

# Out-of-domain GAN inversion via Invertibility Decomposition for Photo-Realistic Human Face Manipulation

Xin Yang<sup>1,2,3</sup> Xiaogang Xu<sup>4,5</sup> Yingcong Chen<sup>1,2,3\*</sup>

<sup>1</sup> HKUST (Guangzhou) <sup>2</sup> HKUST

<sup>3</sup> HKUST (Guangzhou) - SmartMore Joint Lab <sup>4</sup> CUHK <sup>5</sup> SmartMore

xin.yang@connect.ust.hk xgxu@cse.cuhk.edu.hk yingcongchen@ust.hk

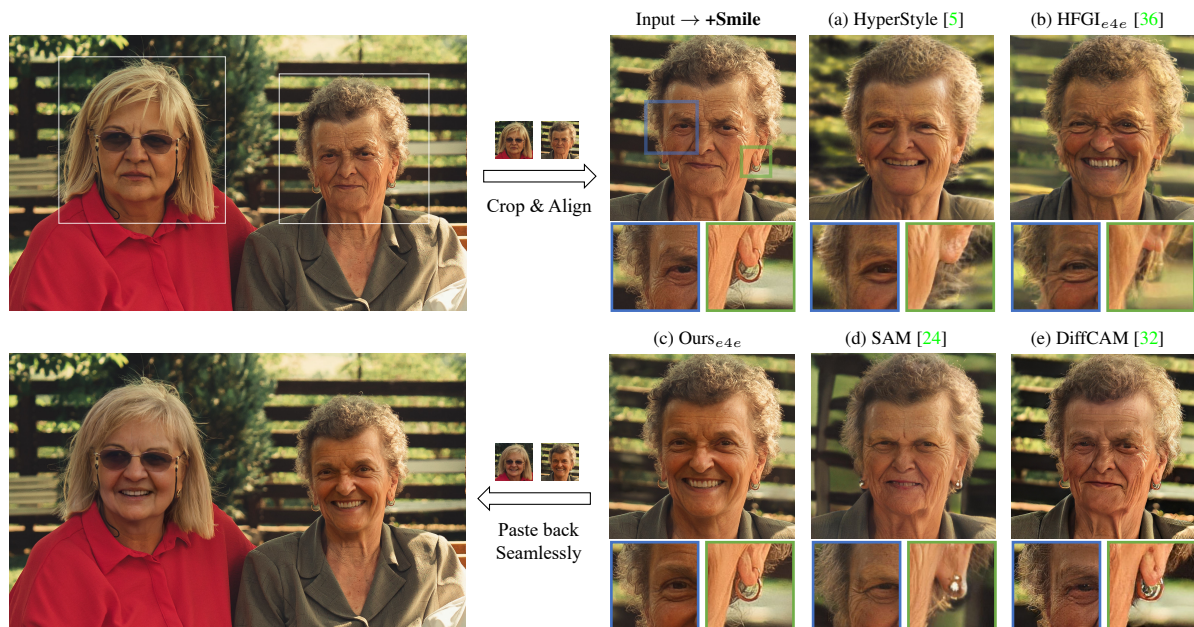


Figure 1: **Out-of-domain GAN inversion and attribute manipulation.** Our method maintains out-of-domain objects in the input image (e.g., background, earrings) and generates high-fidelity attribute-edited results. Please zoom in for detail.

## Abstract

The fidelity of Generative Adversarial Networks (GAN) inversion is impeded by Out-Of-Domain (OOD) areas (e.g., background, accessories) in the image. Detecting the OOD areas beyond the generation ability of the pre-trained model and blending these regions with the input image can enhance fidelity. The “invertibility mask” figures out these OOD areas, and existing methods predict the mask with the reconstruction error. However, the estimated mask is usually inaccurate due to the influence of the reconstruction error in the In-Domain (ID) area. In this paper, we propose a novel framework that enhances the fidelity of human face inversion by designing a new module to decompose the input images to ID and OOD partitions with invertibility masks. Unlike previous works, our invertibility detector is simultaneously learned with a spatial alignment module. We it-

eratively align the generated features to the input geometry and reduce the reconstruction error in the ID regions. Thus, the OOD areas are more distinguishable and can be precisely predicted. Then, we improve the fidelity of our results by blending the OOD areas from the input image with the ID GAN inversion results. Our method produces photo-realistic results for real-world human face image inversion and manipulation. Extensive experiments demonstrate our method’s superiority over existing methods in the quality of GAN inversion and attribute manipulation.

## 1. Introduction

In recent years, there have been efforts to apply generative models, e.g., StyleGAN2 [20], for face image editing [1, 2, 40, 32, 5, 36, 24, 35, 4, 27, 42, 28, 23, 39] and

restoration [37, 14, 44, 41]. The basis of these applications is the GAN inversion. The typical inversion strategy [27, 35] is to train encoders to encode the face images into the latent of the pre-trained generator and reconstruct the images via the generator. However, such approaches can not result in artifact-free and precise reconstruction due to the inevitable information loss when translating the high-resolution image into the limited GAN latent space.

Some improved methods were proposed to further fine-tune the generator for image-specific reconstruction without losing editability [28, 23]. Although these finetuning-based techniques improve the inversion accuracy of identity and style, still, the generator can hardly handle the reconstruction of out-of-domain contents, e.g., the complex background, accessories, and hair. Meanwhile, some works [36, 37, 24] strengthen the potential of out-of-domain inversion capability of pre-trained GAN models by modulating the generator features with the input features extracted from source images. Such methods suffer from the fidelity-editability trade-off [36, 29, 34, 14, 37, 41] since the feature modulation operation breaks the GAN priors. The larger latent space increases the reconstruction quality but undermines the editability of the framework.

