed at finding the latent code of StyleGANs [17–20] that generates a given image have been developed. The majority of GAN inversion studies [11, 14, 42, 44] have focused on reducing reconstruction loss by exploring novel embedding spaces, encoding methods, and optimization algorithms.

Nevertheless, it is still challenging to achieve a good trade-off between reconstruction and editing quality in GAN inversion [56]. The reconstruction quality indicates the degree of similarity between the input and reconstructed images. On the other hand, the editing quality indicates how realistic and plausible the edited image is after performing latent space editing operations. As shown in Fig. 1, the original StyleGAN prior $\mathcal{Z}$ has the hyperspherical constraint, resulting in insufficient reconstruction quality yet high editing quality. In contrast, popular latent spaces such as $\mathcal{W}$ [20], $\mathcal{W}^+$ [2], and $\mathcal{S}$ [44] are derived from $\mathcal{Z}$ by utilizing a mapping network. This approach may enhance the quality of reconstructed images, but it occasionally fails to produce the desired edited images owing to the unconstrained nature of these spaces.

Recent attempts such as SAM [29], PTI [31], and $\mathcal{P}$ [56], which are built upon $\mathcal{W}$ or $\mathcal{W}^+$, aim at maintaining the perceptual quality of edited images during the performance of semantic editing operations. Since they still rely on unconstrained latent spaces, it is impossible to avoid undesirable edited images. To address the trade-offs, our main target is to enhance the reconstruction quality by using a more expressive space or a generator tuning technique while simultaneously preserving the editing quality through the use of

[1]The code is available at: https://github.com/raven38/hypershpere-gan-inversion

the constrained latent space.

To begin with, we revisit the original latent space $\mathcal{Z}$, characterized by its high editing quality. Because employing $\mathcal{Z}$ alone lacks reconstruction quality, we propose to enhance the reconstruction capability by expanding $\mathcal{Z}$ to $\mathcal{Z}^+$ in a similar fashion to prior research [39,45]. Subsequently, the integration of $\mathcal{Z}^+$ with advanced inversion methods, such as $\mathcal{F}/\mathcal{W}^+$ [14], is performed to further improve the overall reconstruction quality.

Qualitative and quantitative evaluations show that our proposed method maintains image quality after performing editing operations without sacrificing reconstruction quality (Fig. 1). Our contributions are as follows:

- We revisit the $\mathcal{Z}$ space for GAN inversion, pointing out that a combination of an extended space of $\mathcal{Z}$ (*i.e.*, $\mathcal{Z}^+$) and highly capable inversion can maintain reconstruction quality while guaranteeing high editing quality for real image editing using GAN inversion.

- We extend cutting-edge 2D/3D GAN inversion approaches (*e.g.*, $\mathcal{F}/\mathcal{W}^+$ [14], PTI [31], SAM [29], and HFGI3D [48]) with $\mathcal{Z}^+$, demonstrating competitive reconstruction quality with baselines and improvements in editing quality over the baselines. In addition, our approach elevates the latent editing methods (*e.g.*, Local Basis) that necessitate $\mathcal{Z}$ from the realm of synthetic image editing to that of real image editing.

2. Related work

GAN inversion aims to project real images into low-dimensional latent codes, which can be mainly classified into encoder-based and optimization-based approaches. The former type of approach [6, 30, 40, 42] trains an encoder network that predicts latent codes that reconstruct input images. The latter one [2, 3, 31] directly optimizes the latent code to reconstruct a target input. Hybrid approaches [14, 29] initialize a latent code with the encoder prediction and then fine-tune the latent code using an optimization method.

Pioneering studies of GAN inversion aim at the faithful reconstruction of target images. For example, extending a latent space leads to high reconstruction quality [2, 3, 44]. Recently, Kang *et al*. [14] and Feng *et al*. [11] have improved the reconstruction performance for out-of-range images. Another recent direction is to increase robustness in downstream tasks [56]. Zhu *et al*. [56] aimed to increase editing quality by introducing the regularization that directs latent codes toward a high-density region. In addition to 2D GANs [15, 32, 33], GAN inversion has also been investigated for the recently developed 3D GANs [1, 4, 6, 8, 10, 21, 24, 37, 38, 40, 48, 50]. Although our study has the same goal as in [56], our approach is to use a latent space where we know the bound of the codes, which is not the case in [56].

Semantic image editing [9, 12, 23, 34, 36, 53] is one of the downstream tasks of embedding real images into a latent space. The task modifies a latent code along semantically meaningful directions to generate an intended image. Supervised [34] and unsupervised [12, 36] approaches are investigated to explore semantic directions. GANSpace [12] finds useful directions by computing eigenvectors on the empirical distribution of a latent code. SeFa [36] factorizes the weights of layers that feed on latent codes. The above methods explore global semantic directions, which are shared among the latent codes and are called global methods. Unlike global methods, local methods [9, 53] explore semantic directions with respect to each latent code. We edit image attributes by applying semantic image editing methods to the latent code obtained by our approach.

3. Approach

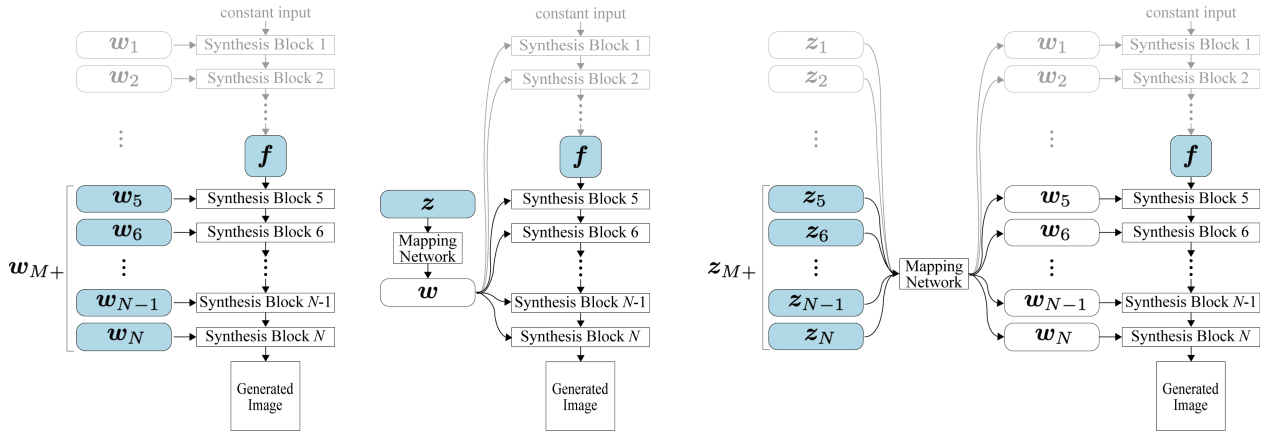
We first review various latent spaces in StyleGANs. Then, we introduce $\mathcal{F}/\mathcal{Z}^+$ that improves editing quality while maintaining reconstruction quality.

