# Diverse Inpainting and Editing with GAN Inversion

Ahmet Burak Yildirim*       Hamza Pehlivan*       Bahri Batuhan Bilecen       Aysegul Dundar
Bilkent University
{a.yildirim, hamza.pehlivan, batuhan.bilecen}@bilkent.edu.tr
adundar@cs.bilkent.edu.tr

## Abstract

*Recent inversion methods have shown that real images can be inverted into StyleGAN's latent space and numerous edits can be achieved on those images thanks to the semantically rich feature representations of well-trained GAN models. However, extensive research has also shown that image inversion is challenging due to the trade-off between high-fidelity reconstruction and editability. In this paper, we tackle an even more difficult task, inverting erased images into GAN's latent space for realistic inpaintings and editings. Furthermore, by augmenting inverted latent codes with different latent samples, we achieve diverse inpaintings. Specifically, we propose to learn an encoder and mixing network to combine encoded features from erased images with StyleGAN's mapped features from random samples. To encourage the mixing network to utilize both inputs, we train the networks with generated data via a novel set-up. We also utilize higher-rate features to prevent color inconsistencies between the inpainted and unerased parts. We run extensive experiments and compare our method with state-of-the-art inversion and inpainting methods. Qualitative metrics and visual comparisons show significant improvements.*

## 1. Introduction

Generative Adversarial Networks (GANs) achieve impressive results on unconditional [13, 40, 14, 37] and conditional image generation [31, 9, 18], inpainting [36, 16, 42], and image editing tasks [1, 4, 8]. Traditionally, each of these tasks has been explored with a dedicated network and training pipeline. However, recently it has been shown that high-quality image editing can be achieved with well-trained GAN models and especially by StyleGAN networks [13, 14] that are trained to generate images without a condition [28, 30]. This approach achieves numerous edits via semantically rich feature representations of well-trained
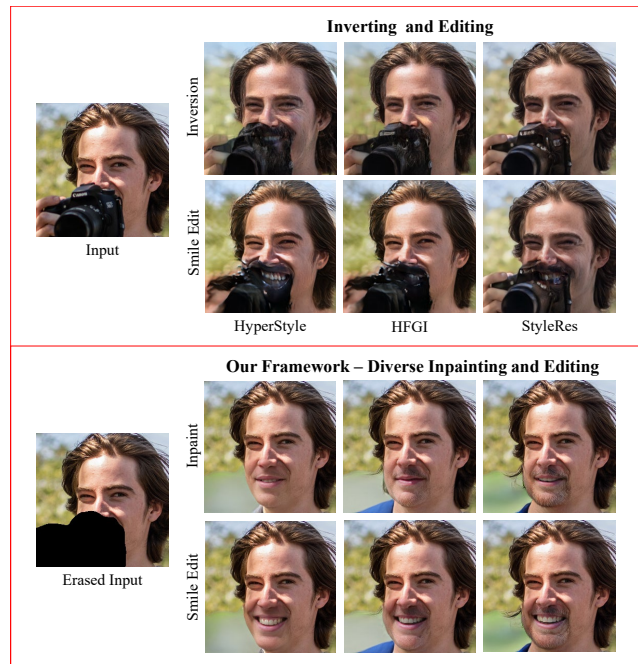


Figure 1. Our framework achieves diverse inpainting and editing with GAN inversion. Compared to other works that are proposed to invert and edit images with StyleGAN (HyperStyle [3], HFGI [30], StyleRes [23]), our framework has advantages, as in the presented example.

GANs [27, 11, 21, 4]. In this work, we are interested in taking this direction one step further and learning an encoder for a pretrained GAN that can achieve high-quality image inversion, diverse inpainting, and editing under one framework.

To benefit from GAN's image editing capabilities, extensive research is conducted on image inversion algorithms to find the corresponding latent codes that will generate a particular real image [7, 25, 43, 28]. It is observed that there exists a trade-off between image reconstruction fidelity and image editing quality [28]. Studies show that low-rate latent spaces ($z$, $W$, or $W^+$ for StyleGANs) are limited in their expressiveness power, and not every image can be in-

---

*Joint first authors, contributed equally.

verted with high-fidelity reconstruction to GAN's natural low-rate latent space [28]. When images are not projected to GAN's natural latent space, even when an image is reconstructed with high fidelity, it will not be editable. Methods are proposed to skip spatially higher resolution features (higher-rate latent codes) from an encoder to a GAN generator for better reconstruction and editing properties [30, 3]. In this work, we are interested in a more challenging scenario; learning an image encoder that can project an erased image to a well-trained GAN's natural latent space. This framework provides new capabilities for image editing, as shown in Fig. 1.

Recently, GAN inversion-based image inpainting methods have been proposed [38, 32]. These methods optimize an image encoder and learn skip connections to the pretrained StyleGAN model. Even though promising results are achieved, these methods model image inpainting in a deterministic way. They are trained to reconstruct the original image from the erased ones without any stochasticity. They are not regularized to project the image into GAN's natural latent space and do not enjoy the editing capabilities of them.

In this work, we propose a framework that achieves high-quality image inversion, diverse inpainting, and editing simultaneously. We design an encoder architecture that takes an erased input image and encodes latent codes for the visible parts. We also design a mixing network that combines randomly sampled latent codes with encoded ones. This setup allows the model to output diverse results. However, we find that the framework learns to ignore the randomly sampled latent codes and has the tendency to output deterministic results. To achieve diversity, we propose to train the framework with a novel design. Our design includes generating images with $W^+$ and erasing a part of it. During training, sometimes the same $W^+$ that generated the input image is fed to the mixing network for the whole input reconstruction, and sometimes a newly sampled $W^+$ is used to reconstruct only the visible part. This way, the network is regularized to use both inputs; erased image and sampled $W^+$. Secondly, to achieve high fidelity inversion of unerased pixels and to prevent the color discrepancy between the unerased and erased parts, we learn higher dimension latent codes. In summary, our main contributions are as follows:

- We propose a novel framework for image inpainting with GAN inversion. Our framework includes an encoder to embed images and a mixing network to combine them with randomly sampled latent codes to achieve diversity. The mixing network has a gating mechanism that improves the results.

