ROBUST GAN INVERSION

Egor Sevriugov
AI Center, Skoltech
Moscow, Russia
egor.sevriugov@skoltech.ru

Ivan Oseledets
AIRI, Moscow, Russia
AI Center, Skoltech
Moscow, Russia

ABSTRACT

Recent advancements in real image editing have been attributed to the exploration of Generative Adversarial Networks (GANs) latent space. However, a key challenge in this process is GAN inversion, which aims to accurately map images to the latent space. Current methods working on the extended latent space $W+$ struggle to achieve low distortion and high editability simultaneously. In response to this challenge, we propose an approach that operates in the native latent space W and fine-tunes the generator network to restore missing image details. This method introduces a novel regularization strategy with learnable coefficients acquired through training a randomized StyleGAN 2 model - WRanGAN, surpassing traditional approaches in terms of reconstruction quality and computational efficiency. It achieves the lowest distortion compared to traditional methods. Furthermore, we observe a slight improvement in the quality of constructing hyperplanes corresponding to binary image attributes. The effectiveness of our approach is validated through experiments on two complex datasets: Flickr-Faces-HQ and LSUN Church

1 INTRODUCTION

The advent of generative adversarial networks (GANs) has significantly advanced the realm of high-fidelity image synthesis. Among the prominent models in this domain is StyleGAN, renowned for its exceptional achievements. Furthermore, a body of research Pfau et al. (2020); Khrulkov et al. (2021); Peebles et al. (2020); Voynov & Babenko (2020); Shen et al. (2020); Härkönen et al. (2020) has underscored the rich interpretability of GANs, laying the foundation for intricate image manipulation. This distinct characteristic empowers the selective modification of specific attributes while preserving the fundamental identity of the image in relation to others. Nevertheless, the practical application of this feature to real-world images has been constrained by the necessity for precise mapping into the latent space. This intricate task, commonly referred to as GAN inversion, originally focused on accurately projecting images into the intrinsic latent space denoted as W. However, as highlighted in Abdal et al. (2019), this methodology has been demonstrated to yield substantial disparities between the original and synthesized images. Subsequent studies have redirected their efforts towards the utilization of the expanded latent space $W+$ Abdal et al. (2020); Richardson et al. (2021); Alaluf et al. (2021a); Tov et al. (2021); Wang et al. (2022); Hu et al. (2022), which enhances image reconstruction quality but compromises editability. This phenomenon, referred to as the distortion-editability tradeoff Tov et al. (2021), imposes limitations on the practical application of codes acquired within the $W+$ space.

Another innovative approach to addressing this challenge was proposed in Roich et al. (2021), introducing a minor adjustment in the generator parameters for operations conducted within the latent space W - the pivotal tuning inversion (PTI).

In this paper, we introduce a new approach based on tuning generator parameters with learnable regularization. Instead of using small, equal regularization coefficients for all model parameters, we propose to learn the optimal ones. This allows for achieving high-quality image reconstructions by tuning only 25% of the model parameters. This approach is based on a randomized version of the StyleGAN 2 model called WRanGAN, in which part of the model weights are assumed to be normally distributed with trainable mean and variance. To apply the learned regularization coefficients during inversion, we utilized the reparameterization trick Kingma & Welling (2014). We evaluated

the effectiveness of our technique on two complex datasets, the Flickr-Faces-HQ Dataset (FFHQ) Karras et al. (2019) and LSUN Churches Yu et al. (2015). Our contributions can be summarized as follows:

- We present a novel adaptive regularization scheme and investigate their effect on reconstruction quality and editability.
- We introduce WRanGAN, a model that learns appropriate regularization coefficients via a randomization of the StyleGAN 2 model.
- We evaluate WRanGAN in terms of generation, reconstruction, binary attributes extraction, and computational cost, and compare it to several baselines.

2 PROBLEM SETTING

2.1 LATENT SPACE MANIPULATION

GANs enable the creation of images that can be manipulated along semantic directions, as evidenced in Pfau et al. (2020); Khrulkov et al. (2021); Peebles et al. (2020); Voynov & Babenko (2020); Shen et al. (2020); Härkönen et al. (2020). Notably, the work presented in Härkönen et al. (2020) introduced the concept of estimating subspaces that remain invariant under random-walk diffusion for identification purposes. Additionally, in the study by Shen et al. (2020), supervision in the form of facial attribute labels was employed to discern meaningful linear directions within the latent space. Furthermore, the utilization of principal component analysis (PCA) for the identification of latent directions was proposed in Härkönen et al. (2020).

2.2 GAN INVERSION

Recent efforts have been dedicated to enhancing the quality of GAN inversion for image reconstruction, a task focusing on precisely identifying the latent code required to accurately reproduce a real image. This pursuit can be broadly classified into two groups: optimization techniques that directly manipulate the latent code to minimize a loss function Abdal et al. (2019); Pernuš et al. (2021), and encoder-based methods that use a trained encoder to generate an image Guan et al. (2020); Alaluf et al. (2021a); Richardson et al. (2021); Tov et al. (2021). Typically, these methods function within the native latent space W, potentially leading to significant visual differences compared to the original image Abdal et al. (2019). Conversely, operating in the extended latent space $W+$ provides increased expressiveness, enabling the recreation of finer image details. Nevertheless, this approach is limited by fixed generator parameters. In response to this constraint, certain strategies have suggested adjustments to the generator network to address visual inconsistencies, as seen in Roich et al. (2021). Other approaches employ hypernetworks to forecast changes in generator parameters, with the objective of reducing distortion and preserving the fidelity of the generated image, illustrated in Alaluf et al. (2021b) and Dinh et al. (2022).

2.3 DISTORTION-EDITABILITY TRADEOFF

The utilization of the extended latent space $W+$ in GAN inversion has shown notable improvement in the reconstruction of real images, albeit at the cost of reduced editability, a phenomenon known as the distortion-editability trade-off Tov et al. (2021). Notably, recent works such as Zhu et al. (2020) and Tov et al. (2021) have proposed methods to seek editable latent codes within the extended latent space $W+$. Conversely, alternative approaches, as presented in the works by Alaluf et al. (2021b) and Roich et al. (2021), offer a completely distinct solution to this challenge. Rather than striving to strike a balance between editability and distortion, these authors advocate for leveraging the advantages of projecting into the latent space W and updating generator parameters to minimize distortion. In our study, we also employed projection into the native latent space to achieve enhanced editability.

2.4 GENERATOR TUNING

Model tuning significantly improves ability to reproduce real image Roich et al. (2021); Alaluf et al. (2021b); Dinh et al. (2022). But changing the parameters of the model can damage its quality.

Figure 1: Comparison between the PTI and WRanGAN regularization strategies. PTI, employing pivotal tuning with a high regularization coefficient, falls short in comparison to WRanGAN, which incorporates optimal regularization coefficients. This is evident in the reconstruction results, particularly the enlarged area around the glasses, where PTI displays inferior performance. Furthermore, the impact on image editing is apparent, as altering the age results in darkened areas around the eyes, and removing glasses fails to produce the desired effect on the image.

In order to improve the realism of generated images after modification of the generator weights, non-saturating GAN loss was used to train hypernetworks Alaluf et al. (2021b); Dinh et al. (2022) (encoders predicting the necessary weight shift) . Despite the significant improvement in the quality of reproduction, these methods are still inferior to the PTI approach Roich et al. (2021) based on direct weight optimization. But optimization of model parameters without any additional constraints requires imposing a regularization with a high coefficient in order not to damage the realism of the generated images and forces to optimize all the parameters of the model to reach low distortion. As a result, it leads to a significant increase in the computational costs.

3 METHOD

The proposed method addresses the issue by utilizing non-uniform learnable regularization. This approach enables the setting of an appropriate regularization coefficient for each parameter based on its impact on model performance (realism of generated images). Initially, we determine suitable regularization coefficients for the inversion task through adversarial training of a generator with partially randomized parameters. Subsequently, these learned coefficients are applied in an inversion procedure that involves encoder projection and regularized optimization to minimize a specific loss function.

3.1 REGULARIZED INVERSION

In order to avoid degradation of realism in generated images, regularization term is added to optimization procedure:

$$\hat{w}, \hat{\theta}_G = \arg\min_{w,\theta_G} \mathcal{L}(G(w,\theta_G),\hat{x}) + \alpha_{\text{reg}}\|\theta_G - \theta_{G,0}\|_2^2$$

where $\hat{x}$ represents a real image, $\alpha_{\text{reg}}\|\theta_G - \theta_{G,0}\|_2^2$ is the regularization term, $G(w,\theta_G)$ is the reconstructed image, $\theta_{G,0}$ are the initial values of the generator weights.For the WRanGAN inversion, we used $\mathcal{L} = 2\mathcal{L}_2 + \mathcal{L}_{\text{LPIPS}}$ and initialized the intermediate latent code w by mapping the output

Algorithm 1 Algorithm of WRanGAN inversion

Input: real image $\hat{x}$, generator parameters $\mu_\theta, \sigma_\theta$
Parameter: regularization coefficient α_{reg}
Output: latent code w and parameterized randomization ϵ

Initialize $w = E(x)$ by the output of encoder network
Initialize ϵ with small value (10^{-4})
for number of iterations **do**
  Set generator weights $\theta_G \leftarrow \mu_\theta + \sigma_\theta \epsilon$
  Update parameters w, ϵ minimizing:

$$\mathcal{L}(G(w, \theta_G), \hat{x}) + \alpha_{\text{reg}} \|\epsilon\|_2^2$$

end for
return w,ϵ

Table 1: Quantitative comparison of memory cost on the number of randomized layers

Number of randomized layers	Relative increase in amount of parameters
4	7%
6	23%
8	39%

of the trained encoder E to the intermediate latent space W: $w = f(E(\hat{x}))$. The regularisation coefficient α_{reg} is chosen to obtain low distortion and not corrupt the model's generative quality. We considered two strategies of regularization to illustrate this paradigm:

- high regularization value (PTI)
- appropriate regularization coefficients (WRanGAN)

Based on the outcomes of our experiments, depicted in Figure 1, it is evident that a high regularization coefficient does not yield the lowest distortion, as illustrated by the enlarged area in the reconstructed image. Moreover, it results in the manipulation of specific image attributes, leading to inaccurate behavior. In the showcased example, the proposed WRanGAN model exhibits superior performance in reconstruction and consistently generates realistic images during editing: altering attributes such as the presence of eyeglasses, gender, and human age.

3.2 WRanGAN inversion

In this segment, we have explored the application of regularization to randomized model parameters $\theta_G \sim N(\mu_\theta, \sigma_\theta)$. For this purpose, we employ the reparameterization trick, which asserts that $\theta_G^i = \mu_\theta^i + \epsilon^i \sigma_\theta^i$ where $\epsilon \sim N(0, 1)$ and i denotes the index of a specific parameter. By imposing regularization on the parameter ϵ, we derive the following equation:

$$\alpha_{\text{reg}} \|\epsilon\|_2^2 = \sum_i \frac{\alpha_{\text{reg}}}{(\sigma_\theta^i)^2} (\theta_G^i - \mu_\theta^i)^2 = \sum_i \alpha_{\text{reg}}^i (\theta_G^i - \theta_{G,0}^i)^2$$

Here, we have utilized the notations $\alpha_{\text{reg}}^i = \frac{\alpha_{\text{reg}}}{(\sigma_\theta^i)^2}$ and $\mu_\theta^i = \theta_{G,0}^i$, and have derived a standard regularization formulation with distinct regularization coefficients for each randomized model parameter. The insights discussed in this section are encapsulated in Algorithm 1.

