PaintInStyle: One-Shot Discovery of Interpretable Directions by Painting

Berkay Doner* Elif Sema Balcioglu* Merve Rabia Barin*
Umut Kocasari Mert Tiftikci Pinar Yanardag

Boğaziçi University
Istanbul, Turkey

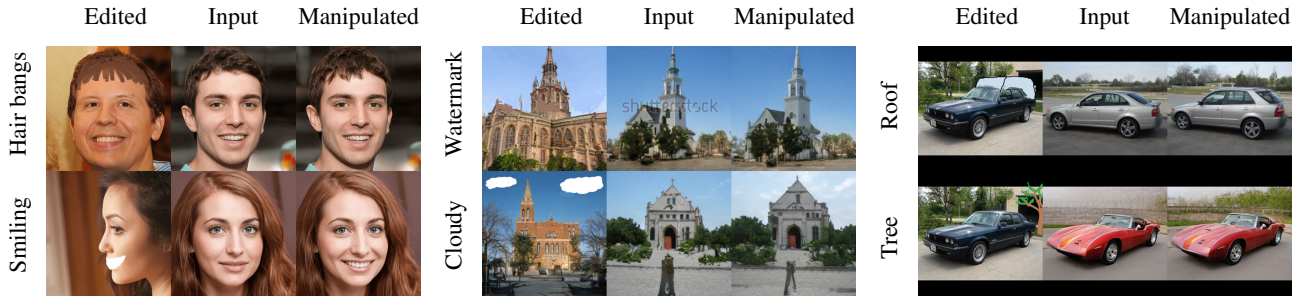

Figure 1. Interpretable directions discovered using our method in the FFHQ [11], LSUN Church, and LSUN Cars [33] datasets. The user drawings on the left are used to manipulate the images in the center.

Abstract

The search for interpretable directions in latent spaces of pre-trained Generative Adversarial Networks (GANs) has become a topic of interest. These directions can be utilized to perform semantic manipulations on the GAN generated images. The discovery of such directions is performed either in a supervised way, which requires manual annotation or pre-trained classifiers, or in an unsupervised way, which requires the user to interpret what these directions represent. In this work, we propose a framework that finds a specific manipulation direction using only a single simple sketch drawn on an image. Our method finds directions consisting of channels in the style space of the StyleGAN2 architecture responsible for the desired edits and performs image manipulations comparable with state-of-the-art methods.

1. Introduction

Numerous studies have been conducted to improve control over the generated images using either supervised or unsupervised approaches. Some supervised studies [7, 24, 31] focus on discovering semantically meaningful directions in the latent space of GANs using pre-trained classifiers or annotated data. Other methods such as [17] achieve some control by training conditional models with labeled data. However, they require significant supervision for each edit. [9] proposes a self-supervised method that is applicable to limited number of pre-defined edit operations. [8, 26, 28] attempt to identify useful directions in an unsupervised manner. However, human interpretation is required to understand which semantic attributes these directions control.

Our goal in this work is to find meaningful latent space directions that can be used to manipulate images in a *one-shot* manner where the user provides a simple drawing (such as *drawing a beard* or *painting a red lipstick*) using basic image editing tools. Our method then finds a *direction* that can be applied to any latent vector to perform the desired edit. Our contributions are as follows:

- We propose a novel method that can discover *directions* in latent space in a one-shot manner with a single drawing provided by the user.
- We demonstrate that our method is able to find several distinct and fine-grained directions in a variety of datasets.

*Equal contribution.

- We make our implementation and demo publicly available to encourage further research in this area: https://catlab-team.github.io/paintinstyle.

2. Related Work

Latent Space Manipulation StyleGAN [11] and StyleGAN2 [12] are two popular GAN models capable of generating high-quality images. A line of research has focused on discovering attribute-specific directions in the latent spaces of GANs to manipulate the semantics of the generated image. Some supervised methods [4, 7, 18, 24] use classifiers trained on specific attributes to discover corresponding directions. [29] identifies attribute-specific control channels using a small number of positive samples. Another type of work [13, 20, 30] enable text-driven manipulations. Alternative methods [8, 9, 25, 28, 35] find meaningful latent directions using unsupervised approaches.

Edit-based Frameworks [10, 16, 22, 32, 34] allow users to edit images by sketching. These methods require the user to specify the area to be edited before sketching the desired attribute in that area, since these methods first perform image completion and then the editing based on the user sketch. [36] eliminates the image completion step, but the user can edit only the object edges. [3] can perform operations to add, remove, or change the appearance of specific objects on regions painted by the user. [5] allows the user to control a limited number of attributes of face images by making changes to their edge map. [14, 19] expect the user to edit the segmentation maps of the images. Some StyleGAN-based methods [1, 2, 15] optimize latent codes of images to manipulate them. [2] takes a drawing and a text input from the user and manipulates the drawn area with the semantics of the given text. [1] performs various types of manipulations, including image editing using *scribbles*. However, the manipulations are limited to a specific image and therefore not applicable to other images. Alternatively, [15] takes an edited segmentation mask of an image and finds global edit directions. [6, 21, 23, 27] focus on encoder architectures that facilitate editing by capturing semantic meanings of latent representations of GANs.