Furthermore, some recent works [32, 24] propose to disentangle the target image into different spatial areas. They refine the regions with low invertibility to improve the inversion fidelity. Such low invertibility regions are the parts that cannot be reconstructed well with the generator, so-called Out-Of-Domain (OOD) areas. E.g., Song et al. [32] estimate a manipulation-aware mask with an attribute classifier. However, they ignore the geometrical misalignment between the inverted and the original images and apply a dehazing module for result refinement, which leads to undesired artifacts in their results (Fig. 1 (e)). Parmar et al. [24] train an invertibility mask prediction module with perceptual supervision. The invertibility masks are then used as guidance for GAN feature modulation. However, their prediction of invertibility is noisy and inconsistent with the face semantic parts. Therefore, they adopt a segmentation model for refinement, but the invertibility in the same semantic area (e.g., occlusions on the face) could be inconsistent. Also, they manually design thresholds to filter out the OOD areas, which is hard to optimize for small objects (Fig. 1 (d)).

In summary, existing invertibility estimation methods mainly adopt the reconstruction error as the reference to judge the OOD regions. However, they ignore that reconstruction errors also come from the In-Domain (ID) areas. Consequently, their predicted mask is noisy and unreliable.

In this paper, we propose a novel strategy for photo-realistic GAN inversion by decomposing the input images into OOD and ID areas with invertibility masks. We focus on the high-resolution ( $1024^2$  pixels) GAN inversion on

human face images and the downstream applications (e.g., attribute editing). Our basic idea is to reduce the reconstruction error of the ID areas and thus highlight the error of OOD regions. The reconstruction error of the ID area comes from both the textural and geometrical misalignment between the input image and the generated image. Although previous works improve the textural accuracy in the reconstruction by predicting or optimizing a better latent vector  $w$ , the geometrical misalignment is rarely discussed, which we believe is also important for invertibility estimation. Hence, we design an invertibility detector learned with an optical flow prediction module to reduce the influence of geometrical misalignment. The optical flow is computed between the features of the encoder and the generator, which is then applied to warp the generated features to alleviate their misalignment with the input features. Compared with feature modulation [36, 24, 42], such warping will not break the fidelity of the generated textures. Along the training, the reconstruction error of the ID area will be minimized, and the invertibility mask prediction will be gradually focused on the OOD regions. The overall procedure needs no extra labels for the mask or flows.

Based on the invertibility prediction, we design an effective approach to composite the generated content with the out-of-domain input feature for a photo-realistic generation with high fidelity. Our framework consists of three major parts: the encoder, the Spatial Alignment and Masking Module (SAMM), and the generator. First, we extract features from the input image and predict its latent vector with a pre-trained image-to-latent encoder [35, 4]. Second, we feed the latent vector into a pre-trained StyleGAN2 [20] model for content generation, acquiring generated features. Third, we estimate the optical flow and the invertibility mask between the input and the generated features at multiple resolutions. Then, we warp the generated features with the flow, aiming to minimize the reconstruction error of ID regions. Finally, we composite the input image with the generated content according to the invertibility mask.

Since only the spatial operation, i.e., warping, is applied to the generated features, we maintain their editability with existing GAN editing methods. Combined with the artifact-free and precise inversion effects, our method has excellent superiority in reconstruction accuracy and editing fidelity over existing approaches. In this paper, we adopt StyleGAN2 as the backbone for experiments, and extensive experimental results demonstrate that our method outperforms current state-of-the-art methods with higher reconstruction fidelity and better visual quality.

In summary, our contributions are listed as follows:

1. We propose a novel framework for out-of-domain GAN inversion on human face images by aligning and blending the generated image with the input image via optical flow and invertibility mask prediction.

2. We investigate the GAN invertibility with a novel Spatial Alignment and Masking Module, which is a new solution for invertibility decomposition.
3. Our proposed framework can produce photo-realistic results in both reconstruction and editing tasks. Experiments show that our framework outperforms existing methods in reconstruction accuracy and visual fidelity.

## 2. Related works

**GAN inversion.** The process of GAN inversion involves encoding a real-world image into a semantic-disentangled latent space before reconstructing the image with a GAN generator. It enables various downstream applications, e.g., face editing with labels [32, 5, 36] or texts [25, 12, 3].

To tackle the non-trivial translation between the image and latent vector, some efforts have been made to design more suitable latent space and better encoders. Abdal et al. [1, 2] analyzes the extended GAN latent space ( $W+$ ) for better inversion and attribute manipulation of real images. Based on the  $W+$  latent space, later works further improve the inversion accuracy with hierarchical encoding [27, 16], progressive training [35] and iterative prediction strategies [4, 39]. More recently, Bai et al. [7] investigated the padding space of the StyleGAN generator to increase the invertibility of the pre-trained GAN model. Roich et al. [28] and Nitzan et al. [23] propose strategies to first encode images into latent vectors, then finetune the pre-trained generator for specific images or people. More recently, Alaluf et al. [5] adopt a hypernetwork [13] to modulate the generator kernel for better GAN inversion. However, such methods can hardly invert the out-of-domain contents such as image-specific backgrounds or accessories.

**The distortion-editability trade-off.** To improve the fidelity of GAN inversion, some works [36, 37, 42, 14, 44, 33] add extra connections between the encoder and generator to further extend the latent space, but such methods fall into a tricky dilemma of balancing the distortion-editability tradeoff [36, 29]. A larger latent space helps increase the reconstruction precision but decreases the editability in the generation due to the increasing semantic entanglement problem. Recently, Parmar et al. [24] proposed the SAM with an invertibility prediction module guided by LPIPS [43] loss to predict the spatial invertibility map for the input image. Nevertheless, their invertibility predictions are noisy and inconsistent with the input image, which needs to be smoothed with a pre-trained segmentation network. Furthermore, since SAM is an optimization-based method, it takes a long inference time to produce a high-fidelity result. Moreover, Song et al. [32] propose DiffCAM to refocus the GAN inversion on the attributes to be manipulated. Instead of inverting the whole image, DiffCAM aims to find the edited area and blend the input image

with the edited one. They adopt a semantic classifier (i.e., the facial attributes) as the prior for a rough editing mask prediction. However, due to the misalignment between the generated content and the input image, they suffer from undesired ghosting artifacts in their composition results. Thus, DiffCAM also needs to employ an extra ghosting removal module, while it can not fix the artifacts thoroughly.

