StyleGAN-based CLIP-guided Image Shape Manipulation

Yuchen Qian Kohei Yamamoto Keiji Yanai
The University of Electro-Communications, Tokyo, Japan
{qian-y,yamamoto-k,yanai}@mm.inf.uec.ac.jp

ABSTRACT

In this paper, we propose a text-guided image manipulation method which focuses on editing shape attribute using text description. We combine an image generation model, StyleGAN2, and image-text matching model, CLIP, and we have achieved the goal of image shape attribute manipulation by modifying the parameters of the pretrained StyleGAN2 generator. Qualitative and quantitative evaluations are conducted to demonstrate the effectiveness of the proposed method.

KEYWORDS

GANs, text-guided image manipulation, image-text cross-modal model, CLIP

ACM Reference Format:
Yuchen Qian Kohei Yamamoto Keiji Yanai. 2022. StyleGAN-based CLIP-guided Image Shape Manipulation. In *CBMI '22: International Conference on Content-based Multimedia Indexing, Sep 14–16, 2022, Graz, Austria.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Generative Adversarial Networks (GANs) [9] have brought revolutionary changes for image processing field, ranging from image synthesis [6, 29], image editing [10, 18], and even some downstream applications like classification and regression [27]. Recently, style-based generative models [12, 13] boosted the ability of generating realistic and diverse images. Latent spaces of StyleGAN often encode abundant and disentangled visual representations. The latent spaces of pretrained StyleGAN model enable a wide variety of image synthesis and manipulations.

To fully utilize the expressive power of GANs, an intuitive and easy-to-use interface for users is needed. Some image synthesis and manipulation models [15, 17, 26] take natural language as their interfaces. Some previous models [15, 16] limit their range of manipulations to the annotations that exist in the training dataset. By using Contrastive Language-Image Pre-training (CLIP) [19] models, some recent methods [17, 26] break this limitation. With the power of the latent space of pretrained StyleGAN, these models enable a wide range of manipulations. However, most of these works mainly focus on controlling the appearance attributes, such as colors and texture, of object on an image, compared to this, editing the shape attributes, like the shape and size of the object on an image, still remains a unsolved problem.

Figure 1: Visual effects of manipulation achieved by our model

In this work, we focus on semantically editing the shape attributes of an image, especially the editing that bring influences to other parts of the image, such as changing the size of the wheels of a car, which will make change not only on the wheels, but also on other parts of this car. To do this, we combine the power of the state-of-the-art generative model, StyleGAN, with the power of the state-of-the-art visual-language model, CLIP. Different from exiting CLIP-guided StyleGAN-based image editing methods that rely on finding latent directions which change the target attributes in latent spaces, our method manipulates images by finding directions in GAN parameter space rather than latent spaces. In this work, we use a shift to modify the parameters of the pretrained GAN model. By changing GAN parameters, some shape attributes editing that cannot be achieved in latent spaces can be achieved. Furthermore, the shift achieved by a single optimization can be reused for the same target attribute editing, which means a large number of images can be edited in a short time using the previously learned shift. Figure 1 shows several examples of manipulation results using our method. By extensive experiments, we demonstrate the effectiveness of our approach on a wide range of datasets. We also compare our method to existing CLIP-guided StyleGAN-based image editing methods and show that our methods have better performance on shape attributes editing.

To sum up, our main contributions are as follows:

- We propose a text-guided image manipulation model which focuses on editing shape attributes.

CBMI '22, Sep 14–16, 2022, Graz, Austria

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

- By shifting the parameters of pretrained StyleGAN2 generator rather than latent codes in the latent spaces, the proposed method can edit the shape attribute and handle the ratio changes due to the shape changes on one part.
- Experiments shows that our method can edit the shape attribute effectively. Compared to the previous methods which modify the GAN latent spaces, our method has better editing effects and image quality.

2 RELATED WORKS

Latent spaces of generative models often encode much disentangled and interpretable visual representations. Early work [20] on GANs observed the changes on generated images gradually by walking in the latent space. In recent years, with the expressive power of StyleGAN, many researches utilize StyleGAN latent spaces for semantic image manipulation [2, 10, 21, 22, 24]. These methods can be roughly divided into two groups: unsupervised and supervised. Unsupervised methods [10, 22, 24] find meaningful latent directions which make interpretable and distinguishable changes to the image. However, found directions require manual check and annotation. Some supervised methods [2, 21] uses pretrained attribute classifiers to get the label for training and transformation. Thus their range of editing is restricted to the annotations that existed in the training dataset. Some recent methods [3, 17] use CLIP to supervise the transformation, which break the limitation and enable a wide range of manipulations.