- To achieve diversity, we propose a novel set-up to train the networks. We use GAN-generated images and train

the network with full image reconstruction and valid pixel image reconstruction depending on if we feed the same latent code to the mixing network that is used to generate the input image or not.

- We conduct extensive experiments to show the effectiveness of our framework and achieve significant improvements over state-of-the-art models for image inpainting. Additionally, we show our framework can achieve diverse inpainting and editing under one framework, as shown in Fig. 1.

## 2. Related Work

**GAN Inversion and Editing.** Recently, extensive research has been conducted on image editing tasks both via end-to-end trained image translation networks [5, 34, 39, 17, 10, 8] and latent space manipulation of pretrained GANs [7, 25, 43, 28, 30]. Pretrained GANs, especially StyleGAN-based methods [13, 14], are shown to organize their latent space in a semantically meaningful way with disentangled representations. This feature enables them to achieve various edits that are beyond what annotated datasets offer [29, 33, 21, 11, 4]. However, the challenge in using these pretrained GANs for real image editing is that one needs to invert images to the natural latent space of GANs and there is known to be a trade-off between reconstruction fidelity of input images and their editability [43, 24, 28]. Different encoders and training schemes are proposed to tackle the inversion problem with great improvements and the inversion problem is still an active research area [30, 2, 3, 23]. This work aims to solve a more challenging inversion problem with input images that have missing pixels. This task is referred to as image inpainting or image completion and to the best of our knowledge, our work is the first to explore it with pretrained StyleGAN with the high reconstruction and editability goal.

**Inpainting.** Inpainting missing pixels, especially large-scale holes, receives a significant amount of attention from both the research community and industry. This task requires filling in missing pixels in a semantically meaningful and multi-modular way. However, most of the proposed methods are deterministic and struggle with large-scale holes [22, 19, 36, 16, 18]. Recently, CoModGAN [42] is proposed with an architecture and training pipeline similar to StyleGAN but with skip connections from the encoder to the generator to pass the valid pixels to the generator without a bottleneck. CoModGAN achieves diverse results. Diffusion models have also shown great success in diverse image inpainting [20, 26, 35]. Nevertheless, none of these methods can achieve diverse inpainting and enable user-controlled editing of images with one network simultaneously, which is the goal of our work.
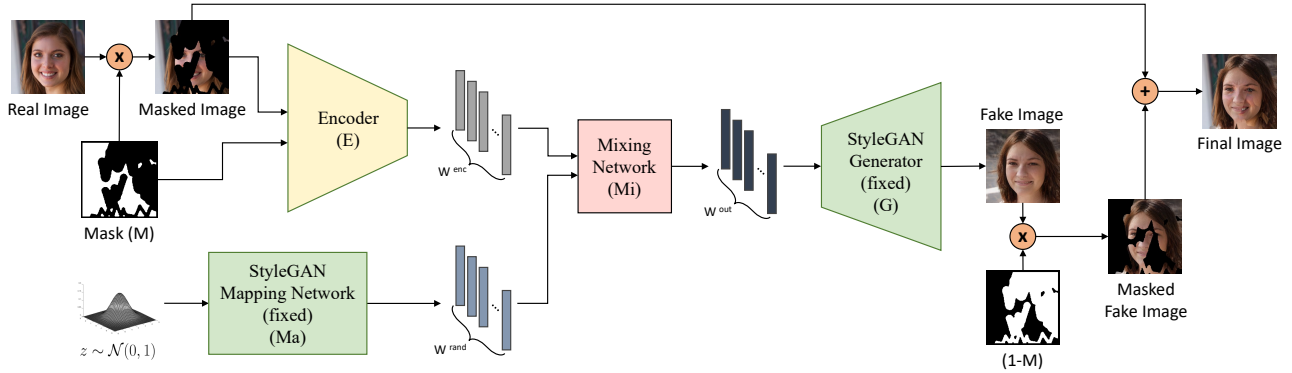
Figure 2. First stage framework includes trainable image encoder and mixing network, and frozen StyleGAN's mapping and generator networks. Our encoder takes an erased image and binary mask to embed the image into StyleGAN's latent space. We also sample a latent code via the mapping network to achieve stochasticity. The mixing network combines the available information of the erased image from the encoder and the missing part from the mapping network. The mixed encoded latent representations are fed to the Generator via the instance normalization layers to output the fake image. There is a final step where the input image and the fake image are combined based on the mask.

## 3. Method

We propose a two-stage training pipeline. In the first stage, we learn a base encoder and mixing network to achieve diverse results as presented in Section 3.1. In the second stage, Section 3.2, we learn skip connections to the generator to achieve high-fidelity reconstructions and seamless transitions between unerased and erased pixels.

### 3.1. First Stage: Diverse Results

Our method utilizes an architecture that includes an image encoder (E), mixing network (Mi), and StyleGAN2's mapping (Ma) and generator (G) networks [14]. We freeze the StyleGAN's mapping and generator networks, as shown in Fig. 2. We obtain erased images by multiplying them with a binary mask M that defines the valid and erased pixels; $I^e = M \odot I$. Before feeding the erased image to the encoder, we concatenate the binary mask M with it. We set a simple encoder [24] to project our erased images. Our encoder embeds erased images into $W^+$ latent space. Our goal is to achieve diversity in our inpainting results. For that, we set a second pathway that utilizes StyleGAN's mapping network to sample random $z$'s and obtains $W^+$ codes that lie in the natural space of StyleGAN. The encoded, $W^{enc}$, and mapped, $W^{rand}$, latent codes are fed to the mixing network. We expect the mixing network to take the information of what is visible from the masked image and combine the missing information with the mapped latent codes. We equip the mixing network with a neural network and gating mechanism as follows:

$$W^{comb}, g = NN(W^{enc}, W^{rand})$$
$$W^{out} = \sigma(g) \cdot W^{comb} + (1 - \sigma(g)) \cdot W^{enc} \quad (1)$$

where $\sigma(.)$ is the sigmoid function. $W^{out}$ goes to the instance normalization layers of StyleGAN generator.