Different from existing GAN inversion methods, we simultaneously improve the ID reconstruction via geometric alignment and OOD reconstruction via image blending with the invertibility mask. In our experiments, we compare the baseline methods’ inversion and attribute manipulation performance with our framework. Our method outperforms existing works in reconstruction quality (Fig. 4) and can produce photo-realistic face editing results (Fig. 8) with off-the-shelf GAN manipulation approaches [30, 25, 17].

## 3. Method

In this paper, we propose a new framework for out-of-domain GAN inversion on human faces. The proposed framework can produce a precise invertibility mask to achieve excellent input and generated image composition. The overview of our framework is shown in Fig. 2 and Fig. 3. In the subsequent sections, we discuss the framework’s overview in Sec. 3.1, our proposed Spatial Alignment and Masking Module (SAMM) in Sec. 3.2, the inversion and editing pipeline using our proposed modules in Sec. 3.3, and our training strategy in Sec. 3.4.

### 3.1. Overview

**Traditional GAN inversion.** The essential part of our framework is the estimation of the GAN invertibility on a pre-trained StyleGAN2 [20]. The StyleGAN2 generator is a multi-scale Convolutional-Neural-Network (CNN), where the results are generated gradually with enlarged spatial resolutions. Meanwhile, the image generation is controlled by the latent vector  $w$  in the generation process. Recent works [27, 35, 4, 36, 24, 32] adopt the pre-trained StyleGAN2 generator and aim to inverse the input RGB image into the well-disentangled  $W^+ \in R^{512 \times 18}$  latent space proposed in [27] with an encoder  $E$ . Given an input image  $x \in R^{3 \times 1024 \times 1024}$ , the GAN inversion of  $x$  is to first encode  $x$  into a latent vector  $w \in W^+$  with the encoder  $E$ , then reconstruct the image  $\hat{x}$  with the generator  $G$ , that:

$$\hat{x} = G(E(x)). \quad (1)$$

**GAN inversion with OOD decomposition.** Previous work [11] analyzed the GAN training strategy and discovered that  $W$  space is a low-rank approximation for the target domain of generator  $G$ . Thus, it is difficult to reconstruct  $\hat{x}$  without distortion. To solve this drawback, we tackle GAN inversion from the perspective of invertibility decomposition. Different from [36, 42], we intend to first decompose

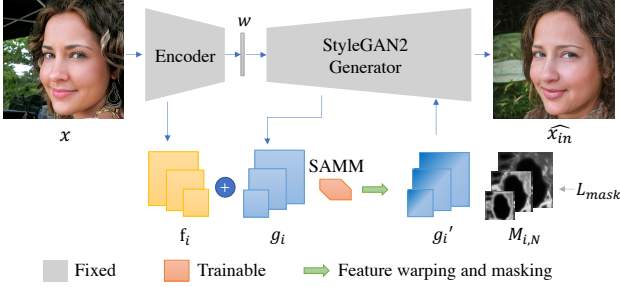


Figure 2: **An overview of our framework.** Our framework begins by extracting features from the input image and aligning the generated features to the input features, which improves the in-domain reconstruction accuracy and eases the invertibility decomposition. Additionally, we predict the invertibility mask for feature and RGB space blending, which enhances the out-of-domain reconstruction quality.

$x$  into the invertible partition  $x_{in}$  and the OOD partition  $x_{out}$  (e.g., accessories, tattoos) with an invertibility mask  $m$ , where

$$x_{out} = x \cdot m, \quad x_{in} = x \cdot (1 - m). \quad (2)$$

Then, we can reconstruct or edit  $x_{in}$  via GAN inversion and preserve the OOD contents  $x_{out}$  for better fidelity, that

$$\widehat{x}_{in} = G(E(x)), \quad (3)$$

$$\widehat{x} = x_{out} + \widehat{x}_{in} \cdot (1 - m). \quad (4)$$

The proposed framework consists of three major parts, i.e., the encoder  $E$ , the generator  $G$ , and SAMM. With the pre-trained and fixed  $E$  and  $G$ , the SAMM estimates GAN invertibility by minimizing the reconstruction error of  $\widehat{x}$ . At the training stage, we adopt the pre-trained image-to-latent encoder  $E$  from [35] and train the SAMM for image-to-image reconstruction. Please refer to the following sections for details.

### 3.2. Spatial Alignment and Masking Module

**The limitation of previous invertibility estimation.** Previous methods for invertibility mask prediction mainly adopt the reconstruction error as the supervision [24]. However, the predictions are usually inaccurate since the subtle ID reconstruction error, which should not be considered as the OOD content, disturbs the prediction of invertibility.

We find these ID errors come from the textural and geometrical misalignment between  $x$  and  $\widehat{x}_{in}$ . While the textural misalignment (e.g., eyes or skin color) can be corrected via latent space optimization [24, 5, 4, 23, 28], it is hard to reconstruct delicate local structures (e.g., the boundary of the face or hair) in  $x$  with  $w$ , especially when the output resolution is high (e.g.,  $1024^2$  and above). In this paper, we

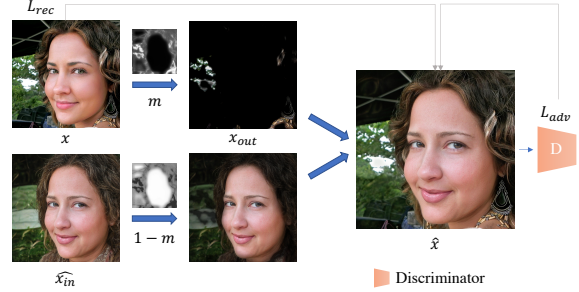


Figure 3: **The invertibility decomposition.** With the predicted invertibility mask  $m$ , we blend the in-domain GAN inversion  $\widehat{x}_{in}$  with the out-of-domain partition  $x_{out}$  to produce the final result  $\widehat{x}$ .

mainly focus on minimizing the geometrical misalignment with spatial operations such as warping.