Unlike previous work, our method finds a *direction* in the latent space using a single drawing provided by the user. This direction can then be applied to perform the same edit on new images. The closest work to ours is [15], however the user must edit the *segmentation map* to identify the directions.

3. Methodology

Background: Generator $\mathcal{G}$ of StyleGAN2 structure functions as a mapping function $\mathcal{G} : \mathcal{Z} \rightarrow \mathcal{X}$ from input latent space $\mathcal{Z}$, to target image domain $\mathcal{X}$. From $\mathcal{Z}$, a mapping-network $f : \mathcal{Z} \rightarrow \mathcal{W}$ that consists of 8 fully connected layers, produces the intermediate latent space $\mathcal{W}$. Each $w \in \mathcal{W}$ is mapped into channel-wise style parameters $s \in \mathcal{S}$ using learned affine transformations. Let $\mathcal{G}_s : \mathcal{S} \rightarrow \mathcal{X}$ denote the part of the StyleGAN2 architecture that maps the style parameters to the output image. The space spanned by style parameters forms *style space* $\mathcal{S}$ [29]. In this paper, we use $\mathcal{S}$ to perform manipulations, as $\mathcal{S}$ space has been shown to be the most disentangled space [29]. Our aim is to find channels in the *style space* that lead to attribute-specific directions that performs the desired edit.

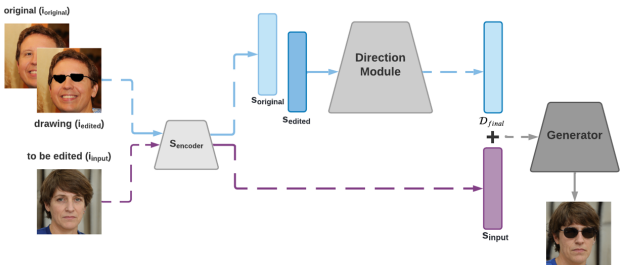

Figure 2. **Illustration of our framework.** Our method takes an original and edited image and identifies a direction using the *Direction Module*. The identified direction can then be used for manipulating new images.

3.1. Finding Directions by Drawing

An illustration of our framework is shown in Figure 2. Our method takes an original and edited image and identifies a direction using the *Direction Module*. The identified direction can then be used for manipulating new images.

Since the edit is made in the image domain $\mathcal{I}$, our method first inverts the images to $\mathcal{S}$ domain using a simple encoder. We train an encoder $E : \mathcal{I} \rightarrow \mathcal{S}$ using a modified architecture of [23, 27], where we invert the images directly into the $\mathcal{S}$ space instead of the $\mathcal{W}$ space. In addition, we did not use a task-specific loss and trained the model with images generated by the random noise vectors $z \in \mathcal{Z}$ instead of real images. Therefore, no additional training data is required and the model can work with only a given pre-trained GAN model.

Our method takes an arbitrary image i_{original}, and its edited version i_{edited}, and finds a direction that can be applied to any new image. A basic approach is to invert original and edited images using the encoder module E and simply find the difference between their style vectors:

$$\mathcal{D}_{\text{basic}} = s_{\text{edited}} - s_{\text{original}} \tag{1}$$

This vector can then be used directly to manipulate the desired attribute of any input image obtained either from a random $z \in \mathcal{Z}$ or using an encoder. The effect of the manipulation can be changed by multiplying the direction