As for image manipulation using StyleGAN latent space, most of the works find latent directions in the $\mathcal{W}$ or $\mathcal{W}+$ latent spaces. Also, some methods use other spaces for finding latent directions, such as StyleSpace $\mathcal{S}$ [25], which is demonstrated to be more disentangled compared to $\mathcal{W}$ or $\mathcal{W}+$ latent spaces. The work of StyleCLIP also proposes a method using this $\mathcal{S}$ space to get a better manipulation effect. Furthermore, some works [4, 5] edit the image by modifying the GAN parameter spaces rather than latent spaces. By exploring the GAN parameter spaces, NaviGAN [5] achieves non-trivial visual effects that cannot be produced in the latent space $\mathcal{W}$ or $\mathcal{W}+$. Many of these found the directions related to shape attribute manipulation in the GAN parameter space.

Natural language is often used as an interface for GAN-based image generation and manipulation. Many of previous works [7, 15, 16] for text-guided image manipulation need to train the models from scratch. Their ranges of manipulation are also limited to the training dataset. In recent years, with the appearance of high-performance models that pretrained with large-scale dataset, some works begin to use pretrained models for higher image quality and lower training cost. For the field of text-guided image manipulation, Contrastive Language-Image Pre-training (CLIP) [19] model has attracted much attention, which has achieved the state-of-the-art performance on image-text matching. CLIP is pretrained with 400 million image-text pair data collected from the Internet and the representations that learned by CLIP are extremely powerful, which can be used in a wide range of tasks from zero-shot image classification to image generation and manipulation. Some works [3, 17] combine CLIP with pretrained image generation models like StyleGAN to enable wide-range high-quality image manipulation, which

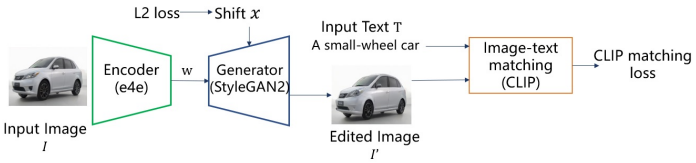

Figure 2: Overview of the proposed model

mainly rely on learned representations in the latent spaces of StyleGAN. Paint by Word [3] lets user edit images with the given text descriptions and masks which indicate the regions to be edited. StyleCLIP [17] gives three approaches for CLIP-guided image manipulation. Two of them find latent directions in the latent space W, and the last one finds latent directions in the latent space S. As the other methods on image synthesis employing CLIP, StyleGAN-NADA [8], CLIP2StyleGAN [1] and DiffusionCLIP [14] have been proposed so far.

In this work, for the purpose of image shape attribute editing, we build our model based on the idea of NaviGAN, which edits the image by modifying GAN parameter spaces rather than latent spaces. We combine this with the powerful image-text matching performance of CLIP to guide the editing with text descriptions.

3 APPROACH

We use a shift x to shift the parameters of one layer of the pretrained StyleGAN2 generator. The proposed model framework is shown in Figure 2. In this section, we first provide a short overview of StyleGAN2 and CLIP, which are the two core components of our model. Next, we describe the architecture and loss function of the proposed model.

3.1 Preliminaries

StyleGAN and its updated version StyleGAN2 have become the state-of-the-art image generators. By feeding different style vectors into each convolution layer of StyleGAN synthesis network, StyleGAN generates high-quality high-resolution images.

StyleGAN has two main components: mapping network and synthesis network. First, a vector z which is sampled from Gaussian distribution is turned into the latent vector w in latent space $\mathcal{W}$ by mapping network. Next, latent vector w is fed into each layer of the synthesis network and the image is generated. Different layers of synthesis network control different parts of features. Thus, some works achieve semantic control on generated images by modifying latent code w or z.

However, for image editing on shape attributes, previous work [5] indicated that some of the shape attributes editing cannot be achieved by modifying latent code w in latent space $\mathcal{W}$. Here, we aim to achieve image shape attributes editing by shifting the parameters of StyleGAN2 convolution layers. For editing real images using pretrained StyleGAN2 model, GAN inversion module which maps a real image into the latent space of GAN model is needed. In this work, we use encoder4editing (*e4e*) [23], which is designed for image manipulation using the pre-trained StyleGAN model.