**Our SAMM for invertibility estimation.** Inspired by previous works done by Collins et al. [9] and Chong et al. [8] where they discovered that simple spatial operations in the GAN feature space maintain the realistic textures in the results. We also find that because  $w$  controls the generation by modulating the variance of the generated feature maps channel-wisely, simple spatial warping in the feature space of  $G$  will not compromise the editability of the framework. Thus, we train SAMM with an optical flow prediction module to predict the flow between  $x$  and  $\widehat{x}_{in}$  in the ID regions. Then, we warp the generated features with grid sampling ( $\mathcal{GS}$ ) to reduce the geometrical error in  $\widehat{x}_{in}$ .

**The details of SAMM.** First, we encode the input image into the latent vectors  $w$  and the input features  $f_i$  ( $i \in \{1, \dots, L\}$ ). Then, feature map  $g_i$  is produced in the  $i$ -th layer of generation. During the generation, we feed both  $g_i$  and  $f_i$  into the SAMM for flow and mask prediction. Meanwhile, we do the prediction in an iterative manner ( $N$  iterations). The pseudo-code of SAMM is shown in Algorithm 1.

**Algorithm 1** The cycle optical flow and invertibility mask prediction with SAMM in the  $i$ -th layer of generation.

- 
- 1: Initialization:  $j = 1, g_{i,0} = g_i$  and  $\Delta_i^x, \Delta_i^y = 0$
  - 2: **for**  $j = \{1, \dots, N\}$  **do**
  - 3:  $\delta_i^x, \delta_i^y, m_i = \text{SAMM}_i(f_i \oplus g_{i,(j-1)})$
  - 4:  $\Delta_i^x \leftarrow \Delta_i^x + \delta_i^x, \Delta_i^y \leftarrow \Delta_i^y + \delta_i^y$
  - 5: **if**  $j = 1$  **then**  $M_i \leftarrow m_i$
  - 6: **else**  $M_i \leftarrow m_i \cdot M_i + M_i \cdot (1 - m_i)$
  - 7: **endif**
  - 8:  $g_{i,j} = \mathcal{GS}(g_i, \Delta_i^x, \Delta_i^y) \cdot M_i + g_i \cdot (1 - M_i)$
  - 9: **end for**
- 

### 3.3. Blending

**Generate  $\widehat{x}_{in}$ .** Because  $g_{i,N}$  is resampled from  $g_i$ , it maintains the local semantic attributes of  $g_i$  and is spatially

aligned to  $f_i$ . The aligned feature  $g_{i,N}$  is then fed into the next styled-convolution layer  $G_i$  and  $g_{i+1} = G_i(g_{i,N}, w)$ . And finally,  $\widehat{x}_{in} = \text{toRGB}(g_L)$ , where  $g_L$  is the last layer’s feature map and  $\text{toRGB}$  is the output convolution in  $G$ . We visualize the generation process of  $\widehat{x}_{in}$  in Fig. 2. In this paper, we align  $g_i$  and  $f_i$  in multiple resolutions of  $32^2, 64^2, 128^2$ , and  $256^2$ . Also, we set  $N = 2$  in most of our experiments. For more details on our architecture, please refer to our supplement.

**Generate  $m$ .** In order to blend the inversed result  $\widehat{x}_{in}$  and  $x_{out}$  in the RGB domain (Eqn. (4)) for a photo-realistic result, it is crucial to find a blending mask  $m$  which precisely distinguish the OOD contents only. As is shown in Fig. 5, we observe that when we set  $\phi_{area}$  to a small value in the training stage, the high-intensity values in  $M_{i,N}$  primarily gather around the OOD area (e.g., earrings, glasses). The mask of different resolutions focuses on slightly different areas because the target texture of each layer in the source generator is different. Usually, the lower layers of the source generator focus on larger structures, such as the rough shape of the face, and the higher layers focus on detailed structures, such as the skin texture and accessories, which is also observed in [27, 35, 5].

Therefore, we could gather  $M_{i,N}$  in each layer of generation for a final blending mask  $m$  during inference. Instead of directly upscale and merging all  $M_{i,N}$  to  $m$ , we consider that consistent high-intensity area as the OOD area. We adopt a merging function to sequentially merge  $M_{i,N}$ ,  $i \in \{1, \dots, L\}$ , as shown in Algorithm 2.

---

**Algorithm 2** Gathering the masks.

---

- 1: Initialization:  $M_{1,N} = \uparrow(M_{1,N})$
  - 2: **for**  $i \in \{2, \dots, L\}$  **do**
  - 3:    $\widetilde{M}_{i,N} \leftarrow M_{(i-1),N} \cdot (\uparrow(M_{i,N}) - \widetilde{M}_{(i-1),N} + 1)$
  - 4: **end for**
  - 5: Update:  $m \leftarrow \widetilde{M}_{L,N}$
- 

Where  $\uparrow$  means up-sampling to the output resolution of the generator, i.e.,  $1024^2$  for the pre-trained StyleGAN2 generator on face images.

**Generate  $\hat{x}$ .** Our final output of GAN inversion result  $\hat{x}$  is produced following Eqn. (3). The blending process is visualized in Fig. 3, where we blend  $\widehat{x}_{in}$  and  $x_{out}$  with  $m$ .

### 3.4. Training Objectives

In Sec. 3.2, we propose the SAMM to align the generated feature  $g$  to  $f$ , where the invertibility is defined by  $m$ . In this section, we demonstrate the loss functions to train SAMM. First, we assume that the subtle misalignment in  $\widehat{x}_{in}$  can be fixed with simple spatial operations in Algorithm. 1. Hence, we train our framework by minimizing the reconstruction loss  $L_{rec}$  on  $\hat{x}$ . Here we adopt the VGG perceptual loss [18]

$L_{per}$ , the MSE loss and the ArcFace [10] identity loss  $L_{id}$  as our reconstruction objectives, that

$$L_{rec} = L_{per}(x, \hat{x}) + \text{MSE}(x, \hat{x}) + L_{id}(x, \hat{x}), \quad (5)$$

please refer to our supplement for detail definition of  $L_{per}$  and  $L_{id}$ . Also, to make  $\widehat{x}_{in}$  look realistic, we keep the vanilla adversarial loss  $L_{adv}$  for GAN model [20] training. Here we skip the definition of  $L_{adv}$  for simplicity. Refer to our supplement for more details.