In our experiments, we observe that the mixing network learns to ignore the $W^{rand}$. To achieve diversity, we propose a training pipeline with GAN-generated data as shown in Fig. 3. We first generate an image with $z_g$, and feed it to the mapping network to obtain $W_g^{rand}$. We synthesize the output image and erase a part of the image with a randomly generated mask, $I_g^e = M \odot I_g$. Our encoder encodes the erased image to $W_g^{enc}$. With these encodings, we generate two images, one that combines $W_g^{enc}$ with mapping of $z_g$ and another one with mapping of a randomly sampled different $z_r$ as given in Fig. 3 (a) and (b), respectively. The generator outputs images $I_g^o$ and $I_r^o$ from these two different paths. The generation of $I_g^o$ has access to the full information that generated the image and therefore, if one can combine the encoded embedding with the mapping, one correctly can generate the input image. On the other hand, the second path, $I_r^o$ image has only access to the information of unerased pixels and is not expected to reconstruct the input image faithfully for the erased parts. To achieve this objective, we set the following losses to train the framework.

**Reconstruction Losses.** For reconstructing the pixels correctly, we use $L_2$ and perceptual losses. We use perceptual losses from VGG ($\Phi$) at different feature layers ($j$) between images. We output two images, $I_g^o$ and $I_r^o$ from the input image ($I_g$) that is erased, $I_g^e$. $I_g^o$ has access to the correct mapping as well as unerased image pixels, therefore this training pipeline is expected to generate the original input image as follows:

$$\mathcal{L}_{rg} = ||I_g^o - I_g||_2 + |\Phi_j(I_g^o) - \Phi_j(I_g)||_2 \quad (2)$$
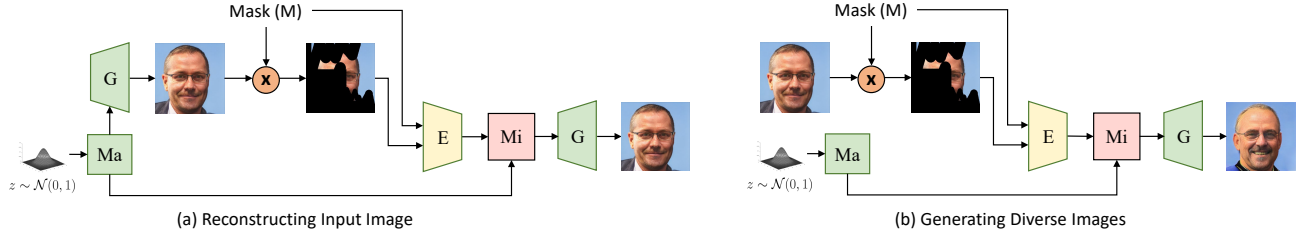
Figure 3. During training, we generate images with the StyleGAN generator. If we feed the mixing network (Mi) with the sampled $z$ that generated the input image, we expect the output to match the input. The reconstruction losses are applied for both valid and invalid pixels. If we sample a new $z$ as in (b), we expect to output diverse results and only apply image reconstruction loss for valid pixels and a GAN loss for the overall image.

On the other hand, $I_r^o$ has only access to the unerased pixels of the original image and is expected to fill the missing information with a randomly sampled latent code. Therefore, we only expect pixel-wise matching on the unerased images. We find the reconstruction losses between the masked output image and erased input image as follows:

$$\mathcal{L}_{rr} = ||(\mathbf{M} \odot I_r^o) - I_g^e||_2 + |\Phi_j(\mathbf{M} \odot I_r^o) - \Phi_j(I_g^e)||_2 \quad (3)$$

If the network is trained with Eq. 2 alone, the mixing network can ignore the input image and encoded features coming from it since full information is provided by the mapping network. Similarly, if the network is trained with only Eq. 3, the mixing network ignores the mapping network. By using these reconstruction losses together, our framework learns to use the information encoded from both paths. Additionally, this training regularizes the encoder and mixing network to output latent codes that are in GAN's natural space which provides editable inversions.

**Adversarial Losses.** We additionally output our final images as follows:

$$I^f = \mathbf{M} \odot I + (1\text{-}\mathbf{M}) \odot I^o \quad (4)$$

This step guarantees that the valid pixels are not modified as given in Fig. 2. We expect these final images to look realistic and use adversarial guidance on the final images for both $I_g^f$ and $I_r^f$. We load the pretrained discriminator from StyleGAN training, $D$, and train the discriminator together with the encoder and mixing network.

$$\begin{aligned} L_{adv} = 2\log D(I_g) + \log\left(1 - D(I_g^f)\right) \\ + \log\left(1 - D(I_r^f)\right) \end{aligned} \quad (5)$$

**Full Objective.** We use the overall objectives given below to optimize the parameters of the Encoder (E) and the mixing network (Mi).

$$\min_{E,Mi} \max_D \lambda_a \mathcal{L}_{adv} + \lambda_{r1} \mathcal{L}_{rg} + \lambda_{r2} \mathcal{L}_{rr} \quad (6)$$

For both of the training stages, we use the same following hyperparameters, $\lambda_a = 8 \times 10^{-2}$, $\lambda_{r1} = 1$, and $\lambda_{r2} = 1$. We set the pixel-wise and VGG reconstruction loss coefficients as 1 and $5 \times 10^{-5}$, respectively for both $L_{rg}$ and $L_{gg}$.

