Data Augmentation for Facial Recognition with Diffusion Model

Zhiqi Huang[1*] Hao Geng[1*] Haoyu Wang[1] Huixin Xiong[2] Zhiheng Li[1✉]
[1]Tsinghua University [2]Megvii
{huangzq22, gengh23, haoyu-wa22}@mails.tsinghua.edu.cn
zhhli@mail.tsinghua.edu.cn xionghuixin@megvii.com
* Equal Contribution ✉ Corresponding authors

Abstract

In recent years, facial recognition technology has made significant progress. However, it also faces challenges in common scenarios of daily life. For example, facial accessories such as masks, glasses, and hats have a negative impact on recognition accuracy. This paper introduces a facial data synthesis pipeline based on the diffusion model, which combines the text-to-image generation method with Mask-ControlNet. The pipeline can generate various common facial occlusions, achieving diverse and high-fidelity facial image generation. By comparing the performance of different models trained with synthetic and real images, extensive experimental results confirm the effectiveness of this method in enhancing the robustness of facial recognition.

1. Introduction

Facial recognition is a technology that matches faces in digital images or videos with faces stored in a database. Currently, facial recognition based on deep convolutional networks such as DeepFace [1], FaceNet [2], and OpenFace [3] are widely used, and they typically achieve high accuracy on normal faces. However, in daily life, factors such as hairstyles, accessories, and clothing may affect recognition performance. The reason is that the collected samples are relatively single and limited in quantity.

To tackle these challenges, researchers have proposed some data augmentation methods. As shown in the first row of Fig 1, they manipulate original images by rotating, flipping, scaling, cropping, adding noise, etc. to generate additional training samples. In addition, some GAN-based methods, such as DiscoGAN [4] and BeautyGAN [5]. These methods can be used for hairstyle transfer and facial makeup transfer. As shown in the second row of Fig 1, diffusion-based generative models like DiFaReli [6] and Diffusionrig [7] have also been utilized for facial image restoration and augmentation.

Before data augmentation, these generative models may need to first obtain prior knowledge of a person's face in order to have a high degree of restoration, such as DreamBooth[8]. Inpainting methods[9], although having high object fidelity without additional training, may result in facial disharmony in the generated images.

Figure 1. Left column is original data. The right shows the images after data augmentation. The upper row is based on geometric transformations and pixel operations. Lower is showing various transfers using Diffusionrig [7].

Our method solves this problem by introducing an additional mask prompt. We fine-tune the pre-trained diffusion model without prior knowledge of specific faces. With the great advancements of large vision models like SAM [10], the facial mask can be easily obtained. Specifically, the reference image is first fed to SAM to produce a mask to segment the regions of the face. Then, the resultant image is concatenated with the reference image as the conditional information for image synthesis. The additional mask prompt facilitates the network to better maintain the facial details and model the pixel relationship of segmented edges, resulting in higher-quality synthetic images.

The main contributions are summarized as below:

1) We propose a framework termed Mask-ControlNet to achieve higher-quality facial image generation by introducing an additional mask prompt. With the help of this mask prompt, different accessories and hairstyles can be synthesized for facial data, while maintaining high-

fidelity facial features.

2) We conduct extensive comparative experiments to test the data augmentation effect of generated images. Quantitative and qualitative results show that our framework can generate diverse and high-quality datasets, which can effectively improve the robustness of facial recognition.

2. Related work

In this section, we introduce some common methods of facial recognition and related datasets. Next, we briefly review recent advances in generative models that can be used for facial data augmentation.

2.1. Facial Recognition Models

Over the past few years, significant advancements in deep neural networks, coupled with the employment of extensive facial datasets, have markedly improved the performance of facial recognition systems. State-of-the-art facial recognition models [11–15] use large-scale face datasets such as DeepFace [1], FaceNet [2], and VGGFace2 [16] to optimize the performance [17–22]. To avert overfitting and to enhance the robustness of these systems, the training dataset needs to simulate the diversity and unpredictability of the real world.