#### 3.4.1 Mask Regularization

We hope to maximize the area of  $x_{in}$  to better utilize the GAN invertibility for the follow-up applications (e.g., face attribute manipulation). Thus, we train SAMM to produce  $m$  with the maximum low-intensity area under the supervision of  $L_{rec}$ , since the region with low-intensity value in  $M_{i,N}$  indicates only minor correction in  $g_i$  is needed, which also implicitly defines ID areas in  $x$ . Inspired by [6], we use a regularization loss  $L_{mask}$  on  $m$ .  $L_{mask}$  consists of the binary regularization  $L_{bin}$  and the area regularization  $L_{area}$ :

$$L_{bin}(M_{i,N}) = \min(M_{i,N}, (1 - M_{i,N})), \quad (6)$$

$$L_{area}(M_{i,N}, \phi_{area,i}) = \min(0, \phi_{area} - \frac{1}{|M_{i,N}|} \sum M_{i,N}), \quad (7)$$

where  $\phi_{area,i}$  is the expect OOD size in the  $i$ -th layer, and  $|M_{i,N}|$  is the pixel count of mask  $M_{i,N}$ . Finally, we have

$$L_{mask} = \sum_{i=1}^L [\lambda_1 L_{bin}(M_{i,N}) + L_{area}(M_{i,N}, \phi_{area,i})], \quad (8)$$

where  $\lambda_1$  is the loss weight for  $L_{bin}$ . In summary, our overall objective  $L_{total}$  is:

$$L_{total} = L_{rec} + L_{adv} + L_{mask}. \quad (9)$$

We found that if we loosen the  $L_{mask}$  with a larger  $\phi_{area}$  and a smaller  $\lambda_1$ , the predicted  $m$  has a high value (close to 1) at most pixels except for the eyes and mouth area. Consequently, the reconstruction error of  $\hat{x}$  can be optimized to be low, while the editability must be harmed. As we hope to maintain the off-the-shelf editability of styleGAN generator, in this paper, we set  $\phi_{area} = 0.3$  for  $32^2, 64^2$  masks and  $\phi_{area} = 0.25$  for  $128^2, 256^2$  masks.

## 4. Experiments

In this paper, we adopt the official checkpoints of e4e [35] or ReStyle [4] encoder  $E$  and StyleGAN2 [20] (config-f) generator  $G$ , which were pretrained on the FFHQ [19] dataset, and fixed their parameters. Then, we train our SAMM with  $E$  and  $G$  on FFHQ for GAN inversion to minimize  $L_{total}$  (Eqn. 9). Please refer to our supplement for more training details.

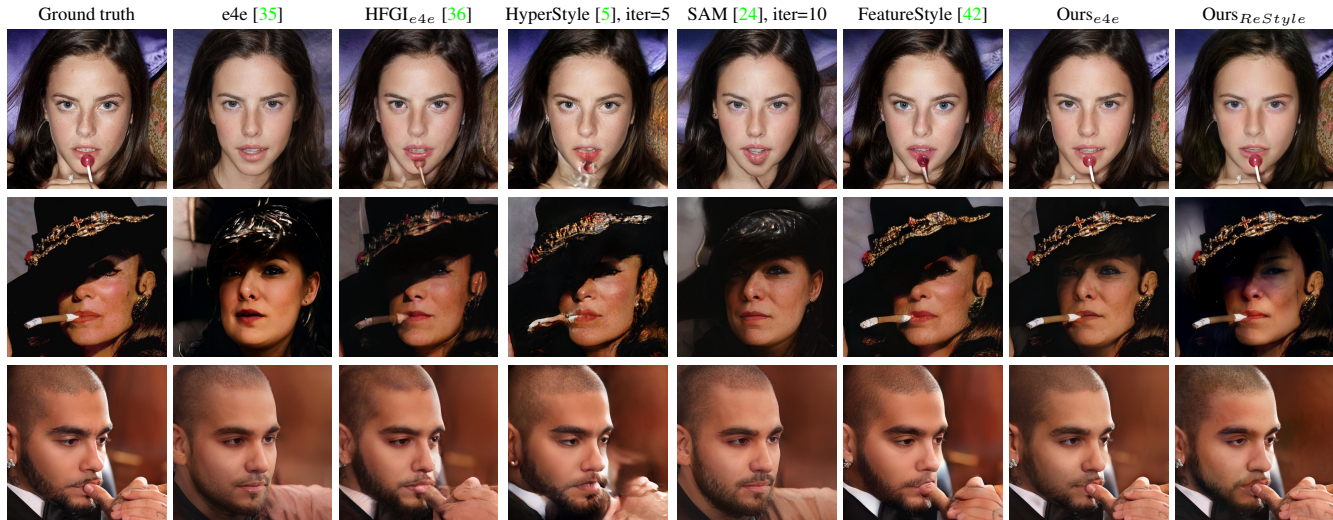


Figure 4: Comparison of GAN inversion quality on CelebAMask-HQ [22] testing dataset. Our method produces the best results in the hats and backgrounds since we skip the generation of the out-of-domain contents. Meanwhile, we eliminate the spatial misalignment in the generator features to better compose out-of-domain content with generated content without ghosting artifacts.