### 3.2. Second Stage: High-Fidelity Reconstruction

Previous works for image inversion show that low-rate latent codes, $W+$, do not have enough capacity for high-fidelity image reconstruction [30, 3]. There is too much information loss due to the bottleneck of projecting high-dimensional images to low-rate latent codes and this bottleneck limits the high-fidelity reconstructions. The same is observed by inpainting with image inversion works [32, 38] as well. The high-fidelity reconstruction of visible pixels is even more crucial in the inpainting task because we combine the erased and generated pixels as given in Eq. 4. When the generator is not able to reconstruct the valid pixels with high fidelity, the color discrepancies become very obvious in the final images. Previous methods [32, 38, 30, 3] suggest encoding information to higher-rate latent codes as well. This is achieved by skip connections from the encoder to the generator on higher-resolution feature maps which are sometimes referred to as $F^+$ space.

The balance of encoding information between $W^+$ and $F^+$ spaces is extensively studied for image inversion methods [30, 3, 23]. That is because by encoding most of the information to $F^+$ space, one can guarantee a high-fidelity reconstruction. However, since the image is not projected to $W+$ properly, it will not be editable which is the main goal of many inversion methods. That is the reason we adopt the two-stage training pipeline similar to state-of-the-art image inversion methods [30, 3, 23]. In the first stage, our method tries to reconstruct and inpaint images only by the encoded low-rate codes, $W^+$. Later, we reduce the information bottleneck by letting higher rate codes skipping from the input image to the generator.
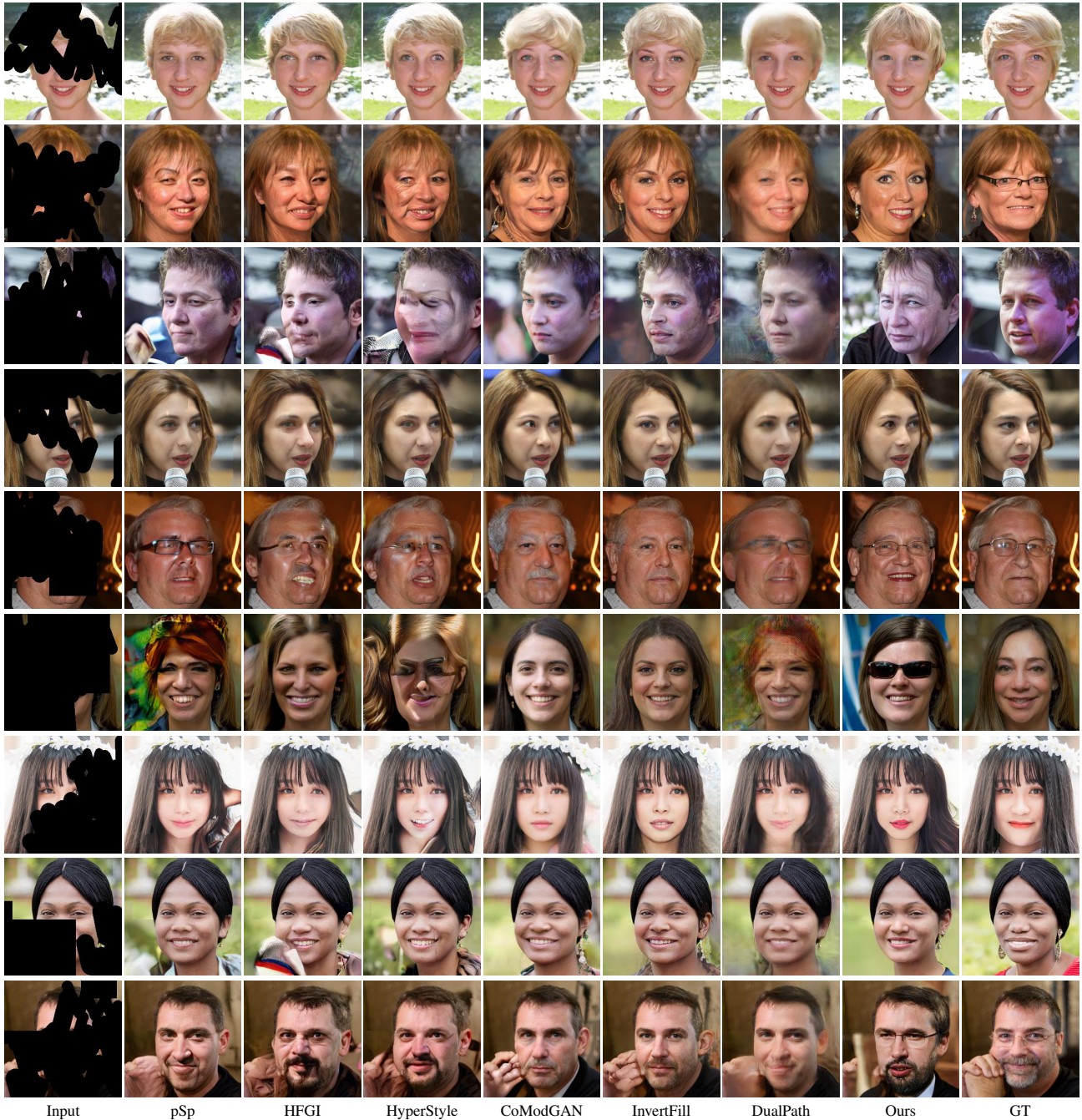
Figure 4. Qualitative results of our and competing methods on the FFHQ dataset. GT refers to original images.