3.1. Analysis of StyleGAN spaces

$\mathcal{Z}$ and $\mathcal{Z}^+$ spaces. The generator $G : \mathcal{Z} \rightarrow \mathcal{X}$ learns to map a simple distribution, called the latent space $\mathcal{Z}$, to the image space, where $\boldsymbol{x} \in \mathcal{X}$ is an image, and $\boldsymbol{z} \in \mathcal{Z}$ is uniformly sampled from a hypersphere. The primitive latent code of the StyleGAN family has 512 dimensions.

AgileGAN [39] and StyleAlign [45] employ the extended space $\mathcal{Z}^+$, which provides a different latent code from $\mathcal{Z}$ for each layer. Each element $\boldsymbol{z}^+$ in $\mathcal{Z}^+$ is defined as $\boldsymbol{z}^+ = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N\}$, where $\boldsymbol{z}_i \in \mathcal{Z}$. A code $\boldsymbol{z}_i$ is an input for each layer of a StyleGAN generator and is transformed by the mapping network and AdaIN [13] before being fed into the generator. The number of layers is $N = 18$ for a 1024×1024 StyleGAN. Note that the extended space has not been explored in GAN inversion, as discussed in [47]. Indeed, AgileGAN [39] addresses stylizing portraits without prioritizing faithful reconstruction, and StyleAlign [45] demonstrates that $\mathcal{Z}^+$ does not yield accurate reconstruction for real image inversion. As shown in [39, 45, 47], while the $\mathcal{Z}$ and $\mathcal{Z}^+$ spaces have the issue of insufficient reconstruction quality, the constraint nature of the spaces (*i.e.*, hypersphere) leads to edited images with less deterioration.

$\mathcal{W}$, $\mathcal{W}^+$, and $\mathcal{S}$ spaces. StyleGANs also use the intermediate latent space $\mathcal{W}$ where each $\boldsymbol{w} \in \mathcal{W}$ is produced by a mapping network consisting of eight fully connected layers denoted as $\mathcal{M} : \mathcal{Z} \rightarrow \mathcal{W}$. Owing to multiple affine transformations and nonlinearity functions in $\mathcal{M}$, the features of $\mathcal{W}$ are more disentangled than those of $\mathcal{Z}$.



(a) $\mathcal{F}/\mathcal{W}^+$ space [14] (b) $\mathcal{F}/\mathcal{Z}$ space (Ours) (c) $\mathcal{F}/\mathcal{Z}^+$ space (Ours)

Figure 2. Latent spaces of StyleGANs. The space $\mathcal{F}/\mathcal{W}^+$ introduced by Kang *et al*. [14] consists of $\mathcal{F}$ and $\mathcal{W}^+$ spaces, and it even leads to the faithful reconstruction of out-of-range images. Using $\mathcal{Z}$ or $\mathcal{Z}^+$ instead of $\mathcal{W}^+$, we introduce $\mathcal{F}/\mathcal{Z}$ and $\mathcal{F}/\mathcal{Z}^+$ spaces without sacrificing reconstruction quality with the aid of $\mathcal{F}$. The base code $\boldsymbol{f}$ is an intermediate output of the StyleGAN generator with spatial dimensions. The detail codes $\boldsymbol{w}_{M+}$ and $\boldsymbol{z}_{M+}$ are the subset of $\boldsymbol{w}^+$ and $\boldsymbol{z}^+$, respectively, and are the inputs of the upper stages of the generator. The optimizing codes are highlighted in blue.

Thereafter, in [2, 3], the $\mathcal{W}^+$ space is introduced, achieving a lower reconstruction loss by allowing the control of details of images. Each element $\boldsymbol{w}^+$ in $\mathcal{W}^+$ is defined as $\boldsymbol{w}^+ = \{\boldsymbol{w}_i\}_{i=1}^N$, where $\boldsymbol{w}_i \in \mathcal{W}$. The $\mathcal{S}$ space [44] is spanned by style parameters, which are transformed from $\boldsymbol{w} \in \mathcal{W}$ using different learned affine transformations for each layer of the generator.

Although the $\mathcal{W}$, $\mathcal{W}^+$, and $\mathcal{S}$ spaces derive faithful reconstruction quality, distortions and artifacts may appear in edited images [39, 46, 56]. This is because the embeddings with these spaces for the images may not correspond appropriately to the StyleGAN prior $\mathcal{Z}^+$, and they cannot guarantee that the edited latent code reaches the original spaces. In this study, we aim at latent editing without suffering quality deterioration with the aid of the nature of $\mathcal{Z}$ or $\mathcal{Z}^+$.

$\mathcal{P}_\mathcal{N}^+$ space. Zhu *et al*. [56] introduced space $\mathcal{P}$ and their normalized space $\mathcal{P}_\mathcal{N}$ to explore the GAN inversion trade-offs. The deactivated space of $\mathcal{W}$ is computed by using the inversion of Leaky ReLU. The $\mathcal{P}_\mathcal{N}$ space is presented by whitening $\mathcal{P}$ using the principal component analysis (PCA) parameter computed on one million latent codes in $\mathcal{P}$. Since the distribution of $\mathcal{P}_\mathcal{N}$ can be interpreted as the Gaussian distribution with zero mean and unit variance, we can calculate the density of the latent codes in $\mathcal{P}_\mathcal{N}$. As discussed in [26], the regularization on $\mathcal{P}_\mathcal{N}$ ensures realistic inversion outcomes. For faithful reconstruction quality, $\mathcal{P}_\mathcal{N}$ can be extended to $\mathcal{P}_\mathcal{N}^+$ similar to $\mathcal{W}^+$ [2, 3] or $\mathcal{Z}^+$.

Although $\mathcal{P}_\mathcal{N}^+$ improves the robustness of the reconstructed latent codes, editing operations are performed on the $\mathcal{W}$ or $\mathcal{W}^+$ space. This means that utilizing the regularization on $\mathcal{P}_\mathcal{N}$ leads to maintaining the image quality only within the neighborhood of an inverted code. Hence, the weaknesses of the unconstrained spaces as discussed above remain. We thus seek the latent space that also guarantees generation quality in the editing phase.