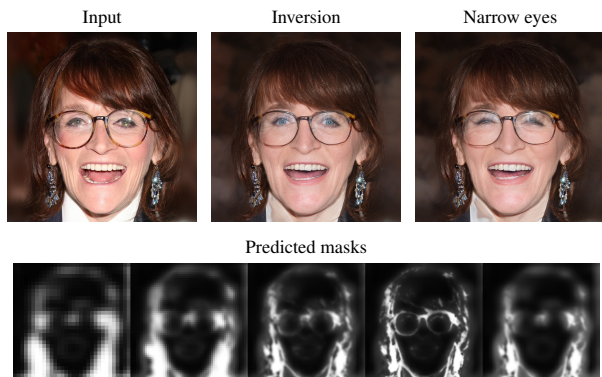


Figure 5: The first row shows our GAN inversion and editing result with our method. The second row shows the predicted masks  $M_{i,N}$  in the resolution of  $32^2$ ,  $64^2$ ,  $128^2$ ,  $256^2$  and  $m$  in the resolution of  $1024^2$ , respectively.

**Face inversion.** Following previous works [36, 32, 24, 42], we evaluate our model and state-of-the-art baseline methods on the first 1,000 images in the testing partition of CelebAHQ-Mask [22] dataset, assessing the inversion quality. We measure the image reconstruction accuracy with PSNR and SSIM [38], the perceptual distance with LPIPS [43], and also the distribution distance between the reconstructed image and the source image that is represented with Fréchet inception distance [15] (FID).

For HyperStyle [5], we set the optimization iteration to 5. For SAM [24], it takes over 131 seconds to finish 1001 iterations of optimization to produce a result, which is too time-consuming for testing. To make a fair comparison, we set the optimization iteration of SAM to 10 in our experiment.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	Runtime(s) $\downarrow$
e4e [35]	20.30	0.665	0.350	34.98	0.145
ReStyle [4]	21.46	0.682	0.345	32.20	0.448
HFGE4e [36]	23.66	0.724	0.285	22.79	0.159
HyperStyle [5], iter=5	23.59	0.728	0.300	30.49	0.915
SAM [24], iter=10	21.95	0.684	0.311	27.38	1.199
FeatureStyle [42]	25.24	0.717	0.188	16.86	0.181
Ours <sub>e4e</sub> , N=1	26.97	0.892	0.147	13.33	0.166
Ours <sub>e4e</sub> , N=2	27.08	0.897	0.143	13.13	0.193
Ours <sub>e4e</sub> , N=3	<b>27.24</b>	<b>0.900</b>	<b>0.139</b>	<b>12.73</b>	0.218
Ours <sub>ReStyle</sub> , N=2	25.37	0.813	0.236	17.68	0.602

Table 1: Quantitative evaluation of GAN inversion quality on the first 1,000 images in the CelebAHQ-Mask [22] testing dataset. The runtime is measured on a single nvidia RTX3090 GPU.

Method	Inversion $\downarrow$	Editing $\downarrow$	Method	Inversion $\downarrow$	Editing $\downarrow$
HyperStyle [5], iter=5	3.67	3.51	SAM [24], iter=10	3.79	3.56
HFGE4e [36]	3.19	4.13	DiffCAM [32]	-	3.75
FeatureStyle [42]	2.68	4.01	Ours <sub>e4e</sub>	<b>1.64</b>	<b>2.04</b>

Table 2: Average ranking of user’s preference on the GAN inversion and attribute editing results of human face images.

Also, because DiffCAM [32] only works for attribute editing, we do not evaluate this model for face inversion. As is shown in Tab. 1, our model outperforms previous methods, achieving the best restoration quality measure with different metrics. In Fig. 4, we present multiple inversion results for comparison. Our method preserves meticulous details in the background, hats, and even the cigarette on the face by conducting the blending for the out-of-domain objects during the generation. We conduct a user study and ask the users to rank the face inversion and editing results based on the image faithfulness, the detail preservation of OOD con-

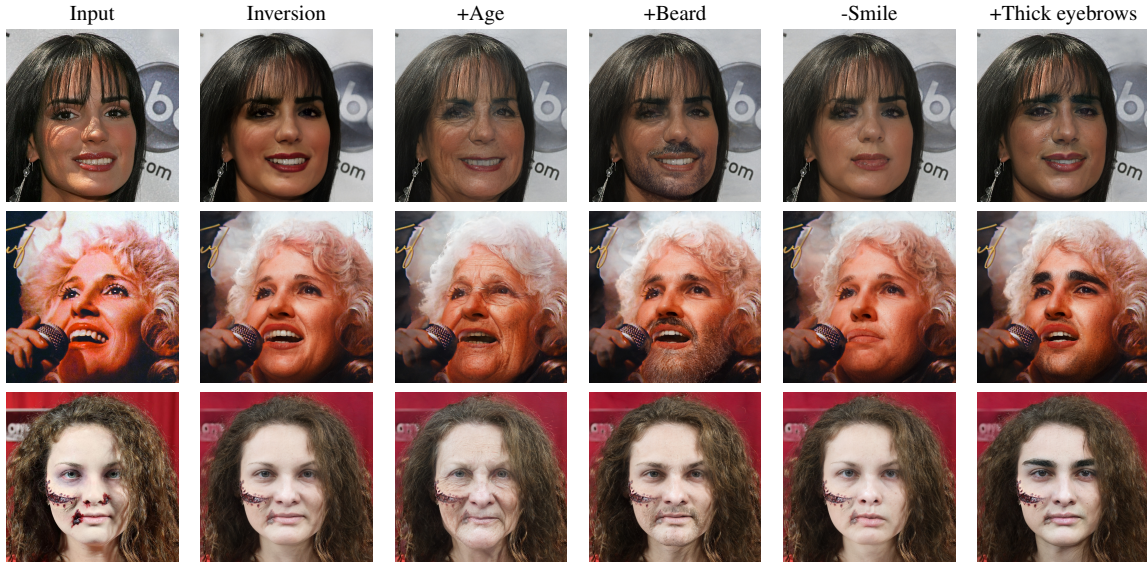


Figure 6: Face manipulation results of our framework on CelebAMask-HQ [22] dataset, please zoom in for detail.