The second-stage training takes the final output of the first-stage model that has gone through the masking as given in Eq. 4. We train an encoder with skip connections to the StyleGAN generator. The encoder also takes the binary mask as input and can detect inconsistencies in the pixel level. The same encoded $W^+$ is used for both generations. However, in the first generation, there are no skip connections ($F^+$), and in the second stage, there are. We also limit the number of skip connection layers and only feed to the feature maps with resolutions of $32 \times 32$, $64 \times 64$, and $128 \times 128$. We fine-tune the models with the same objective from the first stage, given in Eq. 6. We provide the architectural details in Supplementary.

Table 1. Quantitative results of our and competing methods on FFHQ validation dataset for mask ranges between 0-1.

| Models | FID ⇓ | LPIPS ⇑ | U-IDS ⇑ | P-IDS ⇑ |
|---|---|---|---|---|
| e4e [28] | 11.41 | - | 11.96 | 4.84 |
| pSp [24] | 8.23 | - | 14.05 | 6.60 |
| ReStyle [2] | 8.07 | - | 16.46 | 8.61 |
| HFGI [30] | 7.66 | - | 15.63 | 7.76 |
| HyperStyle [3] | 7.46 | - | 16.66 | 8.55 |
| CoModGAN [42] | 7.35 | 0.0790 | 9.78 | 3.79 |
| InvertFill [38] | 7.45 | - | 15.83 | 7.69 |
| DualPath [32] | 17.60 | - | 4.43 | 0.77 |
| Ours | **5.92** | **0.2026** | **18.43** | **10.75** |

## 4. Experiments

**Baselines.** We first compare our method with state-of-the-art image inversion methods pSp [24], e4e [28], HFGI [30], ReStyle [2], and HyperStyle [3]. We use the author's released code and train those models for inpainting tasks with an additional channel in the input for masks. Note that this is a different comparison than the display in Fig. 1 since we train these models for inpainting. pSp model outputs $W^+$ predictions for the generation. HFGI and HyperStyle methods use two-stage training. First, they rely on an encoder to output $W^+$ predictions, then a second encoder takes the input image and StyleGAN generated image with $W^+$ predictions. The goal is to encode the missed information to higher-rate latent codes. We feed erased images and masks to these inversion models and train them to reconstruct the visible pixels faithfully and output realistic images overall.

Next, we experiment with state-of-the-art image inpainting models for our comparisons. We run inferences with CoModGAN's pretrained models [42]. CoModGAN proposes to train a StyleGAN-like model from scratch but with co-modulation and skip connections for the inpainting task. InvertFill [38] and DualPath [32] models are built on pretrained StyleGAN models. InvertFill's authors provide us with their inference results with the masks we provide. We implement DualPath following the set-up provided in their paper. DualPath trains a generator that also takes StyleGAN features.

**Evaluation.** We report Frechet Inception Distance (FID) metric [12], which looks at realism by comparing the target image distribution and reconstructed images. We also look at the diversity scores with Learned Perceptual Image Patch Similarity (LPIPS) [41]. That is for the models that provide diverse images, we generate two images per input image-mask pair and find the distance between these two images. We additionally use U-IDS and P-IDS metrics that are proposed by CoModGAN [42]. They measure unpaired and paired inception discriminative scores, respectively by measuring the linear separability in a pre-trained feature space. Our goal is to output images that are not separable from real
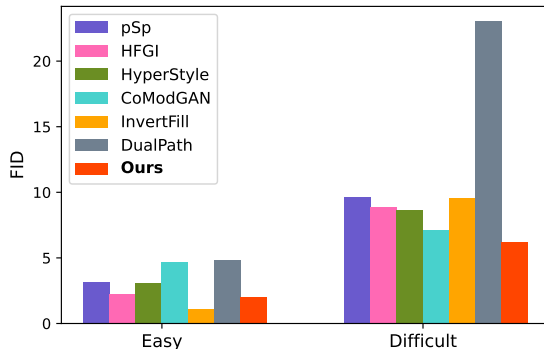


Figure 5. FID plots of ours and competing methods for Easy (0.0-0.4) and Difficult (0.4-1.0) mask ranges on FFHQ validation dataset.

Table 2. Quantitative results of our and competing methods on AFHQ-Dog, AFHQ-Cat, and Stanford Cars validation datasets for mask ranges between 0-1.

| | AFHQ-Dog | AFHQ-Cat | Stanford Cars |
|---|---|---|---|
| Models | FID ⇓ | FID ⇓ | FID ⇓ |
| pSp [24] | 26.46 | 14.26 | 11.77 |
| HFGI [30] | 20.58 | 17.01 | 8.55 |
| HyperStyle [3] | 23.92 | 13.43 | 10.97 |
| Ours | **18.89** | **12.28** | **8.39** |

images. SVM is fitted to the data and we expect a higher error rate.

**Datasets.** We use FFHQ [13], AFHQ-Cat, AFHQ-Dog [6], and Stanford Cars [15] image datasets. We use the StyleGAN2 pre-trained models [14] and use their train and validation splits. We build our extensive comparison with other works on FFHQ human face dataset following previous works. For the other datasets, we train the best single-stage (pSp) and two-stage encoders (HFGI, HyperStyle) for the comparisons.

**Training Details.** We train the first stage for 500k iterations with batch size of 8 on two GPUs. We use learning rate of $1 \times 10^{-4}$ for all networks. We halve the learning rates at each 50k iterations. We use the overall objectives given below to optimize the parameters of the Encoder ($E$) and the mixing network ($Mi$) as was also given in the main paper. We also use the same objective to optimize the parameters of $S$ in the second stage.