$\mathcal{F}/\mathcal{W}^+$ and $\mathcal{F}/\mathcal{S}$ spaces. Kang *et al*. [14] proposed the $\mathcal{F}/\mathcal{W}^+$ space consisting of the $\mathcal{F}$ and $\mathcal{W}^+$ spaces for generalization performance for out-of-range images (Fig. 2a), and the space was also investigated in SAM [29] and Barbershop [55]. The coarse-scale feature map $\boldsymbol{f} \in \mathcal{F}$ is an intermediate output of a generator before taking fine-scale latent codes $\boldsymbol{w}_{M+} = \{\boldsymbol{w}_M, \boldsymbol{w}_{M+1}, \ldots, \boldsymbol{w}_N\}$. An element $\boldsymbol{w}^* = (\boldsymbol{f}, \boldsymbol{w}_{M+})$ of $\mathcal{F}/\mathcal{W}^+$ consists of the base code $\boldsymbol{f}$ and the detail code $\boldsymbol{w}_{M+}$. The information of a noise input and bottom latent codes $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{M-1}\}$ is contained in $\boldsymbol{f}$, and the feature map controls the geometric information. Kang *et al*. [14] have integrated the regularization on $\mathcal{P}_\mathcal{N}$ into $\mathcal{F}/\mathcal{W}^+$ to obtain robust latent codes. The combined space has extended the range of images that can be inverted. The $\mathcal{F}/\mathcal{S}$ space [49] employs the $\mathcal{S}$ space [44] as an alternative to $\mathcal{W}^+$.

While these spaces achieve faithful reconstruction quality, it has the same limitations as $\mathcal{W}$, $\mathcal{W}^+$, $\mathcal{S}$, and $\mathcal{P}_\mathcal{N}^+$. This is because latent editing is performed on unconstrained spaces. Thus, to continue using the $\mathcal{Z}$ space, we leverage the $\mathcal{F}$ space, which complements the lack of representative capacity of the constrained latent space (Figs. 2b and 2c).

3.2. $\mathcal{F}/\mathcal{Z}^+$ space

We introduce the $\mathcal{F}/\mathcal{Z}^+$ space as an example of our approaches. We also introduce our methods based on seminal GAN inversion approaches later. Overall, there is still no existing latent space that can guarantee both reconstruction quality and editing quality. As discussed in [26, 39, 46, 56], leveraging the $\mathcal{Z}$ or $\mathcal{Z}^+$ space leads to high editing quality in exchange for reconstruction quality. To overcome these

limitations, we present an alternative latent space, $\mathcal{F}/\mathcal{Z}^+$, by extending the StyleGAN prior $\mathcal{Z}$. The space consists of the feature space $\mathcal{F}$ for increasing representative capacity and a constrained latent space $\mathcal{Z}^+$ for maintaining editing quality. We use the combination of the spaces $\mathcal{F}$ and $\mathcal{Z}^+$ because we cannot achieve sufficient reconstruction quality when using the space $\mathcal{Z}^+$ to increase the editing quality.

The latent space $\mathcal{F}/\mathcal{Z}^+$ can do semantic editing without image collapse while maintaining the reconstruction quality (Fig. 2c). We define $\mathcal{F}/\mathcal{Z}^+$ by combining the $\mathcal{F}$ and $\mathcal{Z}^+$ spaces. Each element $\boldsymbol{z}^* \in \mathcal{F}/\mathcal{Z}^+$ is defined as $\boldsymbol{z}^* = (\boldsymbol{f}, \boldsymbol{z}_{M+})$, where $\boldsymbol{z}_{M+} = \{\boldsymbol{z}_M, \boldsymbol{z}_{M+1}, \dots, \boldsymbol{z}_N\}$ is a set of latent codes for the fine scales of the generator. We also introduce $\mathcal{F}/\mathcal{Z}$ for comparison, consisting of $\mathcal{F}$ and $\mathcal{Z}$ instead of $\mathcal{Z}^+$.

$\mathcal{F}/\mathcal{Z}^+$ has the desirable properties required for GAN inversion: high reconstruction and editing quality. Our space is characterized by high reconstruction capability attributed to the feature space $\mathcal{F}$ and high editing quality attributed to the primitive space $\mathcal{Z}$. For PULSE, it is discussed in [26] the importance of considering a manifold of a latent space, which controls the content quality. Following this discussion, Zhu *et al*. [56] assume that the deactivated $\mathcal{W}$ follows a Gaussian distribution and picks a latent code located in a high-density region. To greatly benefit from considering the latent manifold, we employ bounded latent codes. Since we know the shape of $\mathcal{Z}$, we completely utilize the information of the $\mathcal{Z}$ distribution.

3.3. Inversion to $\mathcal{F}/\mathcal{Z}^+$

We introduce an inversion algorithm that projects images to the $\mathcal{F}/\mathcal{Z}^+$ space. We obtain latent codes by a hybrid method that first projects a target image to $\mathcal{F}$ by using a pretrained encoder to obtain an initial base code and then directly optimizes the base and detail codes.

Given an input image $\boldsymbol{x}$, we find a latent code $\boldsymbol{z}^*$ that reconstructs $\boldsymbol{x}$ by optimizing an objective function L_{recon}, which is defined as

$$L_{\text{recon}}(\boldsymbol{z}^*) = L_{\text{MSE}}(\boldsymbol{z}^*) + \lambda_{\text{per}} L_{\text{per}}(\boldsymbol{z}^*), \tag{1}$$

where L_{MSE} and L_{per} are the mean squared error (MSE) and perceptual losses [22,52], respectively. The hyperparameter λ_{per} controls the contributions of L_{MSE} and L_{per}. MSE loss is defined as $L_{\text{MSE}}(\boldsymbol{z}^*) = \|\boldsymbol{x} - G(\boldsymbol{z}^*)\|^2$, and perceptual loss is defined as $L_{\text{per}}(\boldsymbol{z}^*) = \|\phi(\boldsymbol{x}) - \phi(G(\boldsymbol{z}^*))\|^2$ with the LPIPS network ϕ, which is a pretrained network with the VGG backbone. MSE and perceptual losses are the distances between the target and inverted images in the data and feature spaces, respectively. We use perceptual loss to enhance reconstruction quality and particularly to avoid blurred images [22, 28, 52].