2.2. Generative Models for synthesizing faces

Due to privacy concerns and resource limitations associated with collecting real facial datasets, there are technical challenges in gathering diverse facial images. Recent research has pivoted towards employing synthetic data as a substitute for real data in the training of facial recognition systems. Major deep generative models, such as Variational Autoencoders (VAE) [23], Generative Adversarial Networks (GAN) [24, 25], Autoregressive Models [26], and Diffusion Models [27–29], are capable of sampling from existing data distributions to generate synthetic data that closely resembles real-world data. Specific GANs [30–32], including CONFIG [33], DiscoGAN [34], VecGAN [35], and Face-ID-GAN [36], can use predefined attributes to regenerate multiple faces of an existing one. In addition, Stable Diffusion [37] has been shown to generate diverse and photorealistic faces, enriching datasets and improving model performance as demonstrated by a recent quantitative analysis [38].

3. Methodology

Given a facial image of a certain person, our goal is to generate an image that maintains high fidelity of facial details, while synthesizing different contexts and compositions based on text prompts.

3.1. Training-time Framework

As shown in Fig. 2, our framework is built on top of a diffusion model and is trained in a self-supervised manner. First, the input image is fed to the VAE encoder to obtain feature maps F and then the noise is progressively added, resulting in F_t. Here, t represents the number of times noise is added. Afterward, the noisy feature maps F_t are passed to the diffusion model to predict the noise and reconstruct the input image.

In parallel to the main path, our framework has an image branch and a text branch to provide additional conditions for the diffusion model. In the image branch, the input image and text prompt are first fed to GroundingDINO [39] to obtain object detection, which is a kind of box prompt used for SAM [10] to produce the object mask. Then, the resultant mask is used to segment the object in the image. Next, the concatenation of the object image and the image is passed to an adapter layer. Afterward, a VAE encoder and ControlNet are employed to control the diffusion model to reconstruct the input image. In the text branch, BLIP[40] is adopted to extract textual descriptions of the input image. Then, the extracted text prompt is fed to CLIP [41] to provide additional control to the diffusion model. Finally, the features extracted from the text and image prompts are connected to the diffusion model with zero convolution layers.

During optimization, only the adapter layer and the ControlNet are trainable while the diffusion model is frozen. The loss function used is defined as:

$$L = E_{z_0,t,c_t,c_f,\epsilon\sim N(0,1)}||\epsilon - \epsilon_\theta(z_t, t, c_t, c_f)||_2^2, \quad (1)$$

where z_0 represents the data in the latent space, c_t and c_f are the text condition and the latent condition, respectively.

3.2. Inference-time Framework

When synthesizing facial data for model training, there are two methods. The first is to preserve facial features and change people's clothing and hairstyle. This requires feeding reference images to SAM to produce a mask to segment the face. Another approach is to generate occlusions on the face, which requires the use of the facial keypoint detection method (MTCNN [42]). For example, if we want to generate a face with a mask, we first need to detect the position of the nose, mask the facial area below the center of the nose, and imply in the text prompt that the person in the image is wearing a mask. In addition, sunglasses, forehead bangs, etc. can also be batch generated through this method. The synthesis effect can be shown in Fig. 4. Then, the concatenation of the mask and the segmented face image is passed to the VAE encoder. Meanwhile, the text that describes the context of the generated image is fed to CLIP. Next, the features extracted from the image and the text prompt are

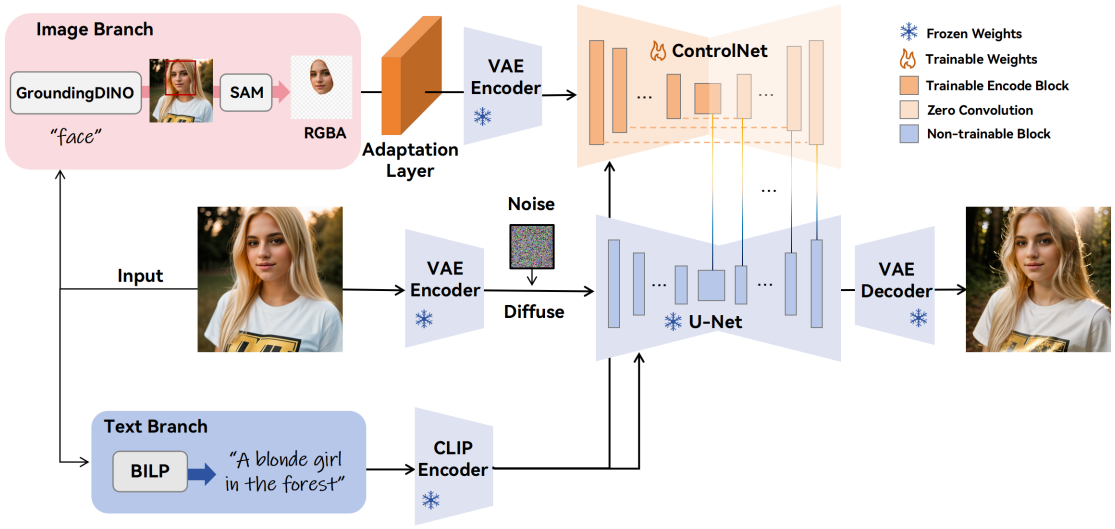

Figure 2. An illustration of our framework during the training phase.