**Quantitative Results.** We provide the quantitative results in Table 1 for FFHQ evaluations. We provide the LPIPS score for the models that provide diversity. The inversion methods specifically pSp, HFGI, and HyperStyle, that are trained for inpainting task achieves reasonable scores. Next, we compare with image inpainting methods. CoModGAN trains a network from scratch for inpainting whereas InvertFill uses a pretrained StyleGAN similar to our work. CoModGAN has a StyleGAN-like training set-
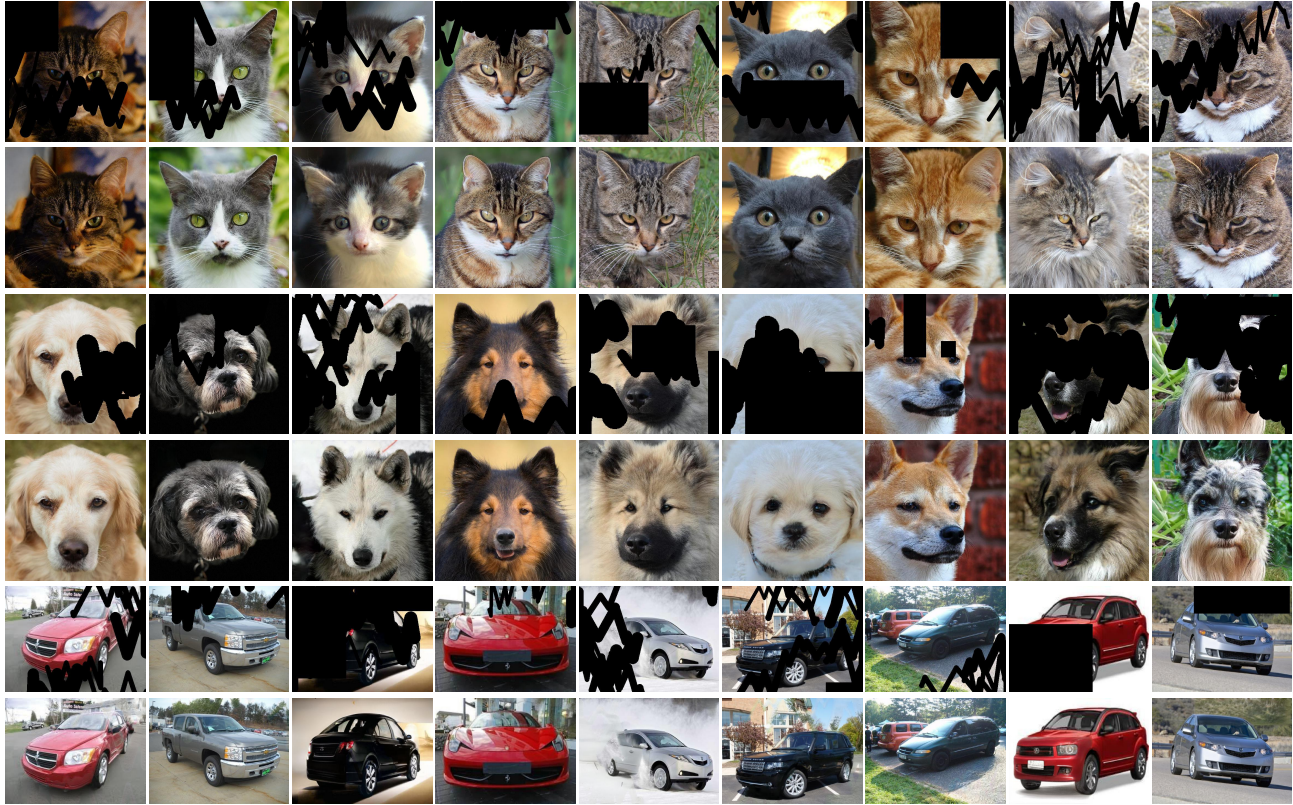
Figure 6. Qualitative results of AFHQ-Cat, AFHQ-Dog, and Stanford Cars dataset trainings.

up, and layers are co-modulated with sampling and encoded image features. CoModGAN achieves diversity but is lower than our model when measured with the LPIPS score. InvertFill, on the other hand, does not have a mechanism to provide any stochastic at all. They achieve similar FID scores but InvertFill achieves significantly better U-IDS and P-IDS scores. We also compare with DualPath which is also built on pretrained StyleGAN. DualPath trains a new generator that takes input from StyleGAN feature maps. We follow the setup provided in their paper and try our best on parameter tuning. However, we do not achieve a good FID score with DualPath. Specifically, the generated images look smooth and semantically consistent, but they do not look sharp. DualPath also does not provide diverse results. We achieve significantly better results than previous methods and also achieve diversity.

We analyze how different methods score with different difficulties of masks in Fig. 5. We set two difficulty levels, an easy one where the erased mask ratio is between (0, 0.4) and a difficult scenario where the erased mask ratio is between (0.4-1.0). When the erased area only contains a small portion, deterministic models also achieve good results since most of the information can be encoded from the unerased parts. InvertFill achieves even a better score than ours and they all achieve better than CoModGAN. How-

ever, when the difficulty level increases, deterministic models struggle. CoModGAN, the other model that can output diverse results achieves the second-best results. InvertFill's FID score significantly increase. Our model achieves good scores in both scenarios.

We also compare with pSp, HFGI, and HyperStyle on AFHQ-Dog, AFHQ-Cat, and Stanford Cars datasets as shown in Table 2. Our method achieves better results than those encoders.

**Qualitative Results.** Fig. 4 shows the results of our and competing methods on the FFHQ dataset. As we mention in the Quantitative Results section, DualPath outputs blurry parts on the inpainted areas in our trainings. The pSp model achieves good semantic consistency. However, the boundaries between the inpainted and original pixels are quite apparent. HFGI and HyperStyle output artifacts on many of the examples. CoModGAN and InvertFill achieve plausible predictions. An interesting observation we make is that all other methods output similar identities for given inputs. For example, the last two row predictions of CoModGAN and InvertFill are similar to each other. On the other hand, since our model samples diverse predictions (more diverse than CoModGAN), our results look quite different. Our results better capture the data distribution as is also measured by FID scores. We also present the inpainting results of our

Figure 7. Qualitative results of the ablation study. Please refer to Table 3 for the training-setting of model IDs.