For efficient optimization, we initialize the base code by employing an encoder. We first compute a rough base code using an encoder and then optimize a precise base code. We train an encoder $E : \mathcal{X} \to \mathcal{F}$ by optimizing the loss function:

$$\begin{aligned} L_{\text{enc}} =& \|G(E(\boldsymbol{x}_\downarrow), \boldsymbol{w}^s_{M+}) - \boldsymbol{x}\|^2 \\ &+ \lambda_{\text{enc}} \|\phi(G(E(\boldsymbol{x}_\downarrow), \boldsymbol{w}^s_{M+})) - \phi(\boldsymbol{x})\|^2, \end{aligned} \tag{2}$$

where $G : \boldsymbol{f} \times \boldsymbol{w}_{M+} \mapsto \boldsymbol{x}$ is a generator (we use a different notation from the above-mentioned G in Section 3 to emphasize the various inputs), $\boldsymbol{f}^s$ and $\boldsymbol{w}^s_{M+}$ are sampled codes corresponding to the sampled latent code $\boldsymbol{z} \in \mathcal{Z}$, and λ_{enc} is the weight of the second term. Training images $\boldsymbol{x} = G(\boldsymbol{f}^s, \boldsymbol{w}^s_{M+})$ are reconstructed from the sampled codes $\boldsymbol{f}^s$ and $\boldsymbol{w}^s_{M+}$. We train the encoder on only the pairs of sampled latent code and generated images (no real image is used) because $\boldsymbol{w}^s_{M+}$ corresponding to given images is unavailable. The downsampled images $\boldsymbol{x}_\downarrow$ have a resolution $8\times$ larger than the feature space $\mathcal{F}$. To consider the encoded latent codes, we use a regularization that penalizes the distance between initial and current latent code in optimization. The regularization term for $\boldsymbol{f}$ is defined as

$$L_{\text{reg}}(\boldsymbol{z}^*) = \|\boldsymbol{f}^0 - \boldsymbol{f}\|^2, \tag{3}$$

where $\boldsymbol{f}^0 = E(\boldsymbol{x})$. The loss prevents the latent code $\boldsymbol{f}$ from straying too far from the encoded code $\boldsymbol{f}^0$. The final objective function for GAN inversion on $\mathcal{F}/\mathcal{Z}^+$ is given by

$$L(\boldsymbol{z}^*) = L_{\text{recon}}(\boldsymbol{z}^*) + \lambda_{\text{reg}} L_{\text{reg}}(\boldsymbol{z}^*), \tag{4}$$

where λ_{reg} is the weight of the regularization loss. We retract the latent codes $\boldsymbol{z}_{M+}$ to the surface of the hypersphere of radius $\sqrt{512}$ after every iteration to compute precise gradients of the latent $\boldsymbol{z}_{M+}$. From the definition of hypersphere, we update each latent code $\boldsymbol{z}_i \in \boldsymbol{z}_{M+}$ independently by calculating

$$\boldsymbol{z}_i = \sqrt{512} \frac{\boldsymbol{z}_i}{|\boldsymbol{z}_i|}. \tag{5}$$

Moreover, in latent editing, our approach ensures the presence of the edited code within the latent space by calculating Eq. (5) after performing editing operations on $\mathcal{Z}^+$.

4. Experiments

We evaluate our latent spaces from two aspects: reconstruction quality and editing quality. For a fair comparison, we note that all experimental settings strictly adhere to previous studies, namely, we compare the editing results on the off-the-shelf latent editing algorithms. Our goal is not to improve the quality of semantic directions, but rather to introduce editing robustness into existing GAN inversion methods. For the reconstruction quality comparison, we verify that our spaces do not underperform the compared spaces.

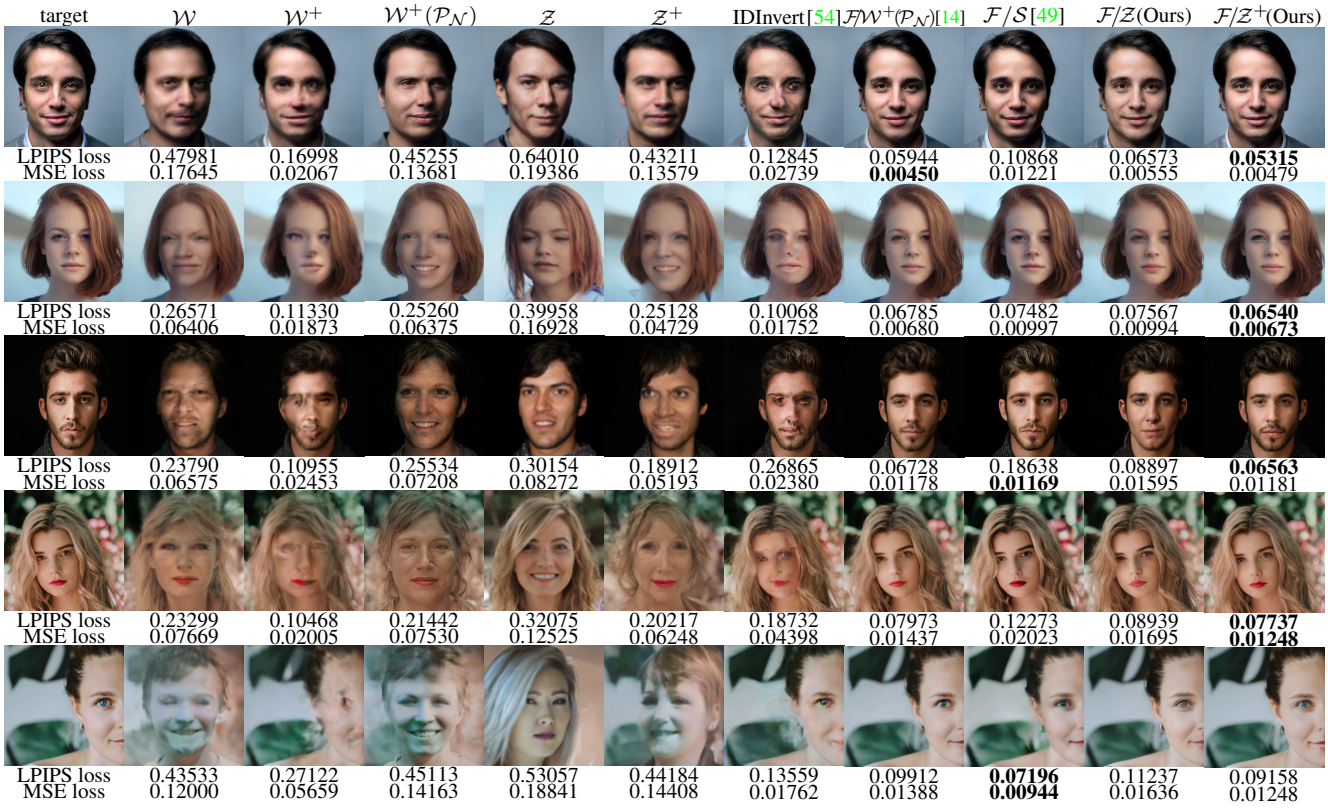


Figure 3. Inverted images with different latent spaces. Any latent space complemented by the $\mathcal{F}$ space achieves faithful reconstruction, whereas the other spaces fail. $\mathcal{F}/\mathcal{Z}^+$ achieves high-quality reconstructions on par with $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$ qualitatively, and the results of $\mathcal{F}/\mathcal{Z}^+$ are competitive with that of $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$ quantitatively. Image credits are listed in Supplementary Material.