passed through ControlNet and used as conditions for the diffusion model to synthesize an image from a noise image.

4. Experiments

In this section, we first introduce the experimental setup and visualize some face data synthesized by Mask-ControlNet. Next, we conduct several comparative studies to test the performance of different models in face recognition tasks, demonstrating the effectiveness of synthesized images in data augmentation.

4.1. Experimental Setup

During the training phase, we collected 18,000 images from numerous websites using keywords such as people, cosmetics, art photos, and clothing. In addition, we selected 20,000 images from the SA1B and COCO datasets. Mask types include people, faces, clothing, various accessories, etc. After data cleaning and annotation, a total of 38,000 valid images and approximately 50,000 valid masks are obtained as the training set.

In the face recognition task, our training and testing sets are sourced from the facial recognition dataset collected from Pinterest. To test the robustness of the model, there are a total of 105 people, each has 100-200 facial images with significant differences, making facial recognition challenging.

We perform various types of enhancements on each face in the training set, including changing hairstyles and clothing, adding facial accessories, and randomly generating facial features. The synthetic images generated by Mask-ControlNet were created using 10-20 images per individual. These synthesized training data are used to train the facial recognition model, and then we test the performance of these models using closed set recognition.

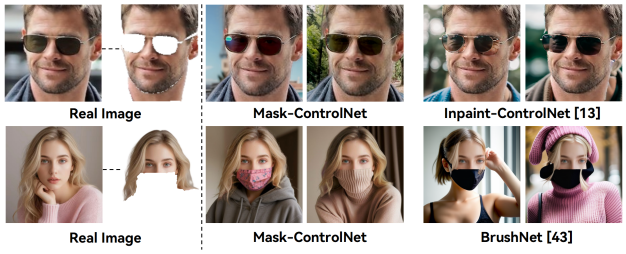

Figure 3. Comparison of synthesized images among Mask-ControlNet, Inpaint-ControlNet [9] and BrushNet [43]

4.2. Performance Evaluation

(1) Facial Data Synthesis

As shown in the first row of Fig. 3, compared to Inpaint-ControlNet [9], our method can generate more diverse and coordinated images, but the background generated by Inpaint is relatively single and the glasses are also very uncoordinated. Compared to BrushNet [43], our method can generate more real and restored images. As shown in Fig. 4, we only need to provide the model with a segmented facial image and a text prompt to control the redrawing of the entire image. From the figure, it is evident that within the mask region, a variety of occlusions, clothing, and hairstyles can be realistically synthesized, while in the non-mask area, the key facial features are still high fidelity.

Figure 4. Four ways to synthesize images. The left of each example serves as the input for Mask ControlNet, while the right contains some images generated under the guidance of the text prompt.

When synthesizing large-scale datasets, the human face can be accurately segmented, but the areas of the eyes, forehead, and mouth are difficult to accurately mask. At present, we only generate glasses, masks, etc. roughly near the key points, which may interfere with training performance. In the following work, we will focus on improving the process of this part to enhance the quality of data synthesis.

(2) Quantitative Results

First, we evaluated the performance of two augmentation methods(random and Mask-ControlNet augmentation) on the three models. Specifically, we trained VGG19 [44] and ResNet50 [45] from scratch and fine-tuned ViT [46] (ViT-B). We compared the test metrics across three levels of dataset augmentation, where the expanded datasets were proportionally scaled to 1x, 2x, and 4x the size of the initial dataset.

The outcomes presented in Tab. 1 demonstrate that, across different dataset augmentation ratios, the Mask-ControlNet augmentation (MCA) consistently achieves superior results compared to random augmentation for all three models. As the proportion of the MCA datasets increases, the effectiveness also improves progressively. However, random enhancement methods such as cropping and rotation can easily lead to overfitting during training. Consequently, we have grounds to believe that our MCA plays a beneficial role in expanding the dataset.