3.3 Weight randomization

The concept of randomizing the model was influenced by Bayesian GAN Saatci & Wilson (2017), in which both the generator and discriminator networks are assumed to have distributions over their internal parameters. However, randomizing the entire network is computationally intensive due to the increased number of parameters needed for training and tuning the generator parameters during

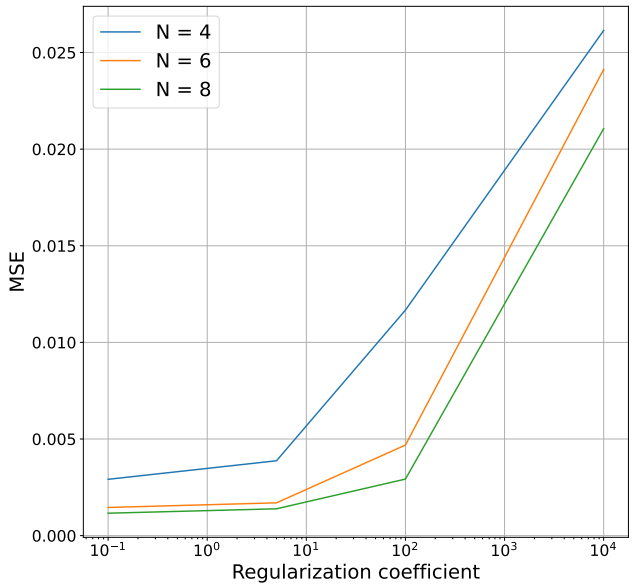

Figure 2: Dependence of MSE on number of randomized layers N versus regularization coefficient. The lower the curve the better chosen number of layers.

Algorithm 2 WRanGAN training algorithm.

Input: pretrained StyleGAN 2 weights $\theta_{G,0}$, dataset $\hat{x}$
Parameter: batch size m
Output: μ_θ, σ_θ

Initialize $\mu_\theta = \theta_{G,0}$
Initialize $\sigma_\theta = 1$ for randomized parameters
for number of training iterations **do**
Sample $z^{(1)}, ..., z^{(m)} \sim N(0, 1)$
Map to intermediate latent space $w^{(i)} = f(z^{(i)})$
Sample $\hat{x}^{(1)}, ..., \hat{x}^{(m)}$ from training dataset $\hat{x}$
Sample $\epsilon \sim N(0, 1)$ and calculate $\theta_G = \mu_\theta + \epsilon\sigma_\theta$
Update discriminator weights θ_D minimizing:

$$\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_D(D(\hat{x}^{(i)}), D(G(w^{(i)}, \theta_G)))$$

Sample $z^{(1)}, ..., z^{(m)} \sim N(0, 1)$
Map to intermediate latent space $w^{(i)} = f(z^{(i)})$
Sample $\epsilon \sim N(0, 1)$ and calculate $\theta_G = \mu_\theta + \epsilon\sigma_\theta$
Update parameters $(\mu_\theta, \sigma_\theta)$ minimizing:

$$\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_g(D(G(w^{(i)}, \theta_G)))$$

end for
return μ_θ,σ_θ

the inversion step. Our previous discussion demonstrated the process of inversion using suitable regularization coefficients, and now we will outline how to acquire these coefficients.

Table 2: Quantitative reconstruction outcomes of the WRanGAN model are juxtaposed with the StyleGAN 2 inversion methodologies, encompassing encoder and optimization-based techniques. The evaluation encompasses various standard metrics, with each metri direction of improvement indicated by the arrow (lower ↓ / higher ↑). The best results for each metric are prominently denoted in **bold**. Notably, metrics highlighted in blue signify instances where we surpass PTI in performance.

Domain	Model	Method	MSE↓	LPIPS↓		MS-SSIM↑	Time (s)↓
				VGG	Alex		
FFHQ	StyleGAN 2	E4E	0.062	0.389	0.235	0.605	1.64
		Restyle	0.035	0.335	0.154	0.72	**0.28**
		SG2 W+	0.04	0.14	0.138	0.783	97.9
		HyperStyle	0.026	0.288	0.105	0.788	0.31
		PTI	0.024	0.293	**0.06**	0.776	35.46
	WRanGAN	WRanGAN inversion	**0.007**	**0.085**	0.083	**0.929**	23.27
LSUN Church	StyleGAN 2	E4E	0.142	0.506	0.418	0.263	1.64
		Restyle	0.087	0.411	0.25	0.489	**0.28**
		SG2 W+	0.107	0.225	0.235	0.543	97.9
		PTI	0.053	0.411	**0.065**	0.643	47
	WRanGAN	WRanGAN inversion	**0.033**	**0.177**	0.224	**0.782**	23.27

Table 3: The assessment of WRanGAN model quality involved the utilization of FID, Precision, and Recall metrics across two domains: FFHQ and LSUN Church. The best results for each domain and metric are distinctly highlighted in bold.

Domain	Model	FID	Precision	Recall
Human Faces	StyleGAN 2	**4.27**	**0.7**	0.42
	WRanGAN	5.61	0.65	**0.45**
LSUN Church	StyleGAN 2	4.3	**0.61**	0.37
	WRanGAN	**3.57**	0.55	**0.42**

How many parameters to randomize? The study conducted in Alaluf et al. (2021b). aimed to identify the most effective parameters to be modified in the generator. It was determined that restricting the randomization to the last few convolutional layers, excluding the toRGB layers, while leaving the discriminator architecture unchanged, was the optimal approach. A grid search was carried out over $N = 4, 6, 8$ with varying regularization coefficients to ascertain the appropriate number of layers for randomization. The findings of this search are detailed in Figure 2, and the associated computational costs are provided in Table 1. The results led to the conclusion that randomizing only the last $N = 6$ convolutional layers produced the best outcomes with a minimal increase in computational expenses.

How to train? When training the WRanGAN model, a pre-trained model was utilized to initialize the mean value of the model parameters, $\mu_\theta = \theta_{G,0}$. A standard deviation of one was then incorporated into each randomized parameter. The generator and discriminator underwent concurrent training to achieve the global optima, as delineated in Algorithm 2.

4 Experiments

This section presents the results of the evaluation of the proposed WRanGAN model. Below are presented the details of conducted experiments: datasets, baselines, and hyperparameters.

Technical details.

- **Models:** We used the StyleGAN 2 Karras et al. (2020) model as a basis, with pre-trained models and base code for implementation taken from an open resource[1].

[1]https://github.com/rosinality/stylegan2-pytorch

Figure 3: Qualitative evaluation of WRanGAN inversion results compared to ones produced by StyleGAN 2 using various approaches for FFHQ domain. For each reconstruction provided zoomed version (interesting regions were cropped) to see the difference in details completely.

Figure 4: Qualitative evaluation of WRanGAN inversion results compared to ones produced by StyleGAN 2 using various approaches for LSUN Church domain. For each reconstruction provided zoomed version (interesting regions were cropped) to see the difference in details completely.

- **Datasets:** We trained using the Flickr-Faces-HQ Dataset (FFHQ) Karras et al. (2019) with pictures resized to resolution 256x256, and LSUN Churches Yu et al. (2015) with pictures center-cropped and resized to 256x256. We randomly sampled 1000 images from both datasets for testing.

- **WRanGAN training details:** We used standard parameters for StyleGAN 2, and trained on 2 GPUs with a batch size of 8 for $200k$ iterations.

- **WRanGAN inversion details:** For the encoder E in Algorithm 1, we used the architecture proposed by Tov et al. (2021) and trained with default parameters. We used the Adam optimizer with a learning rate of $lr = 10^{-3}$, and the number of iterations needed for convergence was set to 500. The randomization parameter was initialized with the value $\epsilon = 10^{-4}$, and we used a regularization coefficient of $\alpha_{\text{reg}} = 10^{-4}$.

Experiments were conducted on 4 Tesla V100-SXM2 GPUs with 16 GB of memory.

4.1 WRanGAN model evaluation

We evaluated the performance of our WRanGAN model by running several metrics such as FID, Precision, and Recall Kynkäänniemi et al. (2019) and comparing our results with those produced by the StyleGAN 2 model (see Table 3). WRanGAN showed an improvement in the Recall metric for both data domains, which suggests that the generator is more likely to reproduce particular real images. However, we observed a slight decrease in the Precision metric. For further details on the randomized parameters of the model and their effect on the generated images, please refer to Appendix A.

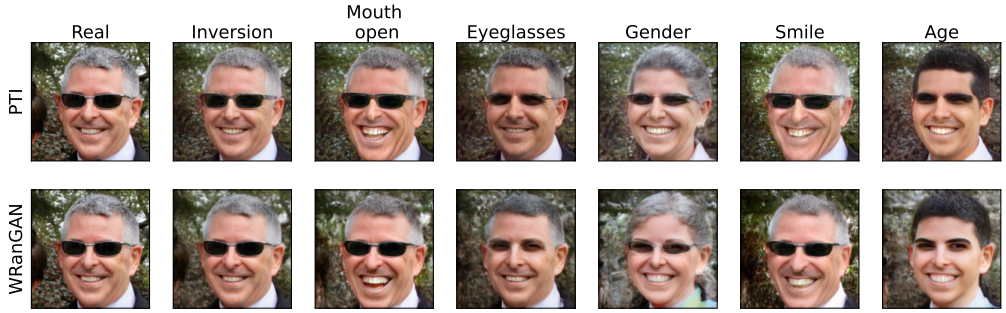

Figure 5: Qualitative evaluation of WRanGAN editing quality compared to PTI approach applied over StyleGAN 2 model.

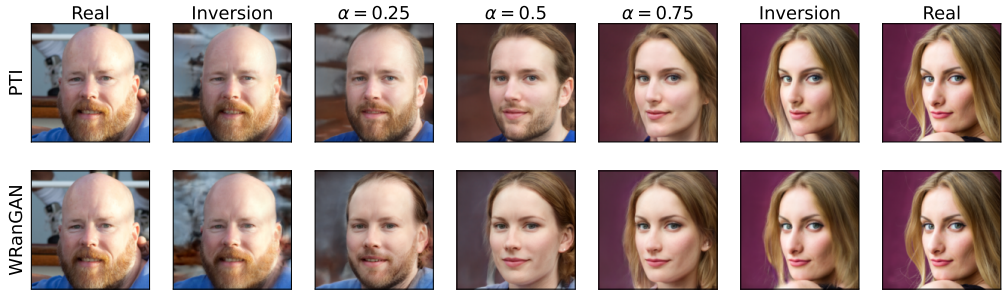

Figure 6: Qualitative evaluation of WRanGAN interpolation quality compared to PTI approach applied over StyleGAN 2 model. Here α denotes interpolation step.