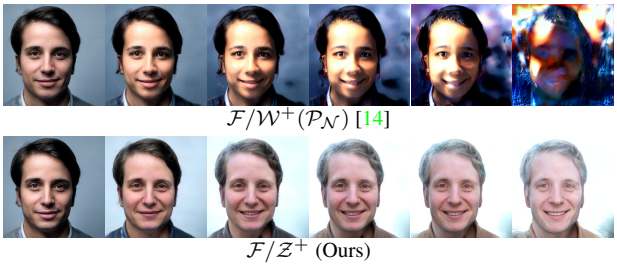


Figure 4. Comparison of editing with random directions with high intensities of up to 20. Editing on $\mathcal{F}/\mathcal{Z}^+$ always produces natural images, whereas editing on $\mathcal{F}/\mathcal{W}(\mathcal{P}_\mathcal{N})$ produces images lacking in quality as shown in the rightmost images.

For the editing quality comparison, we demonstrate that our spaces preserve the perceptual quality of edited images better than other spaces.

Implementation details. For the inversion, we iteratively update the latent code 1200 times with the Adam optimizer. We set the learning rate to 0.01 and $\lambda_{\text{enc}} = \lambda_{\text{per}} = \lambda_{\text{reg}} = 10$.

Reconstruction quality comparison. We first compare the reconstruction performance using StyleGAN2 trained on FFHQ [19] in both qualitative and quantitative ways. Figure 3 shows the reconstructed results, LPIPS loss, and MSE loss for five commonly used benchmark images. We test five standalone spaces without $\mathcal{F}$ (*i.e.*, $\mathcal{W}$, $\mathcal{W}^+$, $\mathcal{W}^+(\mathcal{P}_\mathcal{N})$, $\mathcal{Z}$, and $\mathcal{Z}^+$), IDInvert [54], and four latent spaces with $\mathcal{F}$ (*i.e.* $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$, $\mathcal{F}/\mathcal{S}$, $\mathcal{F}/\mathcal{Z}$, and $\mathcal{F}/\mathcal{Z}^+$). We can see that all standalone spaces fail to reconstruct the image details well. Furthermore, the expansion of $\mathcal{Z}$ to $\mathcal{Z}^+$ demonstrates an improvement in reconstruction quality because $\mathcal{Z}^+$ is 18-fold larger than $\mathcal{Z}$. On the other hand, $\mathcal{F}$-based latent spaces (i.e., $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$, $\mathcal{F}/\mathcal{S}$, $\mathcal{F}/\mathcal{Z}$, and $\mathcal{F}/\mathcal{Z}^+$) reconstruct images effectively because the feature space $\mathcal{F}$ magnifies the latent space's capacity. Among them, $\mathcal{F}/\mathcal{W}^+$, $\mathcal{F}/\mathcal{S}$, and our $\mathcal{F}/\mathcal{Z}^+$ have the finest visual reconstruction quality.

Table 1. Quantitative comparison of latent spaces with the average MSE, SSIM, MS-SSIM, and LPIPS on the test sets of CelebA-HQ. $\mathcal{F}/\mathcal{Z}^+$ yields a comparable performance to $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$.

	$\mathcal{Z}$	$\mathcal{Z}^+$	$\mathcal{W}^+(P_N)$ [56]	$\mathcal{F}/\mathcal{Z}$	$\mathcal{F}/\mathcal{W}^+(P_N)$ [14]	$\mathcal{F}/\mathcal{Z}^+$(Ours)
MSE↓	0.1211	0.0680	0.0772	0.0235	0.0162	0.0165
SSIM↑	0.6062	0.6708	0.6634	0.7352	0.7522	0.7524
MS-SSIM↑	0.5523	0.6376	0.6190	0.7730	0.7974	0.7990
LPIPS↓	0.4721	0.3946	0.4186	0.2862	0.2625	0.2603

The qualitative observations are also validated by the LPIPS, structural similarity index measure (SSIM), multi-scale SSIM (MS-SSIM) [43], and MSE. We show the scores

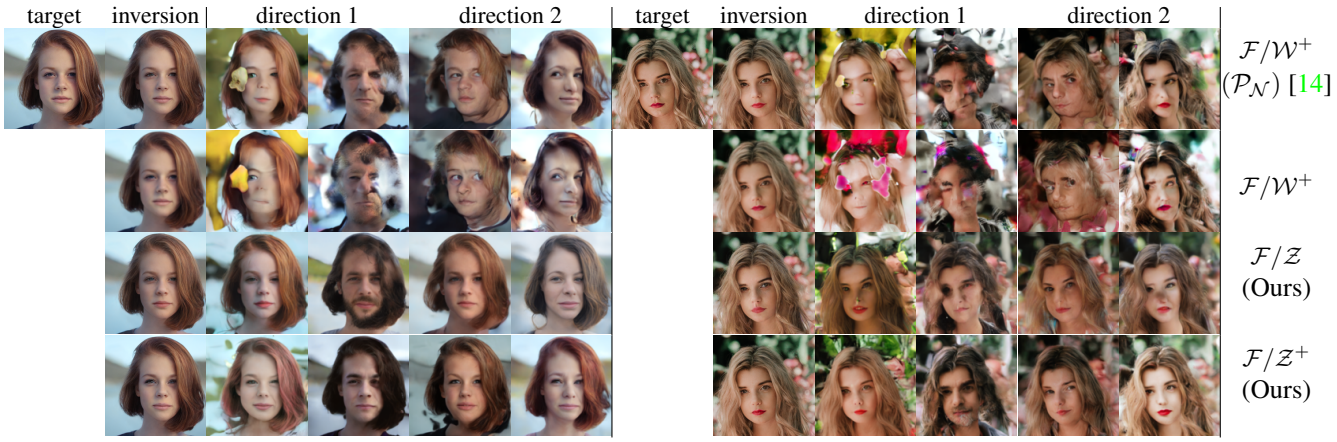


Figure 5. Comparison of editing with GANSpace directions. Although the spaces with $\mathcal{W}^+$ fail to preserve the structures of generated faces, our spaces properly preserve them.

Figure 6. Comparison of editing using $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$ [14], $\mathcal{F}/\mathcal{S}$ [49], and $\mathcal{F}/\mathcal{Z}^+$ with directions obtained by InterfaceGAN [35].