We use CLIP to provide guide for text-guided image manipulation. CLIP consists of an image encoder and a text encoder. CLIP estimates semantic similarity between image and text by learning

a shared embedding space for image features and text features. With the powerful image-text matching ability, CLIP can provide effective feedback for image manipulation process.

3.2 Network Architecture

The inputs of the proposed model are an image I and a text T. The text T describes the target attribute in a short phrase such as "a small-wheel car". The output is an edited image I'. The goal is to translate the input image I into the edited image I' so that the translated image I' matches the input text T.

The shift x is randomly initialized from a Gaussian distribution. By optimizing x, the shifted generator outputs the image that semantically matches input text T. The input image is first mapped into the latent space $\mathcal{W}$ of StyleGAN2 generator by *e4e*. Then the shifted StyleGAN2 generator outputs the image I'. I' is sent to CLIP with input text T to estimate the image-text similarity of them. This provides feedback for the optimization process of shift x.

For a single optimization, shift x modifies one layer of the pretrained StyleGAN2 generator, obtaining the shift that edits a single target attribute. For layer selection, we first choose a certain range of layers, observe the generation results of those layers, and then decide the target layer. Only the shift x is optimized, and the parameters of the rest parts are fixed.

3.3 Loss Function

We use CLIP matching loss L_{clip} and L2 regularization loss L_2 to optimize the shift x. D_{clip} is cosine similarity between image and text features estimated by CLIP. This is the standard way among the CLIP-based image translation such as StyleCLIP [17].

$$\begin{aligned} L &= \lambda_1 L_{clip} + \lambda_2 L_2 \\ &= \lambda_1 D_{clip}(I', T) + \lambda_2 ||x||_2 \end{aligned} \tag{1}$$

To let the output image I' semantically match the input text T, we use CLIP to estimate the image-text matching similarity of them. In order to preserve irrelevant features of the original image while editing target attribute, on top of satisfying the editing of the target attribute, we use L2 regularization loss on x to minimize the editing. The value of λ_1 and λ_2 depends on the nature of target edit.

4 EXPERIMENTS

In this section, we describe the datasets and experimental settings that used in our experiments. Then, we present the qualitative results and quantitative results.

4.1 Datasets and Implementation Details

We use the StyleGAN2 models pretrained on FFHQ [12], LSUN-Car [28], LSUN-Horse [28] datasets. The GAN inversion model, *e4e*, used in the experiment is pretrained on the corresponding dataset. One single optimization takes one pair of input image and text that describes the target attribute. It takes about 90 seconds to optimize the shift x in one 100-step optimization using a NVIDIA GeForce GTX 1080 Ti. For parameter values of loss function, we use $\lambda_1 = 10$ for our experiments, and the value of λ_2 is chosen to be the best value in the range of 0.01-0.1 depending on the nature of target editing. For overall attribute editing such as face width, the value of λ_2 is set to 0.02-0.03, and for fine-grained attribute editing like eyes size, the value of λ_2 is set to 0.07-0.08.