4.2 Inversion quality assessment

The evaluation of inversion quality encompassed several encoder-based approaches, including e4e Tov et al. (2021), ReStyle Alaluf et al. (2021a), and HyperStyle Alaluf et al. (2021b), as well as optimization-based approaches such as SG2 W+ Karras et al. (2020) and PTI Roich et al. (2021). Standard metrics such as mean squared error (MSE), LPIPS Zhang et al. (2018) with the VGG and Alex feature network, and MS-SSIM Wang et al. (2003) were employed for the assessment. The summarized results in Table 2 indicate that the proposed WRanGAN model outperformed all other methods applied to StyleGAN 2 across most metrics. Not only did it achieve the lowest distortion, but its computational efficiency also far exceeded that of PTI, as the tuning procedure necessitated optimizing 4 times fewer parameters. Moreover, the computational speed was 1.5 and 2 times faster for FFHQ and LSUN Church domains, respectively. The visual representation of the results further demonstrated how the enhancement in reconstruction affected the image, with the WRanGAN approach capable of reproducing unique details such as bangs, the outline of the eyes, and small church windows. More detailed visualizations and an exploration of the distribution of randomized parameters for real mapped images can be found in Appendices B and C, respectively.

4.3 Editing and interpolation assessment

Our experiment was designed to verify that the WRanGAN model exhibits a similar excellent property as the StyleGAN 2 model - specifically, for any binary attribute, there exists a hyperplane in latent space such that all samples from the same side have the same attribute Shen et al. (2020). To achieve this, we trained a classifier to predict attributes including Gender, Eyeglasses, Smile, Age, and Open Mouth. Subsequently, we constructed hyperplanes in the latent space corresponding to the selected attributes and evaluated their accuracy, as detailed in Table 4. The findings from our experiment indicate that the WRanGAN model outperforms the basic StyleGAN 2 model, as evidenced in the visualization provided in Figure 5. This visualization clearly shows the significant impact of attributes such as eyeglasses in the original image on attribute editing using the PTI method, while WRanGAN exhibits exceptional performance. Furthermore, the interpolation comparison presented

Table 4: Classification accuracy (%) on separation boundaries in latent space with respect to different face attributes. The best results are highlighted in bold.

Attribute	StyleGAN 2	WRanGAN
Gender	73.9	**75.0**
Eyeglasses	99.8	**99.9**
Smile	99.5	**99.8**
Age	**99.5**	99.4
Mouth open	98.2	**98.5**

in Figure 6 highlights the superior performance of WRanGAN. Additional examples are available in Appendix D for further reference.

5 Conclusion

We introduced WRanGAN, a randomized variant of the StyleGAN 2 model that autonomously learns the optimal scaling (standard deviation) for each parameter to determine the appropriate regularization coefficient. Our non-uniform regularization coefficient approach for GAN tuning exhibited superior performance in terms of distortion and computational efficiency when compared to the highly successful pivotal tuning inversion method. Importantly, our method maintained model integrity, facilitating image editing capabilities. Furthermore, we illustrated the ease of constructing hyperplanes corresponding to standard image attributes in the FFHQ domain within the latent space of a randomized model.

Our approach reduces memory requirements per image during the inversion process, facilitating parallelized computations. Moreover, it exhibits slight dependency on the network architecture, allowing for potential adaptation to other structures such as StyleGAN 3 Karras et al. (2021).

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4431–4440, 2019. doi: 10.1109/ICCV.2019.00453.

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

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement, 2021a.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H. Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing, 2021b.

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 (CVPR)*, 2022.

Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. Collaborative learning for faster stylegan embedding, 2020. URL `https://arxiv.org/abs/2007.01758`.

Xueqi Hu, Qiusheng Huang, Zhengyi Shi, Siyuan Li, Changxin Gao, Li Sun, and Qingli Li. Style transformer for image inversion and editing. *arXiv preprint arXiv:2203.07932*, 2022.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*, 2020.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019. doi: 10.1109/CVPR.2019.00453.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, 2020. doi: 10.1109/CVPR42600.2020.00813.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

Valentin Khrulkov, Leyla Mirvakhabova, Ivan Oseledets, and Artem Babenko. Disentangled representations from non-disentangled models, 2021.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *CoRR*, abs/1904.06991, 2019.

William Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

Martin Pernuš, Vitomir Štruc, and Simon Dobrišek. High resolution face editing with masked gan latent code optimization, 2021.

David Pfau, Irina Higgins, Aleksandar Botev, and Sébastian Racanière. Disentangling by subspace diffusion. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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 (CVPR)*, pp. 2287–2296, June 2021.

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

Yunus Saatci and Andrew G Wilson. Bayesian gan. In *Advances in neural information processing systems*, pp. 3622–3631, 2017.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. In *ACM Transactions on Graphics*, volume 40, pp. 1–14, 2021. doi: 10.1145/3450626.3459838.

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.

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 (CVPR)*, 2022.

Z. Wang, Eero Simoncelli, and Alan Bovik. Multiscale structural similarity for image quality assessment. volume 2, pp. 1398 – 1402 Vol.2, 12 2003. ISBN 0-7803-8104-1. doi: 10.1109/ACSSC.2003.1292216.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, 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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

A INVESTIGATION OF WRANGAN MODEL RANDOMIZATION

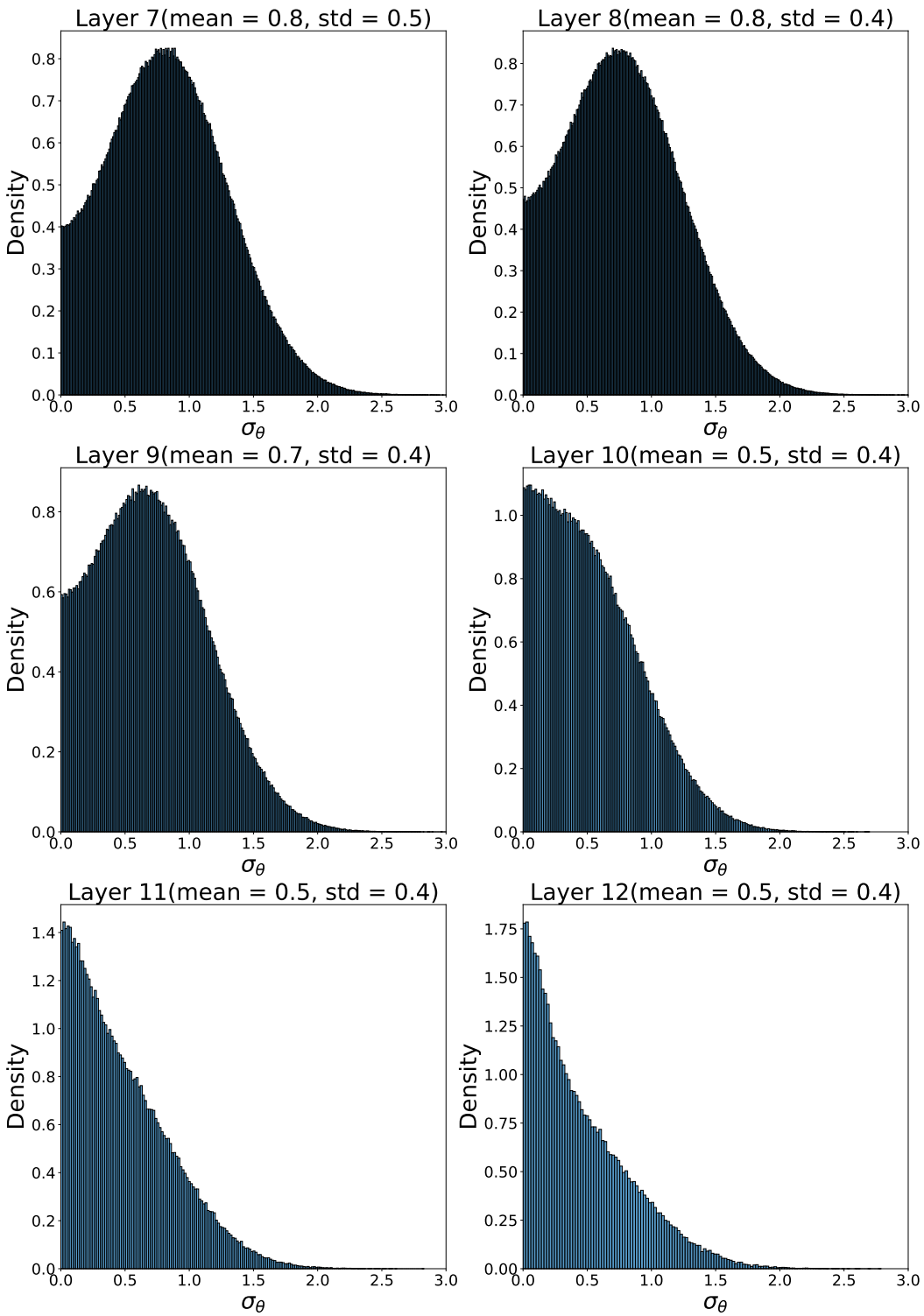

Figure 7: Dependence of distribution of randomized parameters variance with respect to index of convolutional layer for model trained on FFHQ dataset.