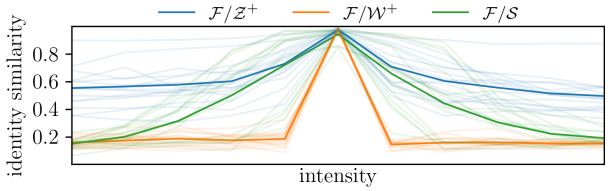


Figure 7. Quantitative comparison of editing quality. We compute the identity similarity between the target and edited images. Each light-color line indicates the results of a semantic direction. Each deep-color line indicates the mean of the results of each method. Achieving high identity similarity in cases with high intensity indicates that our method has a higher editing quality than the compared methods.

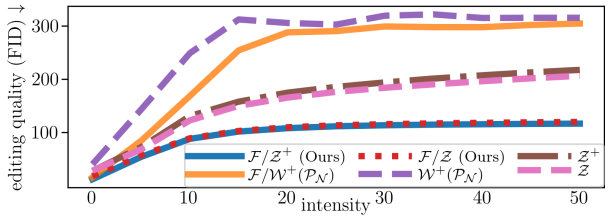


Figure 8. Quantitative comparison of editing quality. We compute the FID between the target and edited images with a high intensities of up to 50. Our method achieves a lower FID in edited images with high intensity, indicating that our method has a higher editing quality than the compared methods.

on the test set of CelebA-HQ [16] in Tab. 1. Similarly to Fig. 3, both the $\mathcal{F}/\mathcal{Z}^+$ and $\mathcal{F}/\mathcal{W}^+(P_N)$ spaces consistently demonstrate comparable levels of reconstruction performance when evaluated using quantitative metrics. However, $\mathcal{F}/\mathcal{W}^+(P_N)$ yields edited images that are less realistic, as will be shown afterward. Later in this paper, we also show that our approach achieves a comparable level of reconstruction performance even on other algorithms.

Editing quality comparison. To assess the robustness of inverted latent codes, we first examine if they can move freely in the latent space using random directions. To this end, we sample a 512-dimensional vector from a Gaussian distribution with a mean of 0 and a variance of 0.04, and then we add the sampled editing vector scaled with the editing intensity to an inverted code. As seen in Fig. 4, when the editing intensity is increased (right-hand images), the edited images produced using $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$ exhibit corruption, indicating that the inverted code deviates from the latent space. In contrast, our $\mathcal{F}/\mathcal{Z}^+$ effectively maintains the quality of edited images regardless of the editing intensity, owning to the hyperspherical constraint imposed by $\mathcal{Z}$.

Subsequently, the images are edited using actual semantic directions. We use GANSpace [12] to discover semantic directions. We present two edited images with intensities of -2 and 2 for each editing direction (Fig. 5). It is obvious that the edited images produced by utilizing $\mathcal{F}/\mathcal{W}^+(P_N)$ and $\mathcal{F}/\mathcal{W}^+$ exhibit deficiencies in facial parts or unexpected water droplets. On the other hand, it is observed that both $\mathcal{F}/\mathcal{Z}$ and $\mathcal{F}/\mathcal{Z}^+$ spaces consistently maintain the editing quality of images. We also show the results obtained using two more directions different from those in Fig. 5. Moreover, for each direction, we edit the images using six differ-

Figure 9. Results of editing with StyleGAN1 trained on the LSUN Cat dataset. Editing using our $\mathcal{F}/\mathcal{Z}^+$ preserves the image content (cat).

ent intensities within the range of [-2.0, 2.0].

Additionally, we compare the edited images with InterfaceGAN [35], as illustrated in Fig. 6. Our methods more effectively mitigate the distortions appearing in edited images than the other competing methods.

Finally, we evaluate the editing quality of our approach quantitatively. For this comparison, we measure the similarity between the original and edited images. The similarity comparison shows editing quality because edited images with collapses lead to low similarity scores. We use MTCNN [51] as the face detector and InceptionResNet V1 [41] trained on VGGFace2 [7] as the feature extractor. We cannot assume that the intensity in the Z distribution is on the same scale as in the W and S distributions and need to carefully design the experiments for a fair comparison. To this end, we use normalized intensity and observe the convergence of the editing performance. We measure the changes in the logit value Δ_a of a pretrained classifier for attribute a [25] between the original and edited images with intensity α, and we normalize the intensity by Δ_a. We use InterfaceGAN to obtain *makeup*, *smiling*, and *eyeglass* directions. To compute identity similarity, we use cosine similarity between the target and edited images. For each method, we plot the identity similarities between the original and edited images with each normalized intensity of editing in Fig. 7. We plot 15 lines for each method (five targets $\times$ three semantic directions). The similarity score of $\mathcal{F}/\mathcal{S}$ gradually decreases compared with that of $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$. Although the similarity scores of the compared methods drop to 0.2, our $\mathcal{F}/\mathcal{Z}^+$ maintains a high similarity score even when the editing intensity is high.

We also provide the FID comparison on CelebA-HQ test set. The images are edited using a wide range of intensities [0, 50] with a step size of 5. For each intensity, we measure the FID between the target and edited images. We plot all FIDs in Fig. 8. In contrast to the latent spaces utilizing $\mathcal{W}^+$, our methods obtain stable FID scores even when the editing intensity is high. The FIDs of $\mathcal{F}/\mathcal{Z}^+$ flatten off after an intensity of more than 20, showing that $\mathcal{F}/\mathcal{Z}^+$ has superior editing quality than $\mathcal{F}/\mathcal{W}^+$ regardless of the scale of the latent space. These observations support the qualitative evaluation finding that $\mathcal{F}/\mathcal{Z}^+$ achieves high editing quality.

Editing comparison on another GAN model. We evaluate the effectiveness of the space on another GAN model. Figure 9 shows the results of editing with StyleGAN1 [19] trained on the LSUN Cat dataset. BDInvert [14] results in a complete corruption of the cat's face in the edited image, which is not the case in our method. Together with the results discussed above, we may conclude that our space maintains robustness and generalization, regardless of the specific GAN model employed. More results on other datasets are provided in Supplementary Material.

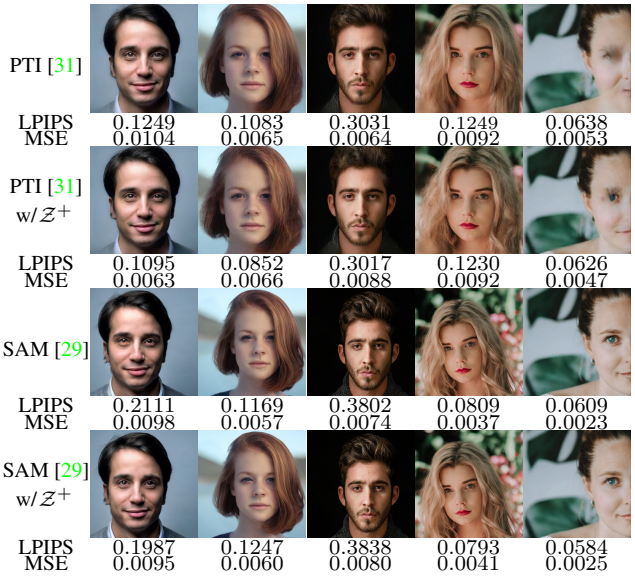


Figure 10. Reconstruction comparisons on state-of-the-arts. The 1st and 3rd rows are inverted results of PTI and SAM. The 2nd and 4th rows are inverted results of the methods that use $\mathcal{Z}^+$ instead of $\mathcal{W}$ or $\mathcal{W}^+$. The results indicate that we can replace the original latent space with $\mathcal{Z}^+$ without losing reconstruction quality to improve editing quality.