Table 3. Quantitative results of our ablation study on FFHQ validation dataset for the mask of range 0-1. We assign model IDs to easily refer to the model in the text and Fig. 7.

| ID | Full Recons | Gated Mi | Second Stage | FID | LPIPS |
|----|-------------|----------|--------------|-------|--------|
| 1  |             | ✓        |              | 10.92 | 0.0073 |
| 2  |             | ✓        | ✓            | 7.87  | 0.0855 |
| 3  | ✓           |          |              | 20.91 | 0.2095 |
| 4  | ✓           | ✓        |              | 16.65 | 0.1440 |
| 5  | ✓           | ✓        | ✓            | 5.92  | 0.2026 |

method on AFHQ cat and dog images and Stanford car images in Fig. 6. Our method achieves successful inpaints on these image domains as well.

**Ablation Study.** We present the quantitative and qualitative results of our ablation study in Table 3 and Fig. 7, respectively. During the development of our model, we track the diversity (LPIPS) and quality (FID) of our results. LPIPS score alone does not reflect the quality of our results because random images can be generated by StyleGAN that are semantically inconsistent with valid pixels of the input image. High LPIPS scores can be obtained from those images. In fact, during training, the LPIPS score starts from a high point and decreases as the framework learns to output consistent pixels with the unerased parts. Therefore, it is important to achieve a reasonable FID and LPIPS together.

We assign IDs to each model in Table 3 to easily refer to them in the text and figures. Model 1 is trained with real images to reconstruct the unerased parts together with adversarial loss. The objective does not have the full-reconstruction image loss based on the sampled $W+$. The model does not achieve any diversity even though the network architecture takes sampled $W+$ via the mixing network. The mixing network ignores the stochastic input as can be seen from Fig. 7. In Model 2, we also add the second stage training which improves the FID results because it removes the color inconsistencies. The diversity is negligible as can be seen from both Fig. 7 and Table 3. Next, we compare methods that are trained with our proposed losses and training pipeline. Model 3 does not have the gated mixing network, instead, it employs a fully connected layer to combine encoded and sampled latent codes. Model 4 has the gated mixing network as given in Eq. 1.

In our experiments, we find that the proposed mixing network does a better job of combining features. Model 3 sometimes generates inconsistent results as can be seen in Fig. 7 and achieves worse FID scores. It has a higher LPIPS score but this diversity result in semantic inconsistencies. As mentioned before, StyleGAN model that randomly outputs images will achieve the best diversity, but will not output consistent pixels with the erased part. For example, the input image may have half-erased eyeglasses and this network may not output eyeglasses for the visible parts. The gated mixing network achieves better in those scenarios. Finally, we add the second stage of training and train skip connections to obtain our final model. The color discrepancies disappear after this stage of training. Interestingly, our not-diverse model - Model 1 achieves better scores than our diverse model - Model 4. However, when second-stage training is added, the diverse model achieves better results. That is because the diverse model in the first stage has more difficulties outputting perfectly semantically consistent pixels but it outputs meaningful and diverse parts. Especially when the erased parts are large, the diverse model does a better job as was discussed for Fig. 5 because it augments the encoded features with sampled ones. The color discrepancies are resolved with the second stage of training bringing the diverse model to a better score than the non-diverse model.