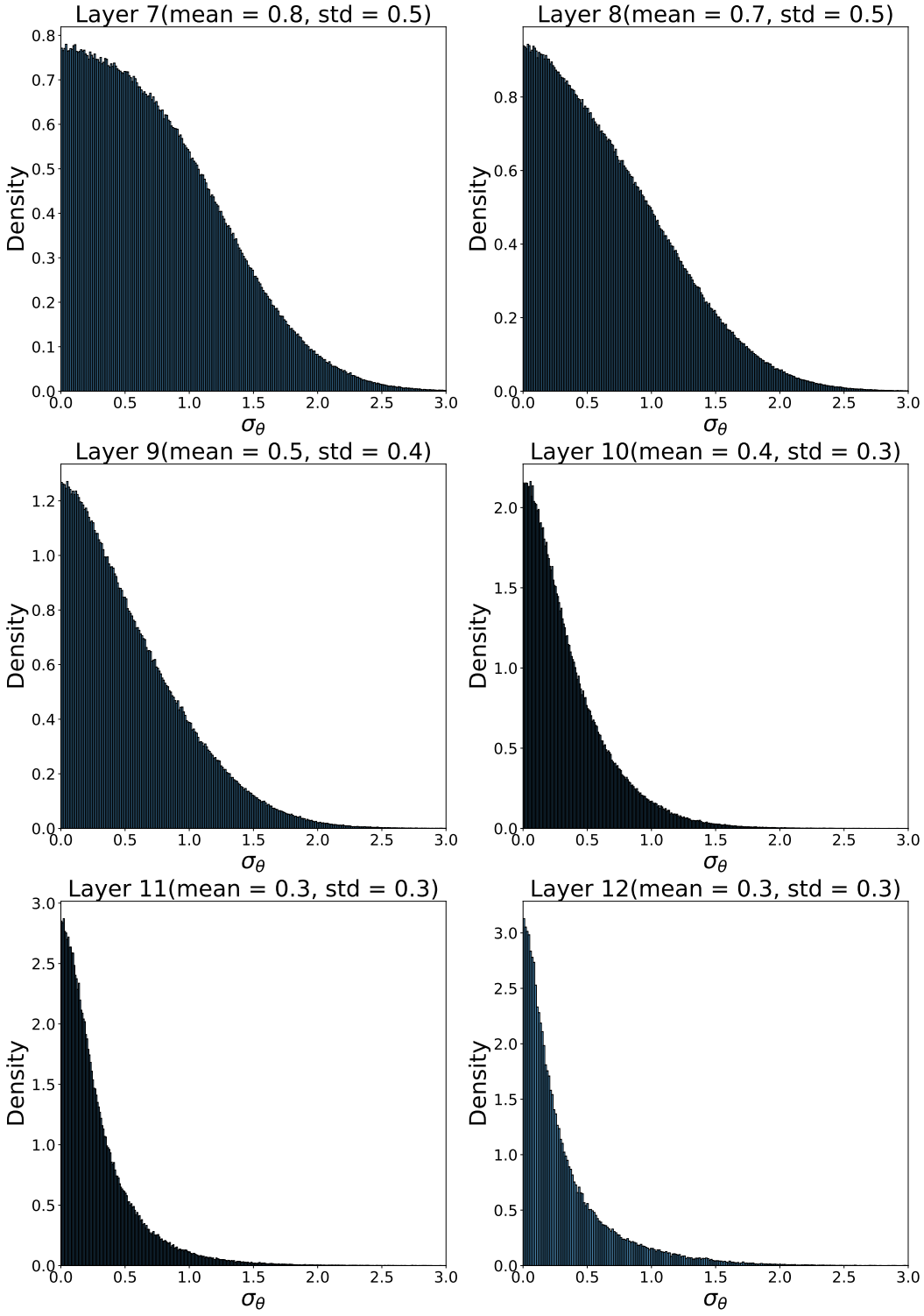

Figure 8: Dependence of distribution of randomized parameters variance with respect to index of convolutional layer for model trained on LSUN Church dataset.

The WRanGAN model is a type of generative model that optimizes two parameters - the mean and variance - in its learning process. It is hypothesized that a larger variance value allows for

Table 5: Percentage of small variances ($\sigma_\theta < 10^{-3}$) for each randomized convolutional layer.

Layer index	Percentage
7	0.13 %
8	0.16 %
9	0.18 %
10	0.29 %
11	0.40 %
12	0.52 %

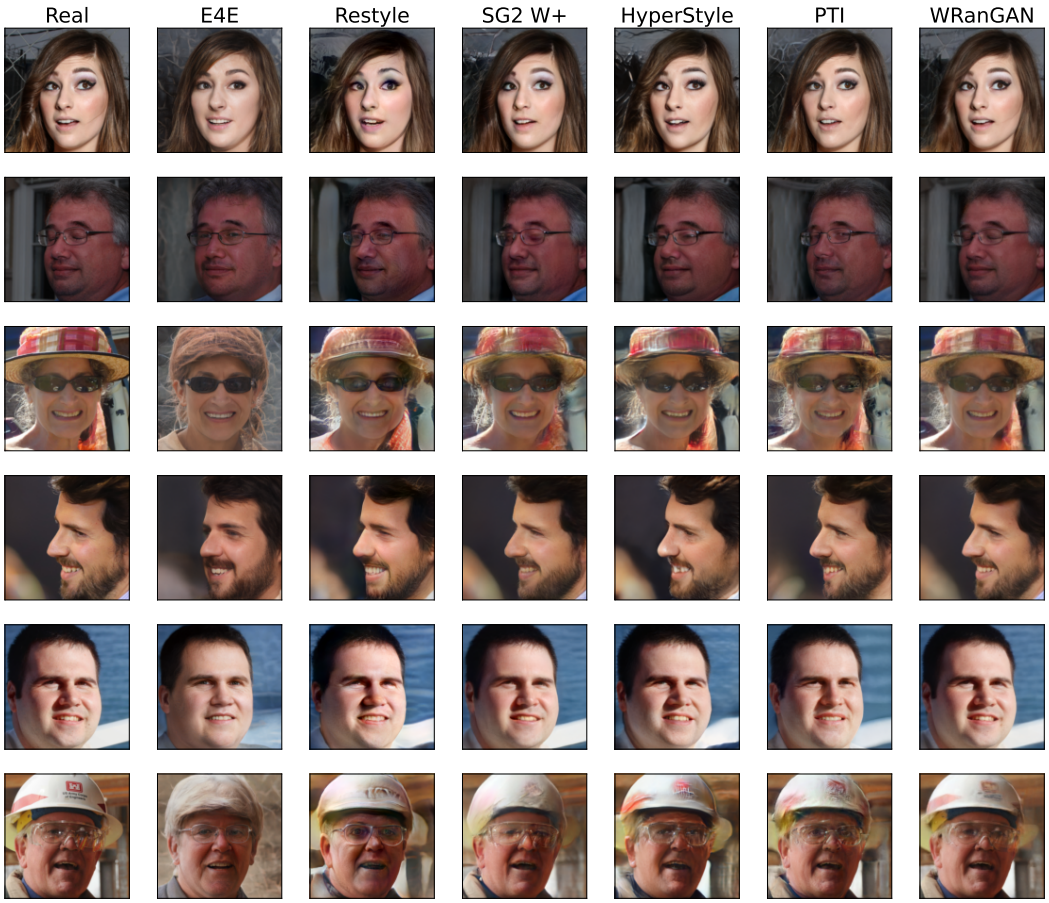

Figure 9: Qualitative reconstruction examples for images taken from FFHQ dataset.

Table 6: Influence of each randomized layer on generator output.

Layer index	MSE
7	0.017
8	0.017
9	0.015
10	0.015
11	0.014
12	0.014

more changes to be made, which can be observed by examining the distributions of the variance values of the trained WRanGAN model. Figures 7 and 8 present the distributions for the FFHQ and LSUN Church domains respectively. It can be seen that the distribution is shifted towards zero

Figure 10: Qualitative reconstruction examples for images taken from LSUN Church dataset.

as the number of randomized layers increases. In order to verify this idea, a test was conducted to measure the number of parameters that have a variance close to zero ($\sigma_\theta < 10^{-3}$) in each layer. The results, as presented in Table 5, indicate that the last layers have a greater effect on the model performance. To further investigate this idea, a procedure was conducted in which the parameters of different layers were changed and the impact of these changes on the output image was assessed by calculating the MSE metric (Table 6). The results confirm the hypothesis that the last layers have the greatest influence.

B ADDITIONAL QUALITATIVE RESULTS ON RECONSTRUCTION

The proposed approach of WRanGAN demonstrates a clear advantage in terms of reproducing unique details, as demonstrated in the additional examples of Figure 9 and Figure 10. Difference maps in Figure 11 and Figure 12 more clearly show the areas of the image that differ from the original, where the WRanGAN model has more accurately reproduced the unique details of facial skin tones, wrinkles, clothing elements, complex hairstyles, and background elements, such as the windows, friezes, pilasters, and clock in the case of churches. However, some unique details are not completely restored even with the use of the new approach.

Figure 11: Qualitative reconstruction examples for FFHQ dataset with difference map, which represents pixel wise difference between reconstructed and original images. Map colors range from purple to yellow - from exact match to maximal difference.

C Investigation of inversion procedure results

The proposed methodology utilizes a reparametrization trick to work with randomized parameters. Specifically, the generator parameter θ_g is represented as $\theta_g = \mu_\theta + \epsilon\sigma_\theta$, where ϵ is sampled from a normal distribution. Figures 13, 14, 15, 16, 17, and Figures 18, 19, 20, 21, 22, provide examples of the resulting distribution of ϵ for images taken from FFHQ and LSUN Church respectively. In general, the parameter is close to normally distributed, albeit with a variance that is 10 times lower than that which was specified at the Wrangan training stage for FFHQ, and 7 times lower for LSUN Church. Despite this reduction, the proposed methodology is still successful.

D Additional qualitative comparisons for editing

We conducted additional comparative experiments of the proposed WRanGAN approach and the PTI inversion method for the StyleGAN 2 model in two domains: the FFHQ domain, with semantic directions corresponding to binary image attributes Shen et al. (2020), and the LSUN Church domain, with the first 4 vectors obtained by PCA approach Härkönen et al. (2020). The results were captured in Figures 23, 24, 25, 26, and 27.

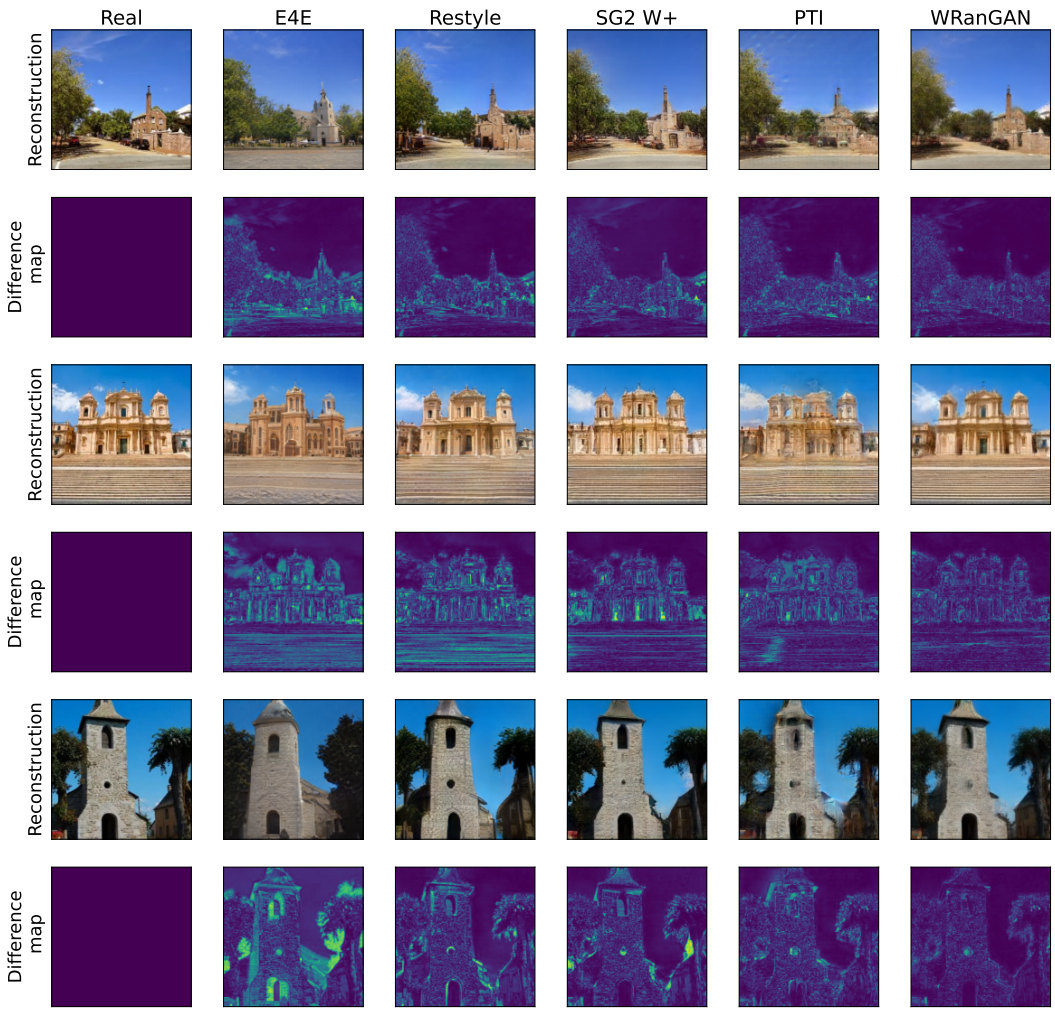

Figure 12: Qualitative reconstruction examples for LSUN Church dataset with difference map, which represents pixel wise difference between reconstructed and original images. Map colors range from purple to yellow - from exact match to maximal difference.