All the comparisons above reveal that $\mathcal{F}/\mathcal{Z}^+$ is superior to the existing spaces. The utilization of the hypersphere space in $\mathcal{F}/\mathcal{Z}^+$ improves editing quality compared with $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$ without sacrificing the reconstruction quality. Indeed, the quantitative performance of $\mathcal{F}/\mathcal{Z}^+$ is comparable to that of $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$. Furthermore, the inverted results of $\mathcal{F}/\mathcal{Z}^+$ and $\mathcal{F}/\mathcal{W}^+(\mathcal{P}_\mathcal{N})$ are nearly identical.

Integration $\mathcal{Z}^+$ into state-of-the-art GAN inversion. We further demonstrate the effectiveness of our $\mathcal{Z}^+$ space. Figure 10 shows the images reconstructed by PTI [31], SAM [29], and their $\mathcal{Z}^+$ version. We can see that the use of $\mathcal{Z}^+$ in PTI and SAM does not sacrifice reconstruction quality. Quantitatively, the reconstruction quality of our PTI extension is consistent with that of PTI. PTI achieves an SSIM of 0.7299 and a MSE of 0.0136 for the CelebA-HQ test set. Our method (PTI with $\mathcal{Z}^+$) achieves an SSIM of 0.7286 and a MSE of 0.0131. Figure 11 shows that integrating the $\mathcal{Z}^+$ space into SoTA inversion methods relaxes editing distortions. We additionally provide examples of editing using

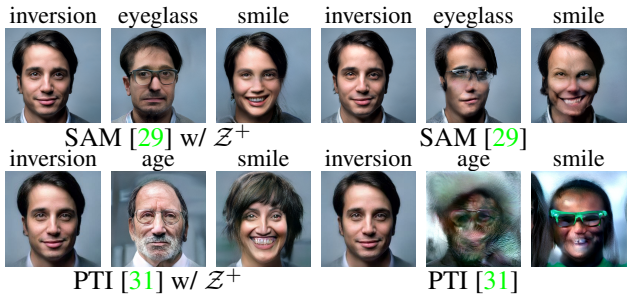


Figure 11. Comparison of editing using SAM and PTI. Replacing $\mathcal{W}^+$ in SAM or $\mathcal{W}$ in PTI with $\mathcal{Z}^+$ prevents the deterioration of the perceptual quality of edited images.

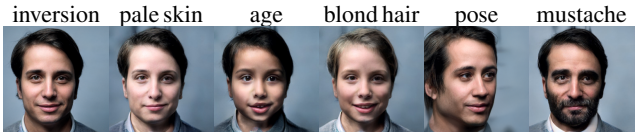


Figure 12. Examples of editing with InterfaceGAN using $\mathcal{F}/\mathcal{Z}^+$. $\mathcal{F}/\mathcal{Z}^+$ can be used in editing with a wide range of attributes including spatial ones.

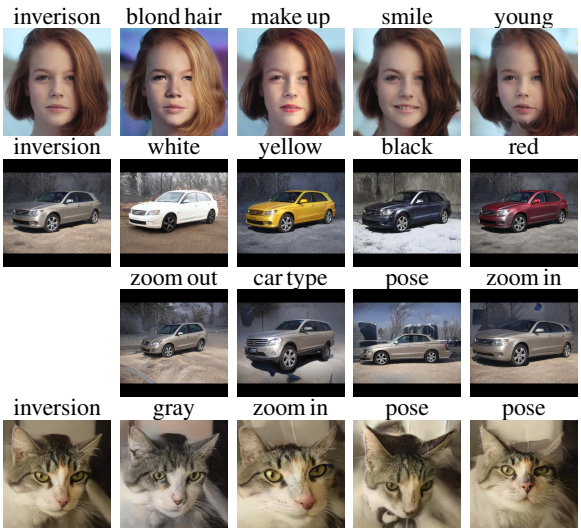


Figure 13. Examples of editing with Local Basis [9]. Local Basis performs well on real image editing as well as synthesized one owing to our method.

$\mathcal{F}/\mathcal{Z}^+$ with InterfaceGAN directions in Fig. 12 including geometric editing. The results show that $\mathcal{Z}^+$ naturally performs latent editing.

Latent editing with Local Basis. Similarly to GANSpace [12], Local Basis [9] finds semantic directions. However, because it requires latent codes in $\mathcal{Z}$, it lacks the capacity to edit real images. In contrast, the latent code in $\mathcal{Z}$ corresponding to real image is known in our method, enabling us to effectively utilize Local Basis for real image editing, as illustrated in Fig. 13.

Inversion for 3D GANs. Since our approach simply replaces latent spaces, it is not limited in 2D StyleGANs. We can naturally integrate our approach into StyleGAN-based 3D GANs. We also evaluate our approach in 3D GANs. We use EG3D [8] trained on FFHQ and set HFGI3D [48] as the inversion method. We first invert the target image with HFGI3D and HFGI3D with $\mathcal{Z}^+$. Then, we generate different view images and edit the obtained latent codes using InterfaceGAN with the *young* direction. Moreover, in 3D GANs, our approach can reconstruct target images well, generate different view images, and edit the image without losing perceptual quality (Fig. 14).

Figure 14. Inversion, novel view, and edited results in HFGI3D [48]. Our method avoids the collapse of an edited image while performing inversion and novel view synthesis on par with the original HGFI3D.

5. Conclusion

In this study, we the revisited $\mathcal{Z}$ space with the hyperspherical prior for GAN inversion. We integrated constrained latent space $\mathcal{Z}^+$ into expressive inversion methods, resulting in the presented methods (*e.g.*, $\mathcal{F}/\mathcal{Z}^+$). Our thorough experiments on PTI, SAM, $\mathcal{F}/\mathcal{W}^+$, and HFGI3D demonstrate that we can preserve the perceptual quality of edited images while maintaining reconstruction quality on par with capable baselines by replacing an unconstrained space (*e.g.*, $\mathcal{W}^+$) to $\mathcal{Z}^+$. Our method also allows editing real images using 2D/3D GANs without concern about image collapse.