**Editing Results.** For our editing results, we first encode the erased image and randomly sample a $W+$ to our mixing network. We edit the output of the mixing network with InterfaceGAN directions [27]. Editing results are shown in Fig. 8. First, we show the erased input. Output rows show the randomly sampled $W+$ results and the next rows show after adding InterfaceGAN directions for different attributes. We show smile, bangs, and beard addition results as well as image edits with blush eyebrows, blonde hair, and lipstick addition. Our model can successfully both inpaint and edit erased images thanks to the semantically rich feature representations of the StyleGAN model bringing more capabilities under one framework.
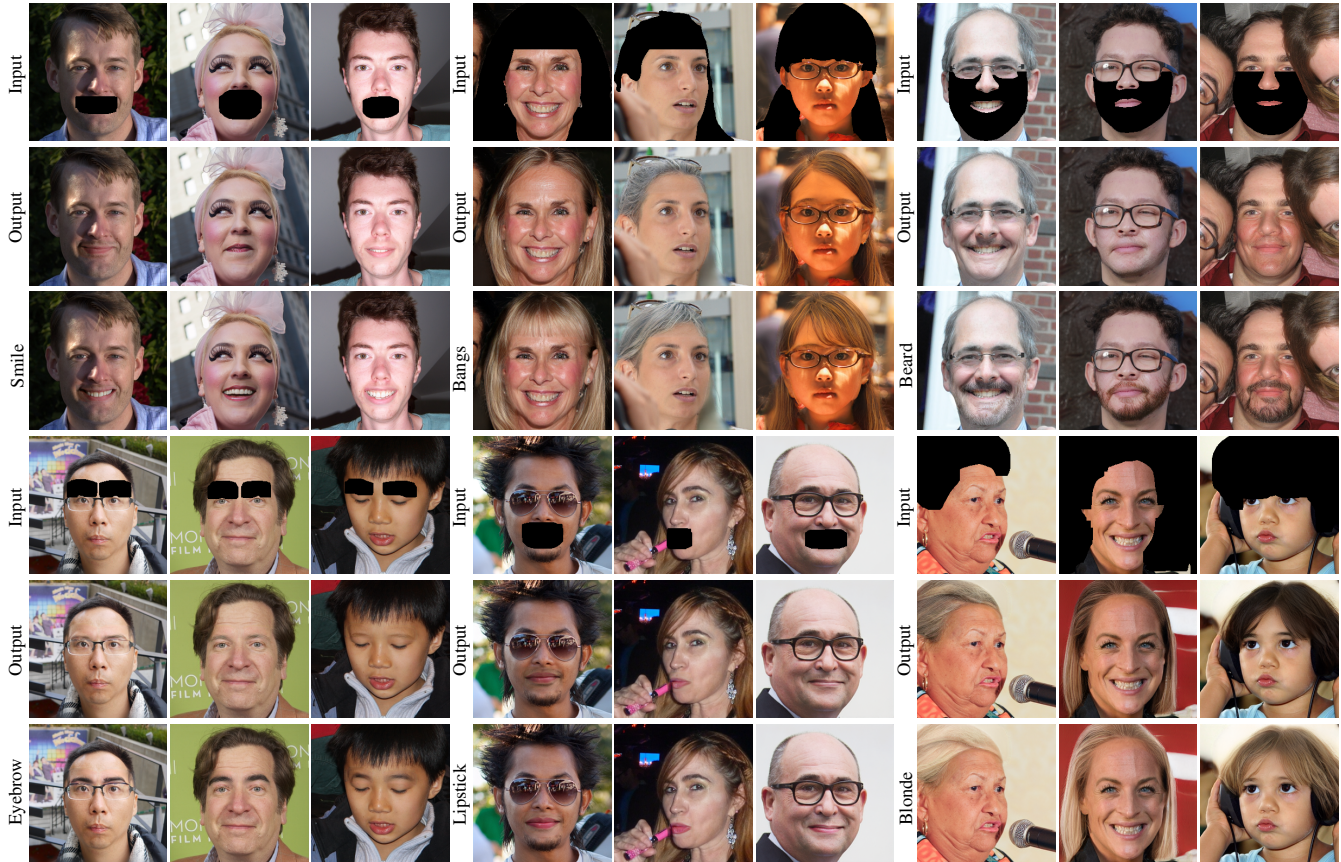
Figure 8. Editing results of our method. Our method achieves diverse inpainting and editing under one framework. First rows show the erased inputs. Output rows show the randomly sampled $W+$ results and the next rows show after adding InterfaceGAN directions for different attributes.

## 5. Conclusion

In this work, we present a novel framework that can achieve diverse image inpainting and editing with GAN inversion. To the best of our knowledge, our framework is the first one that can achieve those simultaneously. We present an extensive ablation study to show the effectiveness of our contributions. Our comparisons with competing methods show that we achieve significantly better scores than previous works both in terms of diversity and quality.

## Acknowledgement

## References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 1

[2] 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*, pages 6711–6720, 2021. 2, 6

[3] 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*, pages 18511–18521, 2022. 1, 2, 4, 6

[4] Zikun Chen, Ruowei Jiang, Brendan Duke, Han Zhao, and Parham Aarabi. Exploring gradient-based multi-directional controls in gans. In *European Conference on Computer Vision*, pages 104–119. Springer, 2022. 1, 2

[5] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[6] 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*, pages 8188–8197, 2020. 6

[7] 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. 1, 2

[8] Yusuf Dalva, Said Fahri Altındiş, and Aysegul Dundar. Vecgan: Image-to-image translation with interpretable latent directions. In *European Conference on Computer Vision*, pages 153–169. Springer, 2022. 1, 2

[9] Aysegul Dundar, Karan Sapra, Guilin Liu, Andrew Tao, and Bryan Catanzaro. Panoptic-based image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8070–8079, 2020. 1

[10] Yue Gao, Fangyun Wei, Jianmin Bao, Shuyang Gu, Dong Chen, Fang Wen, and Zhouhui Lian. High-fidelity and arbitrary face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16115–16124, 2021. 2

[11] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2

[12] 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. 6

[13] 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*, pages 4401–4410, 2019. 1, 2, 6

[14] 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*, pages 8110–8119, 2020. 1, 2, 3, 6

[15] 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*, pages 554–561, 2013. 6

[16] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7760–7768, 2020. 1, 2

[17] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8639–8648, 2021. 2

[18] Guilin Liu, Aysegul Dundar, Kevin J Shih, Ting-Chun Wang, Fitsum A Reda, Karan Sapra, Zhiding Yu, Xiaodong Yang, Andrew Tao, and Bryan Catanzaro. Partial convolution for padding, inpainting, and image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2

[19] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 2

[20] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[21] 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*, pages 2085–2094, 2021. 1, 2

[22] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[23] Hamza Pehlivan, Yusuf Dalva, and Aysegul Dundar. Styleres: Transforming the residuals for real image editing with stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2023. 1, 2, 4

[24] 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*, pages 2287–2296, 2021. 2, 3, 6

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

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[27] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 1, 8

[28] 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. 1, 2, 6

[29] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020. 2

[30] 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*, pages 11379–11388, 2022. 1, 2, 4, 6

[31] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image syn-

thesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1

[32] Wentao Wang, Li Niu, Jianfu Zhang, Xue Yang, and Liqing Zhang. Dual-path image inpainting with auxiliary gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11421–11430, 2022. 2, 4, 6

[33] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 2

[34] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018. 2

[35] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023. 2

[36] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 1, 2

[37] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6731–6742, 2021. 1

[38] Yongsheng Yu, Libo Zhang, Heng Fan, and Tiejian Luo. High-fidelity image inpainting with gan inversion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 242–258. Springer, 2022. 2, 4, 6

[39] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–432, 2018. 2

[40] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 1

[41] 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*, pages 586–595, 2018. 6

[42] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 1, 2, 6

[43] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020. 1, 2