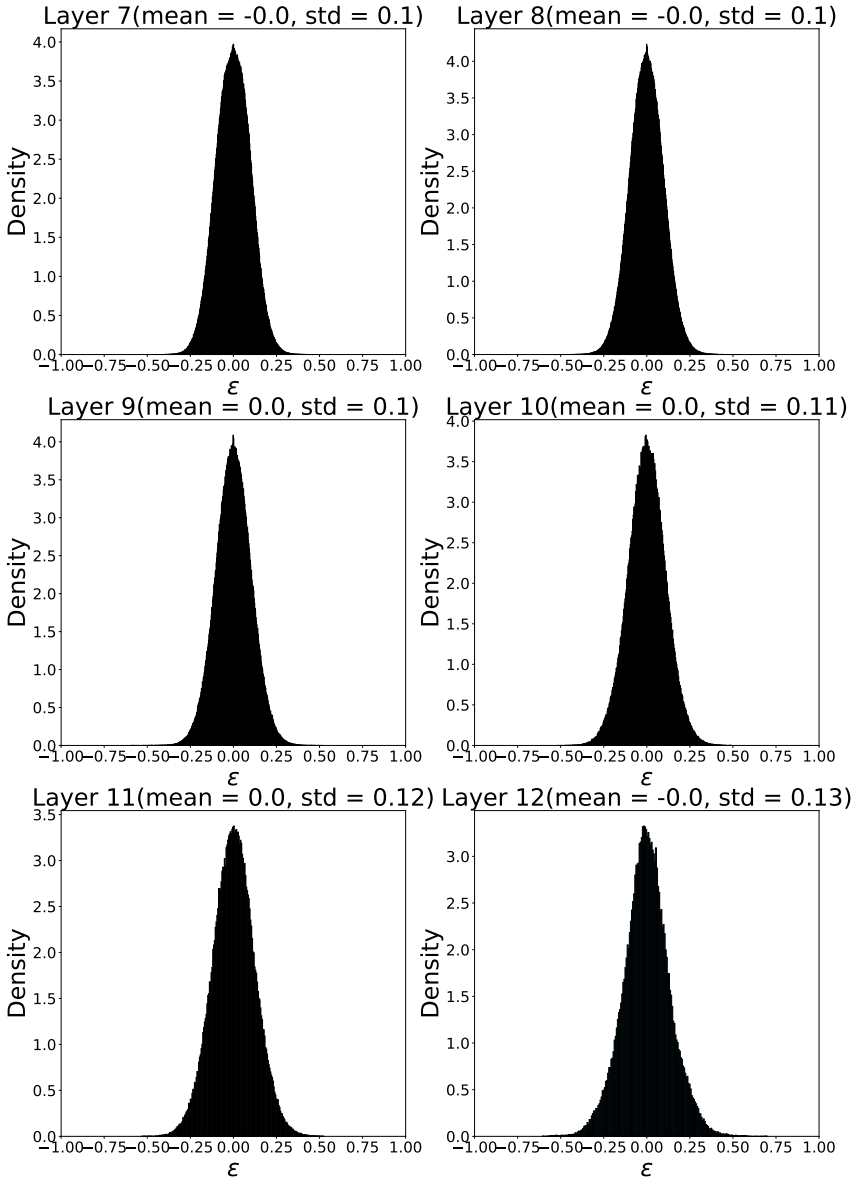

Figure 13: Distribution of parameter ϵ among randomized layers for image taken from FFHQ.

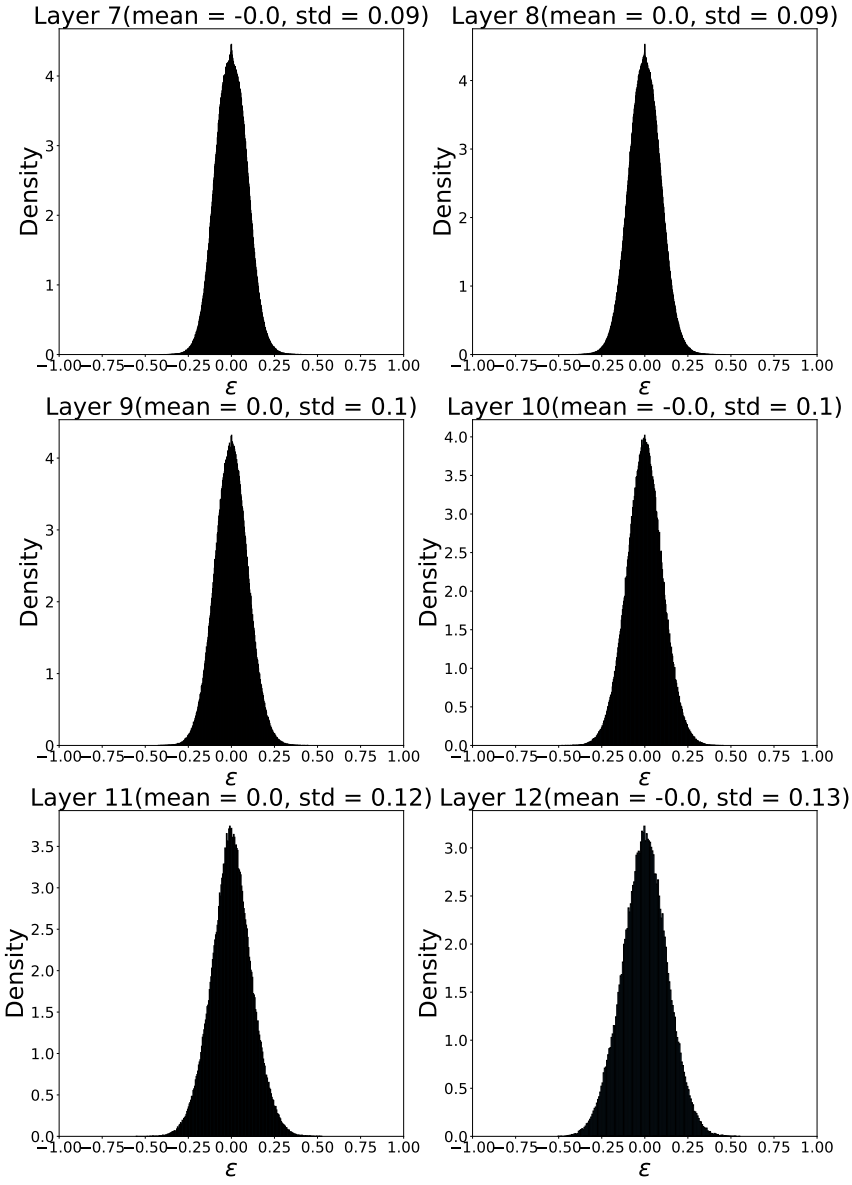

Figure 14: Distribution of parameter ϵ among randomized layers for image taken from FFHQ.

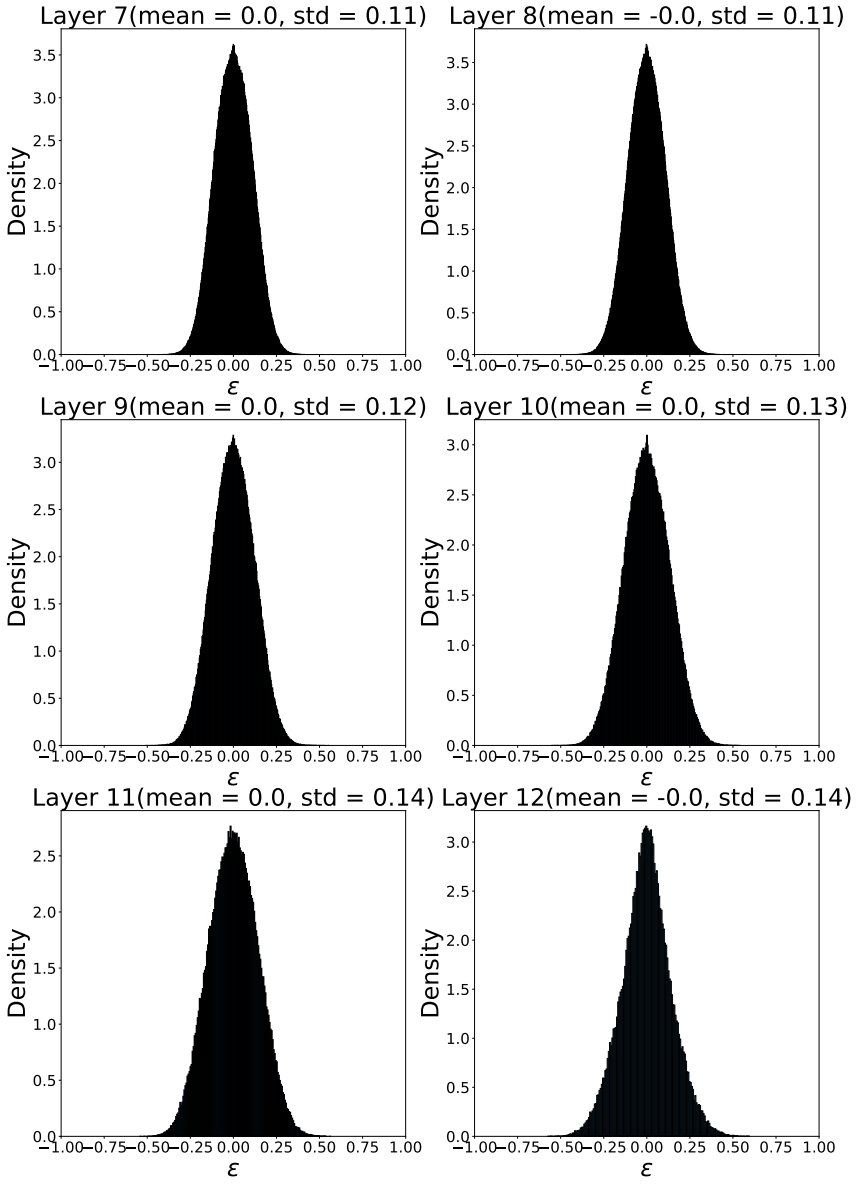

Figure 15: Distribution of parameter ϵ among randomized layers for image taken from FFHQ.