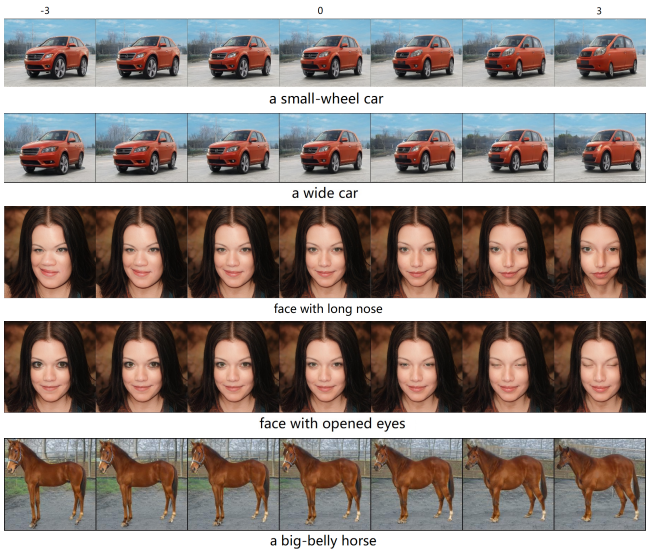

Figure 3: Some visual effects achieved by our model

4.2 Qualitative Results and Analysis

We perform some qualitative analysis to show the performance of image shape attribute editing using our model. First, we present several visual effects achieved by our method. Figure 3 shows five examples. The shift achieved by optimization is multiplied with different magnitude, generating the edited images. Different from appearance attribute editing, shape attribute editing on one part of object may affect other parts, such as the editing on nose length in which the nose becomes longer, the position of the mouth also goes downward. The proposed method can handle such ratio changes on other parts and achieve the target feature editing. Note that the sign of the weights are sometimes reversed regarding the shape change along the text description. On the other hand, some editing such as the result shown in the third row may results in artifact patterns, or editing of irrelevant attributes.

Next, we compare our method with the existing StyleGAN-based CLIP-guided image manipulation method, StyleCLIP, on image shape attribute editing. For StyleCLIP, we use all three methods introduced in the work, two of which perform modification in latent space $\mathcal{W}$ and one of which perform modification in latent space $\mathcal{S}$. Results are shown in Figure 4.

As can be seen, the two methods of StyleCLIP which modify in the latent space W can hardly achieve the goal of shape attribute manipulations. Global directions, which modifies in the latent space S, can achieve some shape feature editing, but it may change the irrelevant style and texture. In the case of nose length editing, our method can achieve the transformation of target attribute better than StyleCLIP. For wheel size editing and horse belly size editing, the images generated by StyleCLIP sometimes change the style and texture of original images. However, the proposed method also

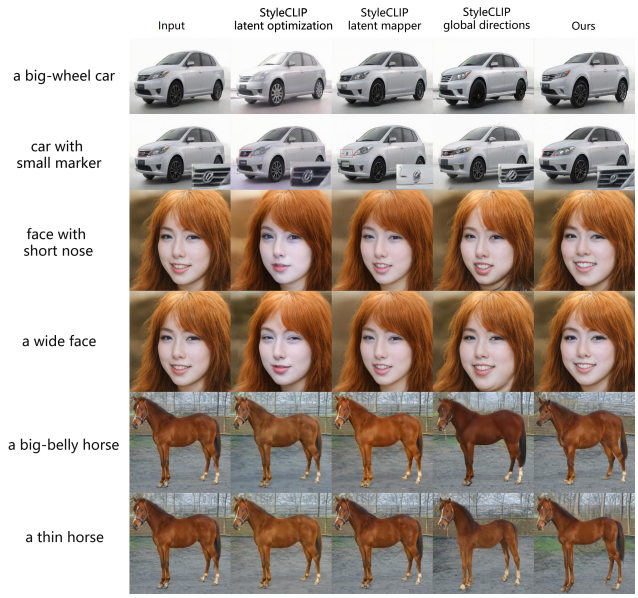

Figure 4: Comparisons of qualitative results with StyleCLIP

sometimes makes changes to irrelevant features, such as changing the shape of the mouth while editing the length of the nose.

4.3 Quantitative Results and Analysis

We conduct quantitative evaluations to show the quality of edited images achieved by our method. The metric we use for quantitative evaluations is Fréchet Inception Distance (FID) [11].

First, in order to show that our method does not harm image quality significantly, we estimate the FID score between the baseline set and the evaluation set generated by our model. The baseline set contains 3000 real images. For the evaluation set, we first perform a single optimization for a target attribute to get a shift. Then the obtained shift is applied on the pretrained generator, and generated 3000 images by editing 3000 real images that used in the baseline set.

We do this experiment on two target manipulations: "wheel size" and "cheek size". For each target attribute, we multiply the shift with different magnitudes to generate the evaluation sets under different strengths. In this experiment, magnitudes ±3, ±5, ±10 are used. The example images that achieved by these magnitudes are shown in Figure 5. Results are shown in Table 1. Since image reconstructed by GAN inversion is not the same as the original image, the FID values of the image set reconstructed by inverse mapping are also included, and the shift magnitude is set to 0.

Next, we conduct quantitative evaluations on images edited by StyleCLIP, and compare it with our model. As before, we measured the FID values of 3000 real images of the baseline set and 3000 images manipulated by StyleCLIP. The baseline set uses the same images as previous, and the evaluation set is obtained by using the global direction method of StyleCLIP. For the parameters of StyleCLIP, we select the parameters which can achieve similar visual effect as the corresponding magnitudes of the proposed method. The example images achieved by StyleCLIP are shown in Figure 6.