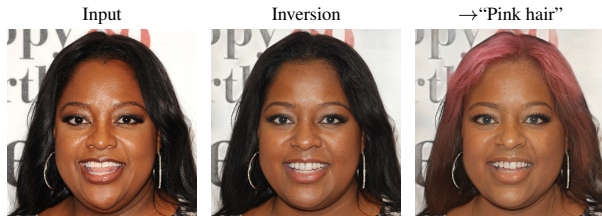


Figure 7: **Edit with text guidance.** Our method also support off-the-shelf attribute manipulation with text guidance [25] for in-domain area (i.e., hair color).

tents, the overall visual satisfaction, and the editing quality. The average rankings are summarized in Tab. 2. Most participants rated both the results of face inversion and attribute manipulation of our framework as the best. Please refer to our supplement for more details of the user study and more visual results for comparison.

**Attribute manipulation.** Also, we have compared our method with state-of-the-art methods [36, 4, 5, 24] on face attribute manipulation [17, 25]. In Fig. 6, we apply off-the-shelf GAN editing approach [31, 31] with our framework on CelebAHQ-Mask [22] images, our method preserves the out-of-domain contents regardless of the editing direction. Moreover, as shown in Fig. 7, we also perform text-guided semantic editing on the hair color with the CLIP [26, 25] model. In Fig. 1<sup>1</sup> and Fig. 8, we compare our attribute editing performance with baseline approaches. Our framework produces high-fidelity editing results without undesired artifacts and provides the best editing quality against existing works. Compared with [5, 24, 36, 42], our work preserves more details in the microphone, hat, and background.

<sup>1</sup>Image license: [www.pexels.com/license](http://www.pexels.com/license)

In addition, we notice an unnatural over-sharpened artifact in the result of DiffCAM [32] (Fig. 1(e)). We hypothesize their ghosting removal module introduces such an artifact. Instead, our SAMM module helps decrease the ID reconstruction error along with the generation of  $\hat{x}_{in}$ . Thus we do not need a post-processing process after the blending. Furthermore, ours works better than [32, 24] in cases with occlusions on faces. For more visual results, please refer to our supplement.

#### 4.1. Ablation Study

In this section, we conduct ablation studies on our spatial alignment and masking module.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	Avg. OOD area(%) $\downarrow$
Ours <sub>e4e</sub> , w/o SA	26.32	0.875	0.171	13.53	31.34
Ours <sub>e4e</sub> , N=2, skip SA	26.68	0.888	0.156	14.26	23.97

Table 3: Ablation of spatial and iterative alignment. The average OOD area is calculated by averaging the intensity of the predicted invertibility masks on the testing dataset.

**Spatial alignment (SA).** We study SA by training the model without SA (“Ours<sub>e4e</sub>, w/o SA”) and by skipping the grid sampling operation in Algorithm 1 on our full model (“Ours<sub>e4e</sub>, N=2, skip SA”). As shown in Tab. 3, we found that the reconstruction quality drops compared to our full model for both settings. Additionally, the predicted average OOD area (AOA) is 7.37% larger for “Ours<sub>e4e</sub>, w/o SA”, while a lower AOA is vital to our framework since we hope to maximize the ID area in the decomposition to better facilitate downstream applications such as attribute editing. In Fig. 9 we provide some visual results of skipping the SA, where we detect undesired ghosting artifacts in the blending result. Please refer to our supplement for more analysis.

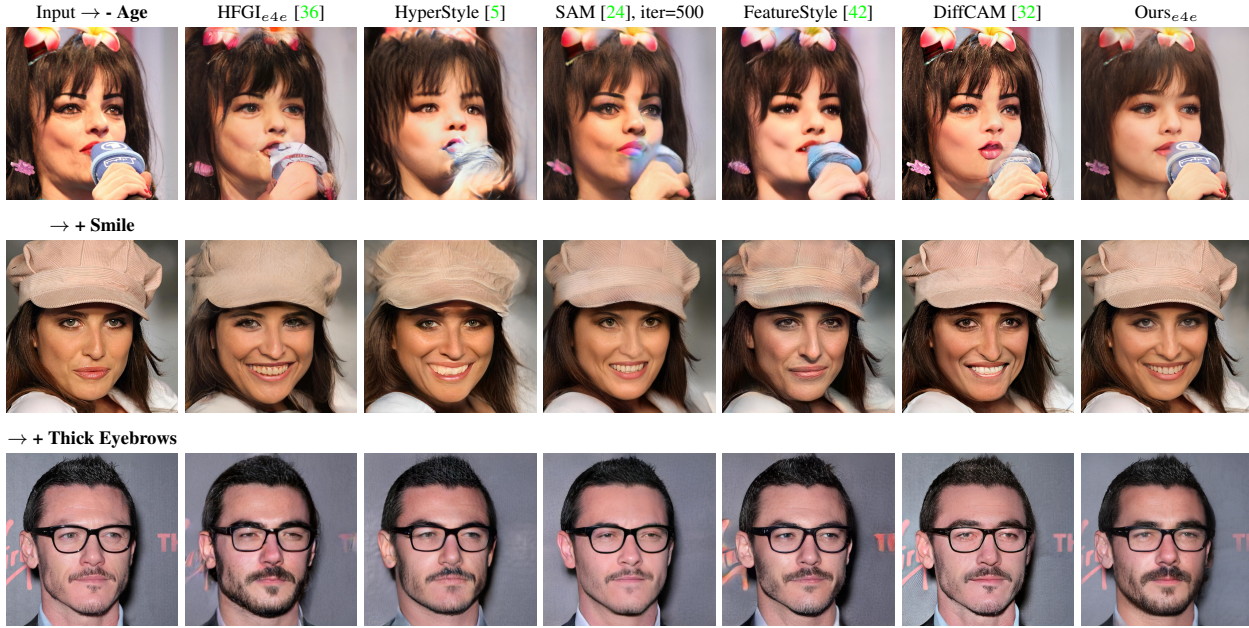


Figure 8: Comparison of attribute manipulation quality via GAN inversion. We apply the same edit on the latent vectors and generate the editing results with baseline methods and our method. Our method produces the best result with well-preserved out-of-domain content and high-fidelity generated content. Please zoom in for detail.