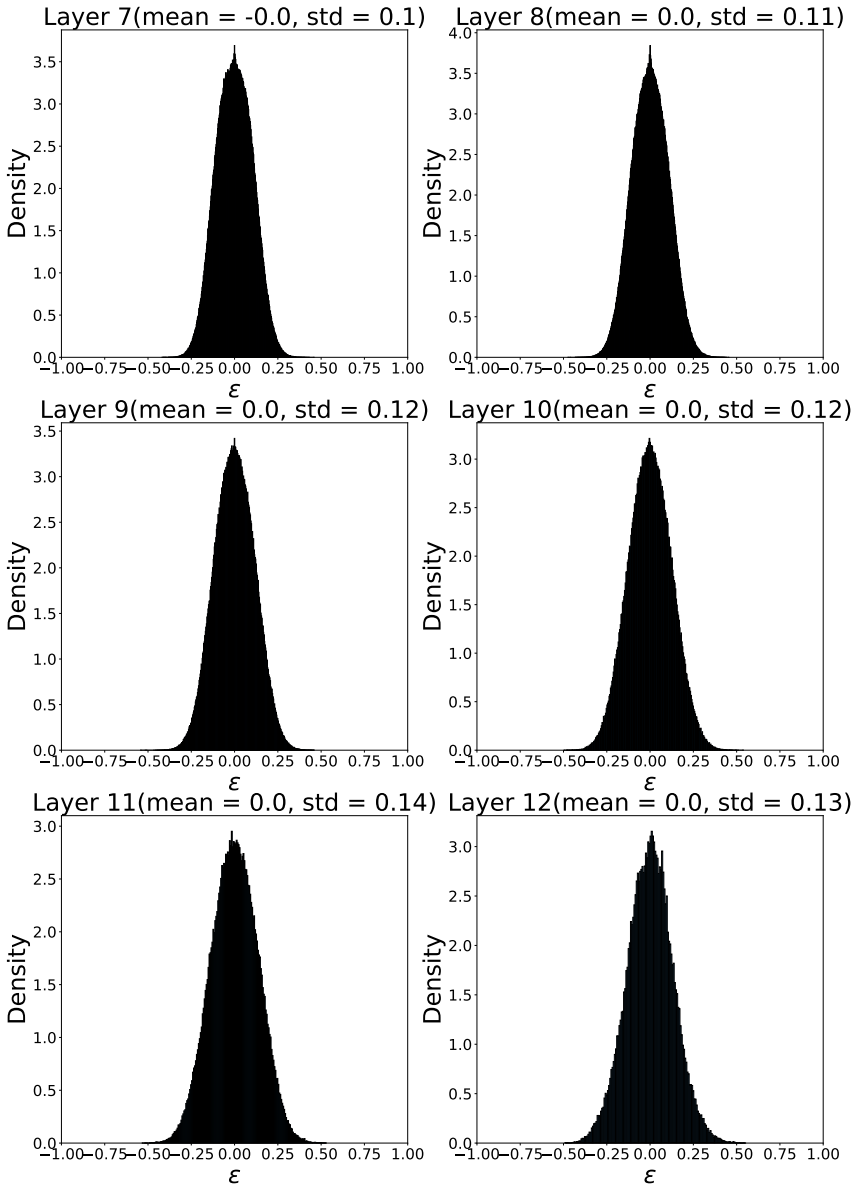

Figure 16: Distribution of parameter ϵ among randomized layers for image taken from FFHQ.

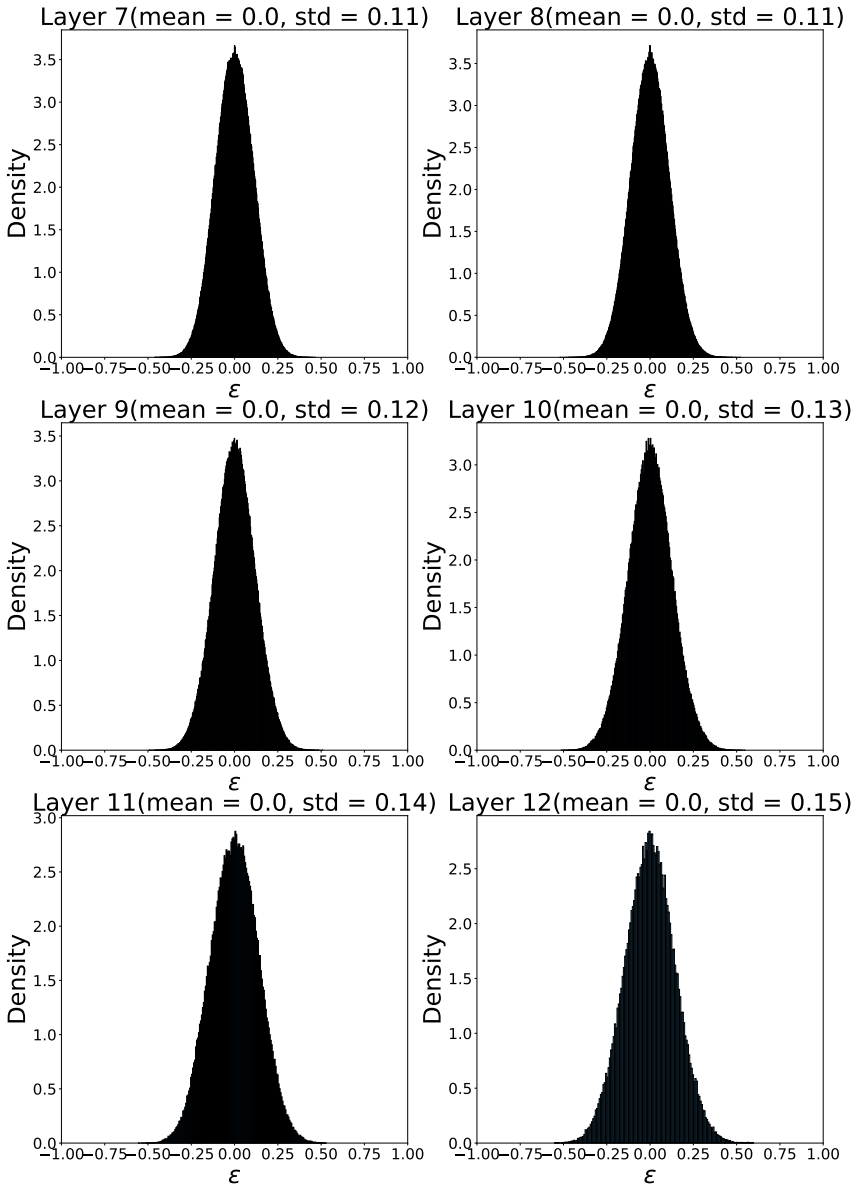

Figure 17: Distribution of parameter ϵ among randomized layers for image taken from FFHQ.

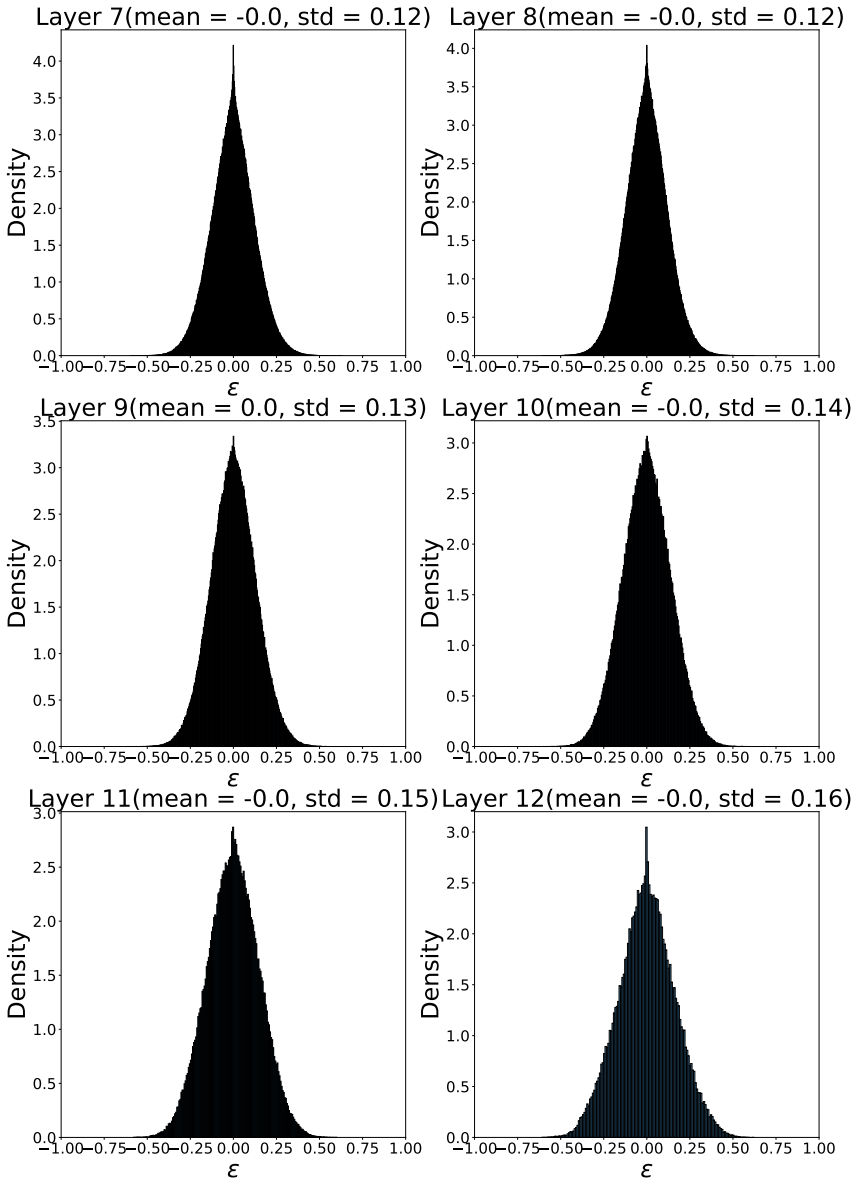

Figure 18: Distribution of parameter ϵ among randomized layers for image taken from LSUN Church.