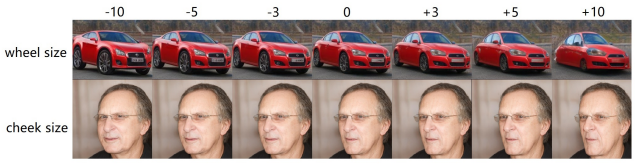

Figure 5: Examples of images generated by proposed method

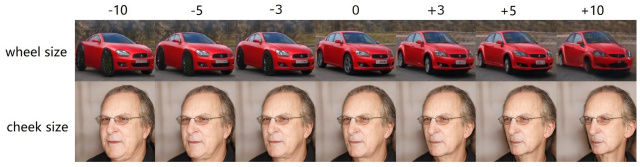

Figure 6: Examples of images generated by StyleCLIP

Table 1: Results for qualitative evaluations

		magnitude	FID	magnitude	FID
Wheel Size	StyleCLIP	-3	33.36	+3	18.16
		-5	42.35	+5	23.33
		-10	54.34	+10	67.57
	Ours	-3	**15.34**	+3	**15.22**
		-5	**17.96**	+5	**21.30**
		-10	**26.27**	+10	**62.39**
	GAN inversion	0	12.54		
Cheek Size	StyleCLIP	-3	28.90	+3	28.40
		-5	**29.16**	+5	29.20
		-10	30.55	+10	30.12
	Ours	-3	**28.89**	+3	**27.96**
		-5	29.34	+5	**28.41**
		-10	**30.21**	+10	**29.86**
	GAN inversion	0	25.6		

Table 1 shows the results of quantitative evaluations between StyleCLIP and our method. The FID values of our proposed method are better than StyleCLIP in the most cases. For wheel size, compared to StyleCLIP, our method preserves the texture of the wheels. For cheek size, our method achieves similar results to StyleCLIP. Overall, the results of the proposed model are proved to be better than StyleCLIP in preserving the image quality.

5 CONCLUSION

In this work, we propose a text-guided image manipulation method that makes use of the pretrained StyleGAN2 and CLIP models. By modifying the parameter of pretrained StyleGAN2 generator, the proposed model can achieve the goal of editing shape attribute. The experimental results showed that, compared to the existing method which modified the GAN latent spaces, our method was effective on image shape attribute editing. Also, our method can preserve the image quality while editing target attribute.

As future works, we plan to select the layer in which the parameters are tuned automatically and to combine multiple layers with adaptive weights for more natural image translation. **Acknowledgment** This work was supported by JSPS KAKENHI Grant Numbers, 21H05812 and 22H00548, 22K19808.

REFERENCES

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy J. Mitra, and Peter Wonka. 2022. CLIP2StyleGAN: Unsupervised Extraction of StyleGAN Edit Directions. In *Proc. of SIGGRAPH.*

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

[3] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. 2021. Paint by word. *arXiv preprint arXiv:2103.10951* (2021).

[4] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. 2020. Rewriting a Deep Generative Model. In *Proceedings of the European Conference on Computer Vision.*

[5] Anton Cherepkov, Andrey Voynov, and Artem Babenko. 2021. Navigating the gan parameter space for semantic image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 3671–3680.

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8188–8197.

[7] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. 2017. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision.* 5706–5714.

[8] Rinon Gal, Or Patashnik, Gal Maron, Haggaiand Chechik, and Daniel Cohen-Or. 2021. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *arXiv:2108.00946* (2021).

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems.*

[10] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. *Advances in Neural Information Processing Systems* 33 (2020).

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).

[12] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4401–4410.

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8110–8119.

[14] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[15] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2020. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7880–7889.

[16] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. 2018. Text-adaptive generative adversarial networks: Manipulating images with natural language. *arXiv preprint arXiv:1810.11919* (2018).

[17] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2085–2094.

[18] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. 2020. SemanticAdv: Generating Adversarial Examples via Attribute-conditioned Image Editing. In *European Conference on Computer Vision.* Springer, 19–37.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).

[20] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations.*

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

[22] Yujun Shen and Bolei Zhou. 2021. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1532–1540.

[23] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–14.

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

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

[26] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2256–2265.

[27] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. 2021. Generative hierarchical features from synthesizing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4432–4442.

[28] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365* (2015).

[29] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5104–5113.