Limitation. Since this study focuses on the editing quality, whether semantic directions correspond to a certain attribute is not evaluated. Latent editing for $\mathcal{Z}$ and $\mathcal{Z}^+$ is less explored than that for $\mathcal{W}$, $\mathcal{W}^+$, and $\mathcal{S}$. Investigating local editing methods (*e.g.*, Local Basis) and nonlinear editing methods may contribute to accurate latent editing in $\mathcal{Z}^+$.

Acknowledgements. This work was supported by the Institute for AI and Beyond of the University of Tokyo, D-CORE Grant from Microsoft Research Asia, ROIS NII Open Collaborative Research 2023_23FC01, JSPS KAKENHI Grant Numbers JP22K17947, JP23KJ0381, JP23H03449, and JP22H00540.

References

[1] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3DAvatarGAN: Bridging domains for personalized editable avatars. In *CVPR*, pages 4552–4562, 2023. 2

[2] 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, 3

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN++: How to edit the embedded images? In *CVPR*, pages 8296–8305, 2020. 1, 2, 3

[4] Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Hao Tang, Gordon Wetzstein, Leonidas Guibas, Luc Van Gool, and Radu Timofte. 3d-aware video generation. *Transactions on Machine Learning Research*, 2023. 2

[5] Amit H Bermano, Rinon Gal, Yuval Alaluf, Ron Mokady, Yotam Nitzan, Omer Tov, Oren Patashnik, and Daniel Cohen-Or. State-of-the-art in the architecture, methods and applications of StyleGAN. In *Computer Graphics Forum*, volume 41, pages 591–611, 2022. 1

[6] Ananta R. Bhattarai, Matthias Nießner, and Artem Sevastopolsky. TriPlaneNet: An encoder for EG3D inversion. In *WACV*, 2024. 2

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2018. 7

[8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, pages 16123–16133, 2022. 2, 8

[9] Jaewoong Choi, Junho Lee, Changyeon Yoon, Jung Ho Park, Geonho Hwang, and Myungjoo Kang. Do not escape from the manifold: Discovering the local coordinates on the latent space of GANs. In *ICLR*, 2022. 2, 8

[10] Zijian Dong, Xu Chen, Jinlong Yang, Michael Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to generate 3D avatars from 2D image collections. In *ICCV*, pages 14916–14927, 2023. 2

[11] Qianli Feng, Viraj Shah, Raghudeep Gadde, Pietro Perona, and Aleix Martinez. Near perfect GAN inversion. *arXiv preprint arXiv:2202.11833*, 2022. 1, 2

[12] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. In *NeurIPS*, pages 9841–9850, 2020. 1, 2, 6, 8

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *CVPR*, pages 1501–1510, 2017. 2

[14] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. GAN inversion for out-of-range images with geometric transformations. In *ICCV*, pages 13941–13949, 2021. 1, 2, 3, 5, 6, 7

[15] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *CVPR*, pages 10124–10134, 2023. 2

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 6

[17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 33:12104–12114, 2020. 1

[18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. pages 852–863, 2021. 1

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1, 5, 7

[20] 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

[21] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3D GAN inversion with pose optimization. In *WACV*, pages 2967–2976, 2023. 2

[22] 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 *CVPR*, pages 4681–4690, 2017. 4

[23] Minjun Li, Yanghua Jin, and Huachun Zhu. Surrogate gradient field for latent space manipulation. In *CVPR*, pages 6529–6538, 2021. 2

[24] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *arXiv preprint arXiv:2306.08768*, 2023. 2

[25] Ji Lin, Richard Zhang, Frieder Ganz, Song Han, and Jun-Yan Zhu. Anycost GANs for interactive image synthesis and editing. In *CVPR*, pages 14986–14996, 2021. 7

[26] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, pages 2437–2445, 2020. 3, 4

[27] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag Your GAN: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 1

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019. 4

[29] 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 *CVPR*, pages 11399–11409, 2022. 1, 2, 3, 7, 8

[30] 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. 2

[31] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 42(1):1–13, 2022. 1, 2, 7, 8

[32] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In *NeurIPS*, pages 17480–17492, 2021. 2

[33] Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2022. 2

[34] 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

[35] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterfaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE TPAMI*, 44(4):2004–2018, 2020. 6, 7

[36] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *CVPR*, pages 1532–1540, 2021. 1, 2

[37] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3D generation on ImageNet. In *ICLR*, 2022. 2

[38] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. EpiGRAF: Rethinking training of 3D GANs. *NeurIPS*, 35:24487–24501, 2022. 2

[39] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. AgileGAN: stylizing portraits by inversion-consistent transfer learning. *ACM TOG*, 40(4):1–13, 2021. 2, 3

[40] Michael Stengel, Koki Nagano, Chao Liu, Matthew Chan, Alex Trevithick, Shalini De Mello, Jonghyun Kim, and David Luebke. AI-mediated 3D video conferencing. In *ACM SIGGRAPH Emerging Technologies*, 2023. 2

[41] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 7

[42] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM TOG*, 40(4):1–14, 2021. 1, 2

[43] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402, 2003. 5

[44] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for StyleGAN image generation. In *CVPR*, pages 12863–12872, 2021. 1, 2, 3

[45] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. StyleAlign: Analysis and applications of aligned StyleGAN models. In *ICLR*, 2022. 2

[46] Jonas Wulff and Antonio Torralba. Improving inversion and generation diversity in StyleGAN using a gaussianized latent space. *arXiv preprint arXiv:2009.06529*, 2020. 3

[47] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN inversion: A survey. *IEEE TPAMI*, 45(3):3121–3138, 2022. 1, 2

[48] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3D GAN inversion by pseudo-multi-view optimization. In *CVPR*, pages 321–331, 2023. 2, 8

[49] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A style-based GAN encoder for high fidelity reconstruction of images and videos. In *ECCV*, pages 581–597, 2022. 3, 5, 6

[50] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3D GAN inversion through geometry and occlusion-aware encoding. In *ICCV*, pages 2437–2447, 2023. 2

[51] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Sign. Process. Letters*, 23(10):1499–1503, 2016. 7

[52] 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. 4

[53] Jiapeng Zhu, Ruili Feng, Yujun Shen, Deli Zhao, Zheng-Jun Zha, Jingren Zhou, and Qifeng Chen. Low-rank subspaces in GANs. *NeurIPS*, pages 16648–16658, 2021. 2

[54] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, pages 592–608, 2020. 1, 5

[55] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Barbershop: GAN-based image compositing using segmentation masks. *ACM TOG*, 40(6):1–13, 2021. 1, 3

[56] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved StyleGAN embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020. 1, 2, 3, 4, 5