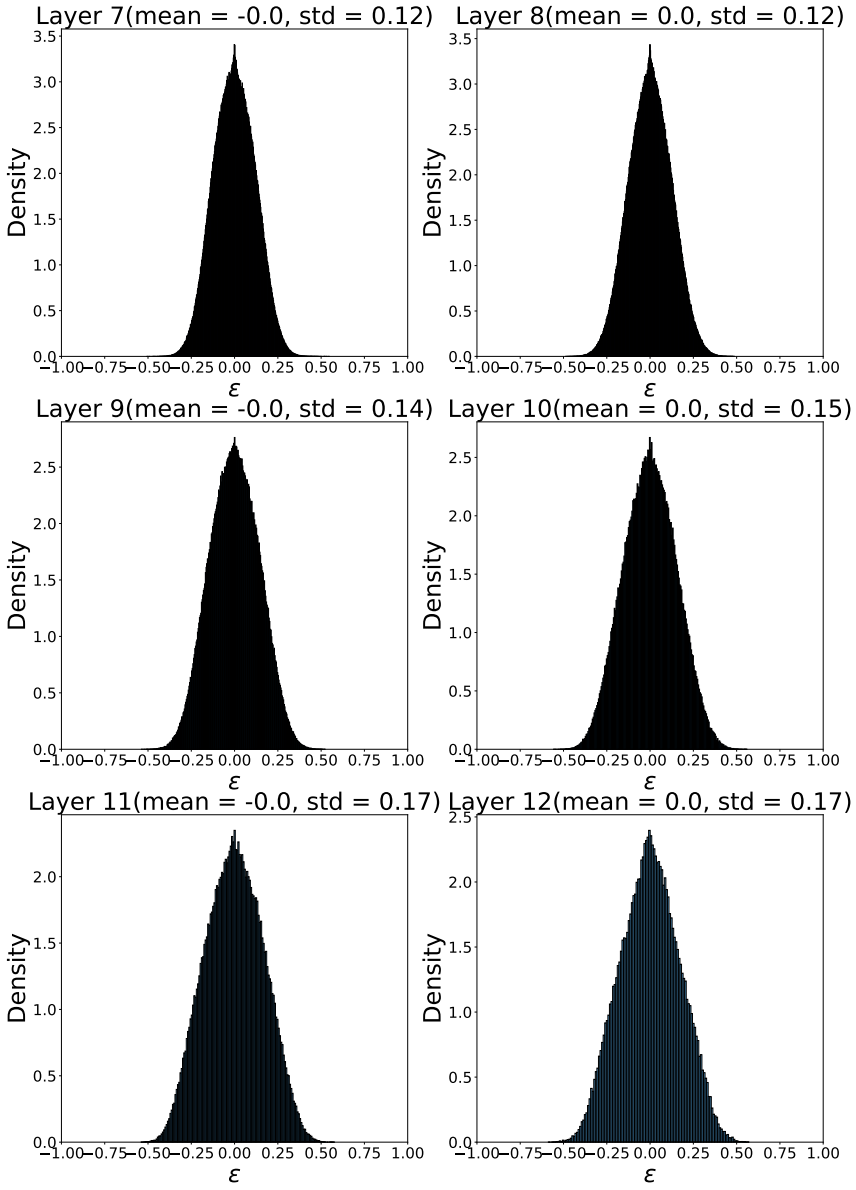

Figure 19: Distribution of parameter ϵ among randomized layers for image taken from LSUN Church.

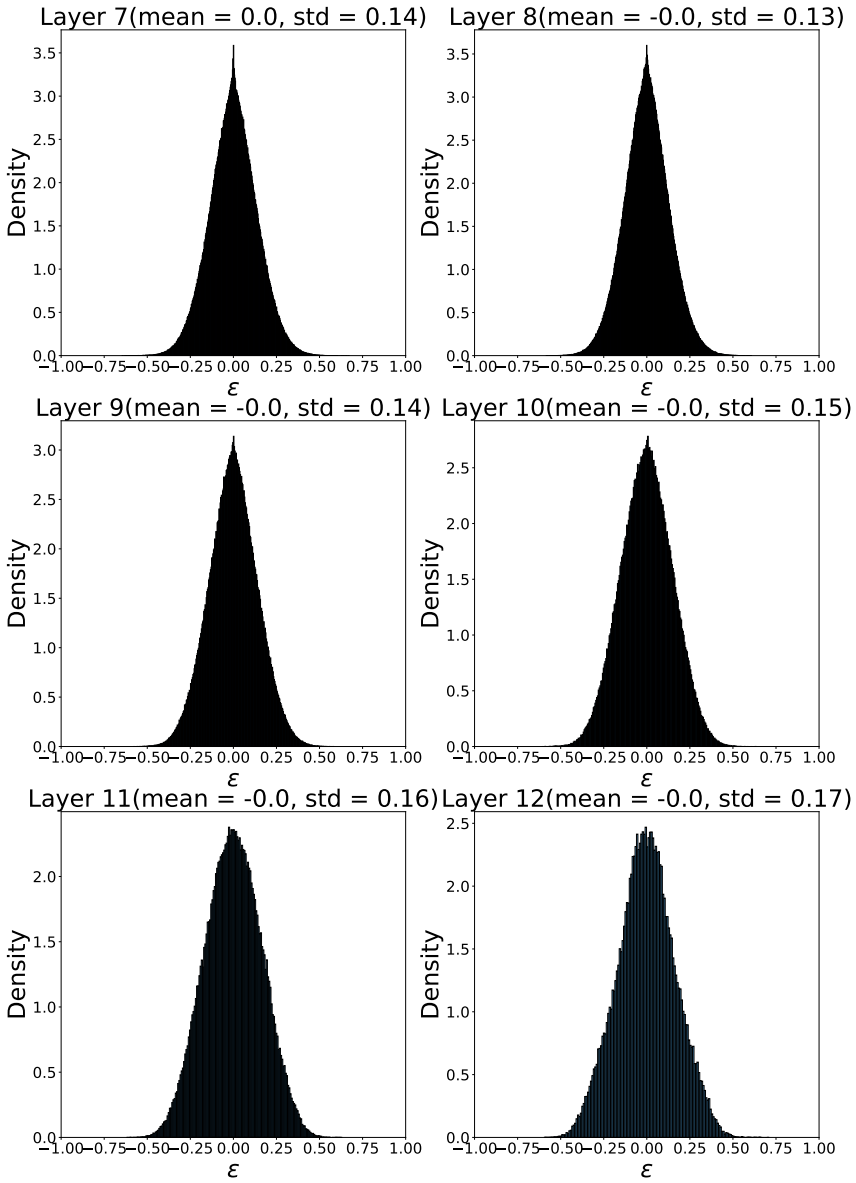

Figure 20: Distribution of parameter ϵ among randomized layers for image taken from LSUN Church.

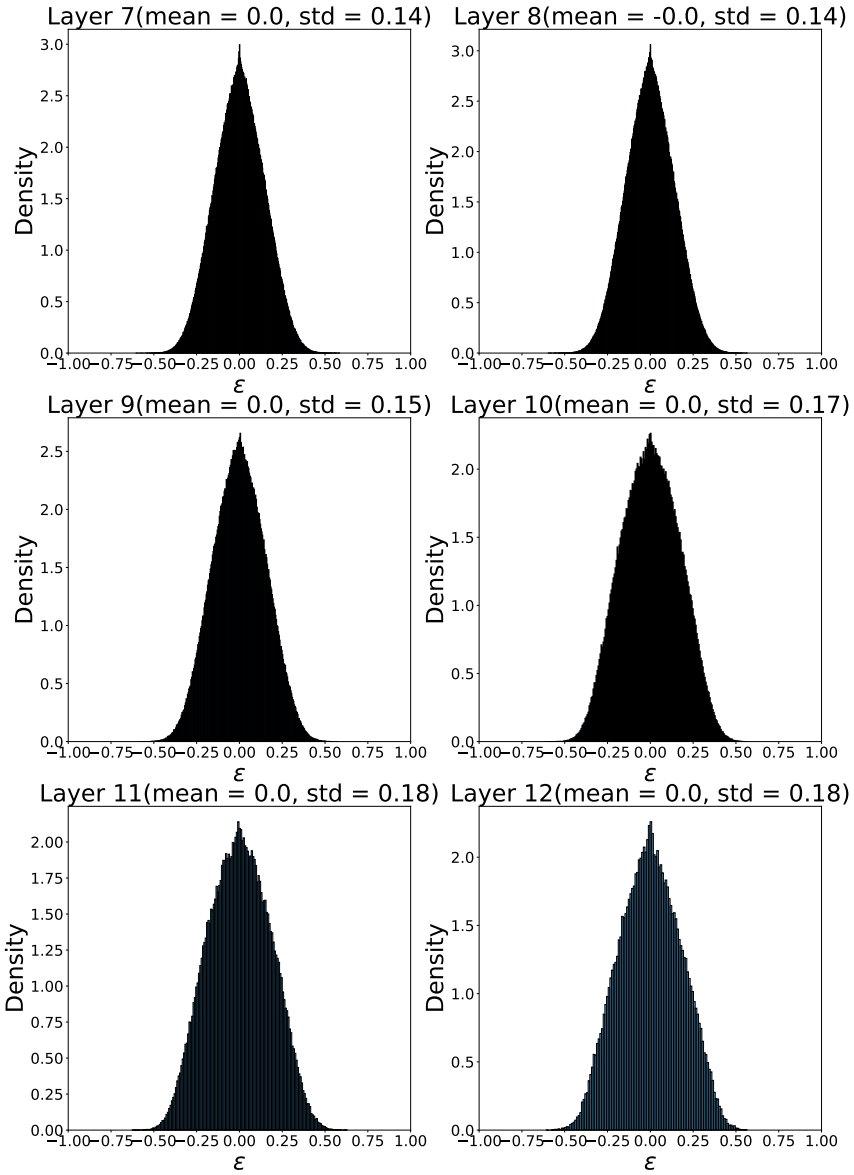

Figure 21: Distribution of parameter ϵ among randomized layers for image taken from LSUN Church.

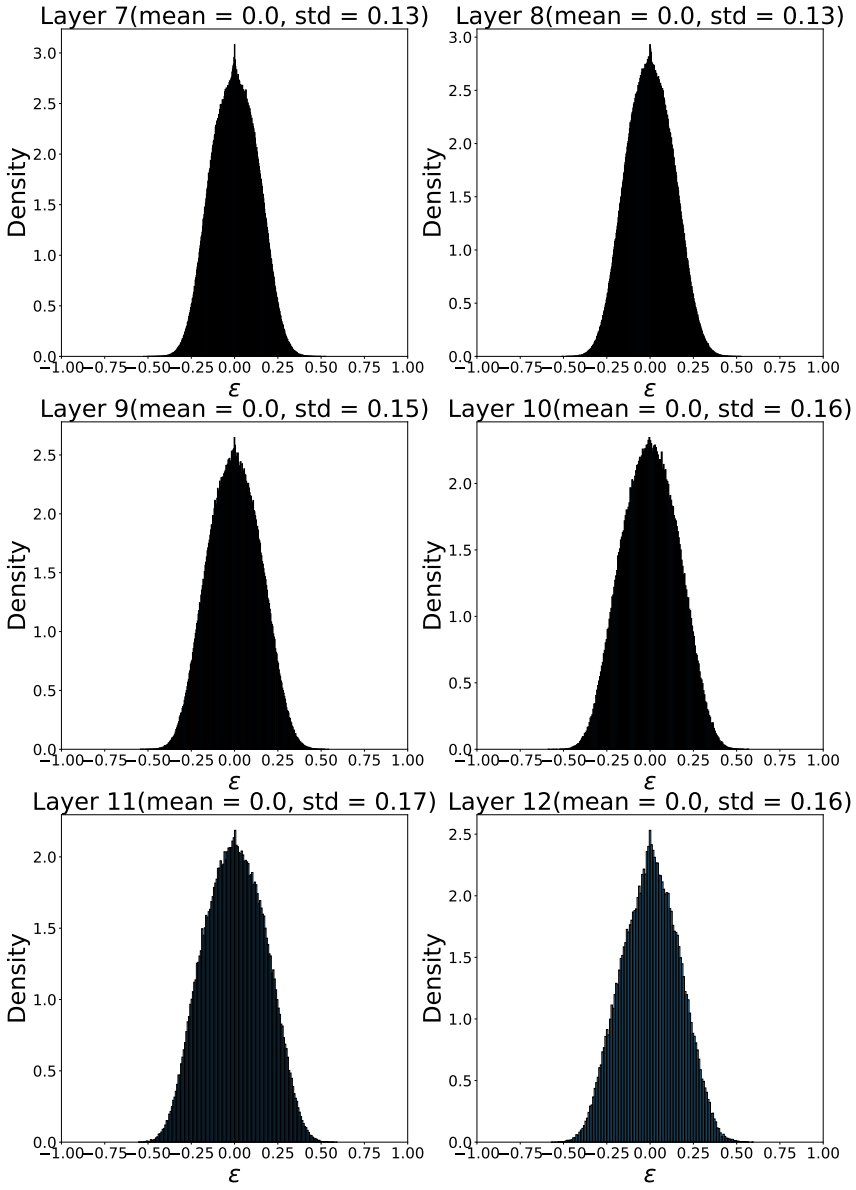

Figure 22: Distribution of parameter ϵ among randomized layers for image taken from LSUN Church.