Additionally, we compared four representative face recognition methods: FaceNet [2], ResNet50 [45], InceptionResnetV1 [47, 48], and ViT [46]. The experimental data consists of 100×105 real images and 100×105 generated images. Mask ControlNet generates occlusion of random areas on the face by generating accessories, with the ratio of accessories to face area set to 0%, 20%, and 40%, respectively. We trained FaceNet and fine-tuned the other three models and evaluated their recognition accuracy on the same test set. Additionally, we compared the similarity between the feature vectors(ViT-B encoded) of the generated images and the original images.

Table 1. Comparison of accuracy across different data augmentation methods and dataset enhancement proportions in two models. The ratio in the table denotes the ratio of the original dataset to the augmented dataset. RA indicates data enhancement using the Random Augmentation method, while MCA refers to augmentation employing Mask-ControlNet.

	1:1		1:2		1:4	
	RA	MCA	RA	MCA	RA	MCA
VGG19	0.386	**0.403**	0.377	**0.466**	0.453	**0.483**
ResNet50	0.404	**0.592**	0.519	**0.622**	0.511	**0.656**
ViT	0.875	**0.886**	0.890	**0.912**	0.914	**0.929**

Table 2. Comparison of feature similarity and accuracy on four models for synthetic data with different proportions of mask. IRV1 in the table represents InceptionResnetV1. 0% of the mask indicates that the face in the original image has been regenerated without any changes.

Mask	Similarity	FaceNet↑	ResNet↑	IRV1↑	ViT↑
0%	68.94	**0.257**	0.947	0.865	0.887
20%	65.40	0.251	**0.958**	**0.903**	**0.931**
40%	62.25	0.225	0.953	0.856	0.906

As shown in Tab. 2, as the area of the mask increases, the similarity between images gradually decreases, but the difference is not significant, indicating that the synthesized image can still retain most features of the face. Among the four methods, except for simple networks like FaceNet(CNN) which perform slightly worse than the original data (0.251 vs 0.257), all other methods perform best on synthetic faces with a mask of 20%. This indicates that masking a smaller portion of the face is of great help in supplementing the dataset, while masking a larger area of the face may introduce some noise, thereby interfering with the training of the model.

5. Conclusion

In this paper, we present a simple yet effective framework to synthesize high-quality facial images with an additional mask prompt. With this conditional information, the network can well capture the relationship between object edge pixels. From the quality of generated images, our method can synthesize real and high-fidelity facial images, including various facial occlusions, clothing, and hairstyles. Extensive experiments demonstrate the effectiveness of our synthesized data. Additionally, we expect to explore how to more accurately recognize different regions of the face and achieve more efficient and high-quality facial data synthesis in the future.

References

[1] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 1, 2

[2] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1, 2, 4

[3] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20, 2016. 1

[4] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017. 1

[5] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 645–653, 2018. 1

[6] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22646–22657, 2023. 1

[7] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023. 1

[8] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 3

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[12] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference (ASYU)*, pages 1–5. IEEE, 2020.

[13] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6817–6826, 2020.

[14] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[15] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1578–1587, 2022. 2

[16] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2

[17] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2746–2754, 2015. 2

[18] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.

[19] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7044–7053, 2017.

[20] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018.

[21] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021.

[22] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 2

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[25] 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. 2

[26] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 2

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[28] Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, et al. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. *arXiv preprint arXiv:2403.13304*, 2024.

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[30] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020. 2

[31] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14083–14093, 2021.

[32] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020. 2

[33] Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 299–315. Springer, 2020. 2

[34] Xiaoke Zhang, Zongsheng Hu, Guoliang Zhang, Yongdong Zhuang, Yuenan Wang, and Hao Peng. Dose calculation in proton therapy using a discovery cross-domain generative adversarial network (discogan). *Medical physics*, 48(5):2646–2660, 2021. 2

[35] 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. 2

[36] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 821–830, 2018. 2

[37] 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

[38] Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*, 2022. 2

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul 2022. 2

[41] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[42] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 2

[43] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024. 3

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[46] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[47] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 4

[48] Shuai Peng, Hongbo Huang, Weijun Chen, Liang Zhang, and Weiwei Fang. More trainable inception-resnet for face recognition. *Neurocomputing*, 411:9–19, 2020. 4