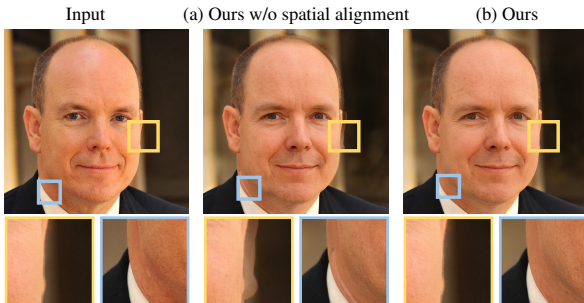


Figure 9: **Ablation study on spatial alignment.** We generate (a) by skipping the grid sampling operation on  $g_i$  in SAMM. Compared to the result of our full model (b), the ghosting artifact is evident in (a). Please zoom in for detail.

**Iterative alignment.** In Sec. 3.2, we introduce the iterative alignment strategy for flow and mask prediction. In Tab. 1, we quantitatively evaluate the inversion quality of our model with  $N=\{1, 2, 3\}$ , and the reconstruction performance increase with the counts of iterative alignment. Besides, we measure the AOA for different  $N$ . With  $N=2$ , 23.97% of pixels in our testing dataset is recognized as the OOD partition and the number increases by 2.4% and 0.65% when  $N=1$  and  $N=3$ . It suggests that the iterative alignment strategy also improves the editability of our framework. Also, we conduct experiments to investigate the usefulness of our loss function on the masks. Please refer to our supplement for more details.



Figure 10: Our GAN inversion and manipulation on Standford Cars [21] dataset.

## 4.2. Extension to non-face domains

Conceptually, our framework could be extended to other image domains such as cars for better OOD GAN inversion quality. As shown in Fig. 10, our framework also improves the car inversion result, compared to e4e [35]. We will leave more investigation of non-face domains to our future work.

## 5. Conclusion

In this paper, we propose a novel framework for photo-realistic OOD GAN inversion via invertibility decomposition on human face images. We design the SAMM to predict the invertibility mask and the optical flows along with the generation process, which can work as a general plugin with different pre-trained encoders (i.e., e4e [35], ReStyle [4]). We only invert the ID area in the input image using the encoder and generator, then align the generated features with the input features for better reconstruction accuracy without compromising the editability. Then, we can manipulate the attributes in the ID area with off-the-shelf approaches. Lastly, we seamlessly merge the OOD contents with the generated image. Our framework produces photo-



realistic results and outperforms previous works in terms of both inversion accuracy and editing fidelity, as demonstrated in our experiments.

## References

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 1, 3
- [2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020. 1, 3
- [3] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. *ACM TOG*, 2022. 3
- [4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *ICCV*, 2021. 1, 2, 3, 4, 5, 6, 7, 8
- [5] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H. Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *CVPR*, 2021. 1, 3, 4, 5, 6, 7, 8
- [6] Jeongmin Bae, Mingi Kwon, and Youngjung Uh. Furrygan: High quality foreground-aware image synthesis. In *ECCV*, 2022. 5
- [7] Qingyan Bai, Yinghao Xu, Jiapeng Zhu, Weihao Xia, Yujin Yang, and Yujun Shen. High-fidelity gan inversion with padding space. In *ECCV*, 2022. 3
- [8] Min Jin Chong, Hsin-Ying Lee, and David Forsyth. Stylegan of all trades: Image manipulation with only pretrained stylegan. *arXiv*, 2021. 4
- [9] Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. Editing in style: Uncovering the local semantics of GANs. In *CVPR*, 2020. 4
- [10] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5
- [11] Ruili Feng, Deli Zhao, and Zheng-Jun Zha. Understanding noise injection in gans. In *ICML*, 2021. 3
- [12] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM TOG*, 2022. 3
- [13] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. In *ICLR*, 2016. 3
- [14] Jingwen He, Wu Shi, Kai Chen, Lean Fu, and Chao Dong. Gcfsr: a generative and controllable face super resolution method without facial and gan priors. In *CVPR*, 2022. 2, 3
- [15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6
- [16] Xueqi Hu, Qiusheng Huang, Zhengyi Shi, Siyuan Li, Changxin Gao, Li Sun, and Qingli Li. Style transformer for image inversion and editing. In *CVPR*, 2022. 3
- [17] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, 2020. 3, 7
- [18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5
- [19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 5
- [20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1, 2, 3, 5
- [21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *3dRR-13*, 2013. 8
- [22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 6, 7
- [23] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *arXiv*, 2022. 1, 2, 3, 4
- [24] 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*, 2022. 1, 2, 3, 4, 6, 7, 8
- [25] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 3, 7
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7
- [27] 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*, 2021. 1, 2, 3, 5
- [28] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 2021. 1, 2, 3, 4
- [29] Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, 1959. 2, 3
- [30] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 3
- [31] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE TPAMI*, 2020. 7
- [32] Haorui Song, Yong Du, Tianyi Xiang, Junyu Dong, Jing Qin, and Shengfeng He. Editing out-of-domain gan inversion via differential activations. In *ECCV*, 2022. 1, 2, 3, 6, 7, 8
- [33] Xuansong Xie Tao Yang, Peiran Ren and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 3

- [34] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *itw*, 2015. 2
- [35] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM TOG*, 2021. 1, 2, 3, 4, 5, 6, 8
- [36] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *CVPR*, 2021. 1, 2, 3, 6, 7, 8
- [37] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 2, 3
- [38] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 6
- [39] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. E2style: Improve the efficiency and effectiveness of stylegan inversion. *TIP*, 2022. 1, 3
- [40] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *CVPR*, 2021. 1
- [41] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 2
- [42] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A style-based gan encoder for high fidelity reconstruction of images and videos. In *ECCV*, 2022. 1, 2, 3, 6, 7, 8
- [43] 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. 3, 6
- [44] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. Blind face restoration via integrating face shape and generative priors. In *CVPR*, 2022. 2, 3