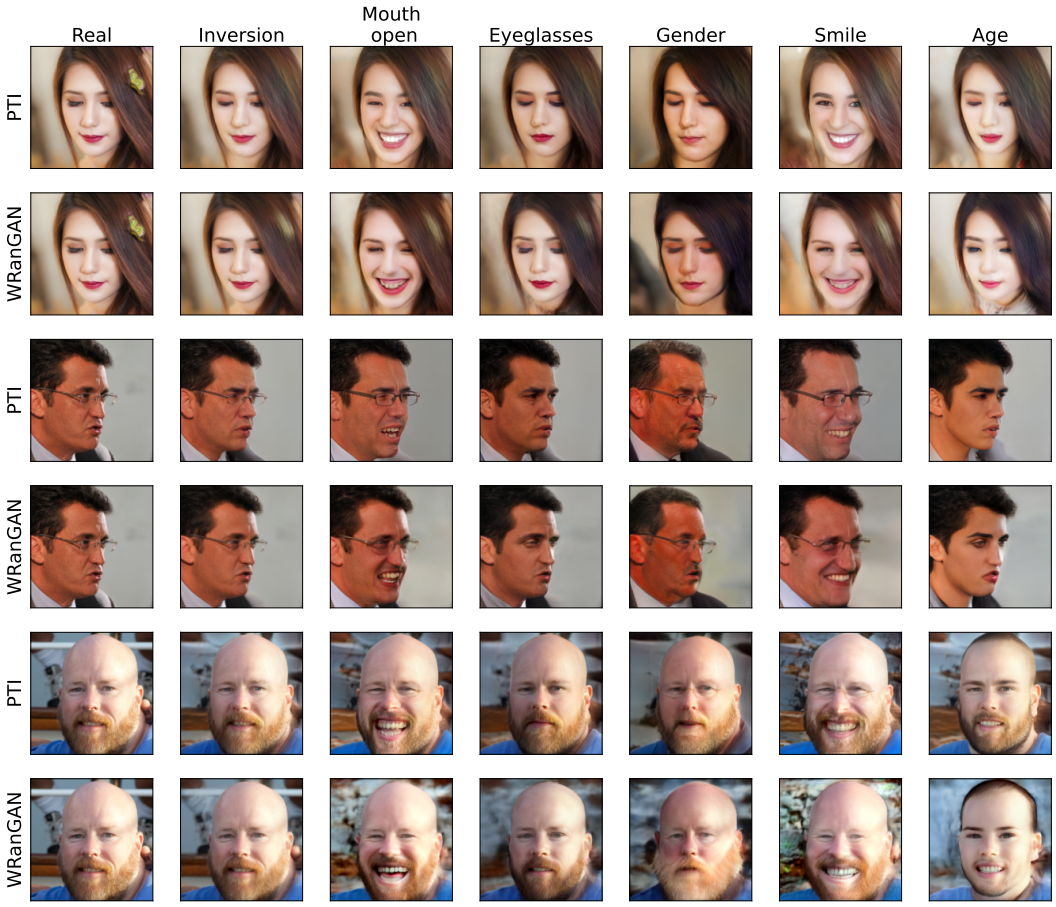

Figure 23: Qualitative editing comparisons for FFHQ dataset

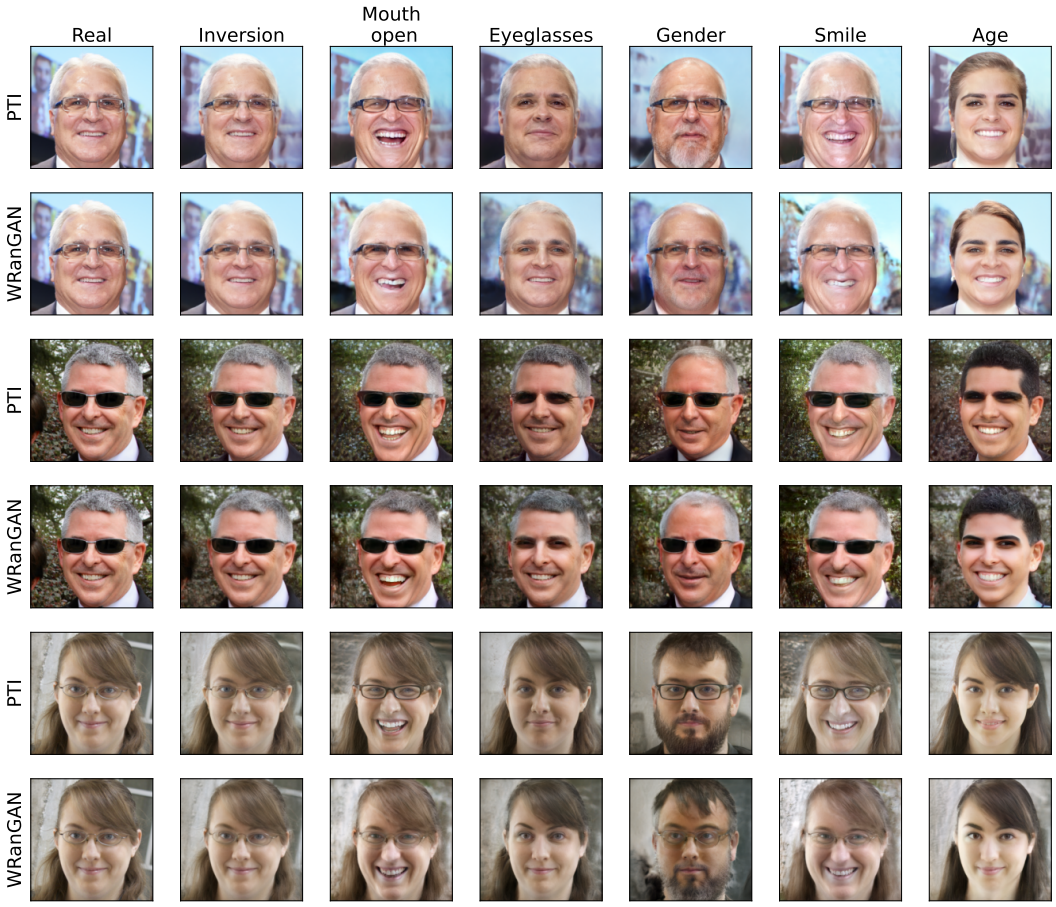

Figure 24: Qualitative editing comparisons for FFHQ dataset

Figure 25: Qualitative editing comparisons for FFHQ dataset

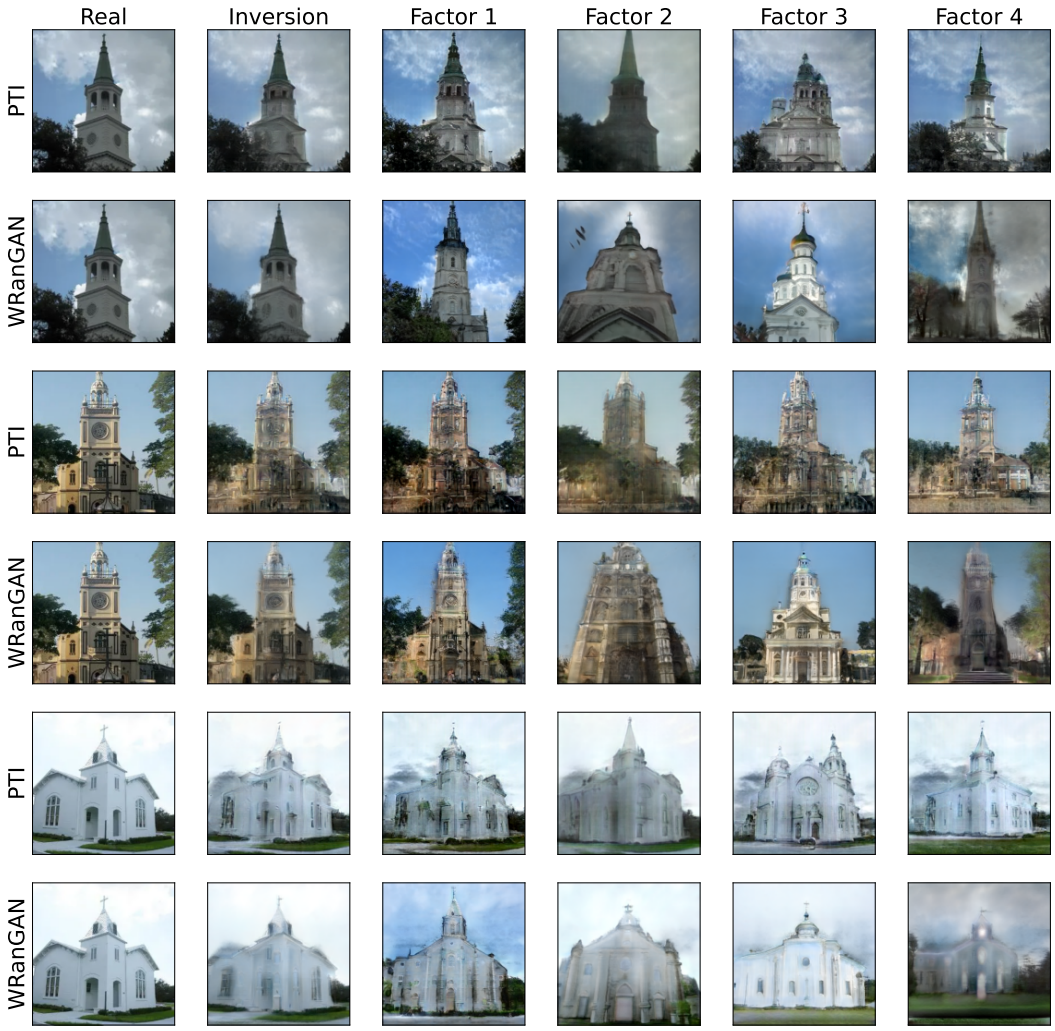

Figure 26: Qualitative editing comparisons for LSUN Church

Figure 27: Qualitative editing comparisons for LSUN Church