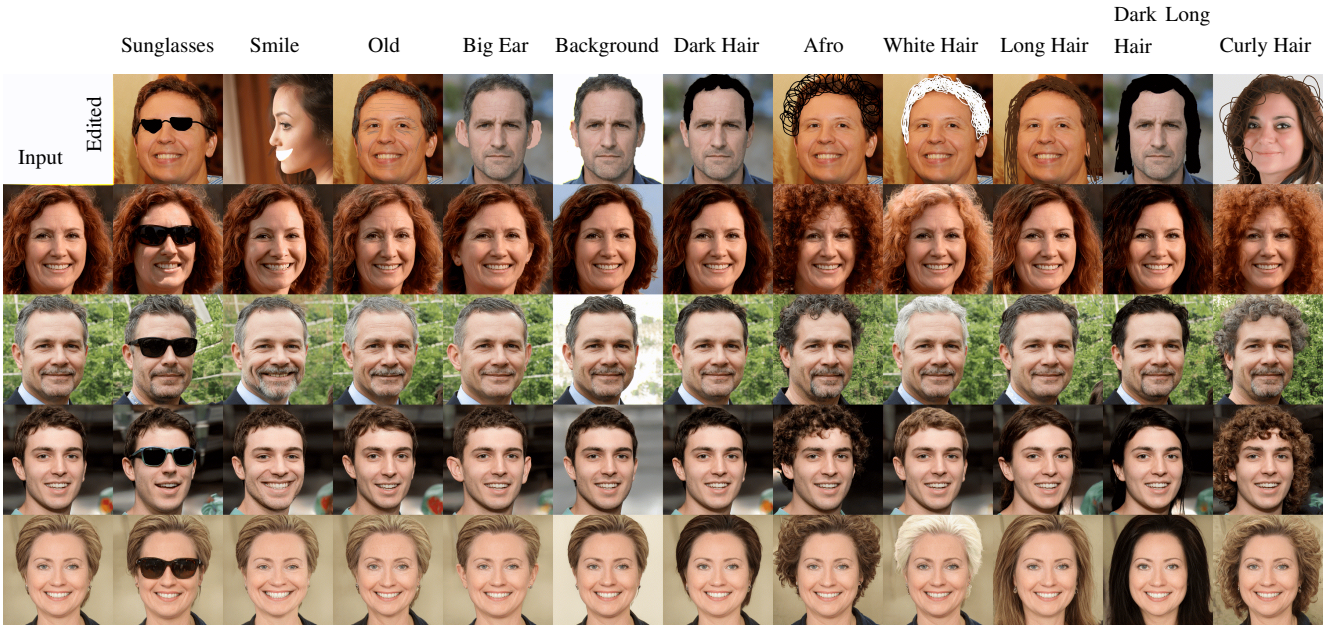

Figure 3. Various edits made on StyleGAN2 FFHQ model using our method. User-drawn sketches are shown in the top row. Other rows show randomly generated human faces and their corresponding edits. The last row shows the manipulations on a real image.

by a scalar, λ:

$$i_{\text{manipulated}} = \mathcal{G}_s(s_{\text{input}} + \lambda \mathcal{D}_{\text{basic}}) \tag{2}$$

While this method allows us to manipulate the image to a certain extent, it does not perform disentangled edits since many irrelevant channels are also obtained (as shown in the ablation study in Figure 4).

Direction Module: Our method uses the *Direction Module* that restricts channels in the obtained direction to the region of interest, i.e., the semantic region of the edit drawn by the user. We use a pre-trained segmentation network [14] in which we identify the segmentation regions that most closely match the drawing, and use them as the region of interest. We extract a boolean mask of the user drawing based on the pixel differences between the original image and the edited image. Then we obtain the segmentation map of $\mathcal{G}_s(E(i_{\text{edited}}))$. We take the top k segmentation regions that have the highest intersection over union (IOU) values with the boolean mask. Using the segmentation map of $i_{\text{manipulated}}$, we select k regions previously identified with i_{edited} and obtain the corresponding boolean mask m on the input image.

To limit the changes to the regions extracted above, we use the *split generator* inspired by [2]. We provide s_{input} and $s_{\text{manipulated}} = s_{\text{input}} + \lambda \mathcal{D}_{\text{basic}}$ and m to the generator. We then upsample and propagate two sets of x and img tensors, representing the area inside and outside the mask, where x refers to the output of the each style block of StyleGAN2 and img is the output of the tRGB blocks. In each layer, we combine these tensors as follows and pass them to the next layer:

$$x = x_{mask\_in} * m + x_{mask\_out} * (1 - m) \tag{3}$$

At the output level, we obtain the image tensor where the region inside the mask has the features from the $s_{manipulated}$ and the region outside the mask has the features from the s_{input}. The encoder network can then be used with the new region-constrained image generated by the split generator, using the same approach that was used to find the basic direction. However, the direction found with a single image may contain image-specific biases. Therefore, this step is performed with N randomly generated images and the average of their style vectors is used as the final direction. To eliminate outliers, the values outside the interquartile range (IQR) for each style channel are removed. This process can be formalized as follows. First, N random z vectors are sampled with their corresponding style vectors s. Let s_j denote the style vector and i_j denote the image generated using the split generator. Then, the direction is determined as follows:

$$\mathcal{D}_j = E(i_j) - E(\mathcal{G}_s(s_j)) \tag{4}$$

Final direction can be represented as follows:

$$\mathcal{D}_{final} = \frac{\sum_{j=1}^{N} \mathcal{D}_j}{N} \tag{5}$$

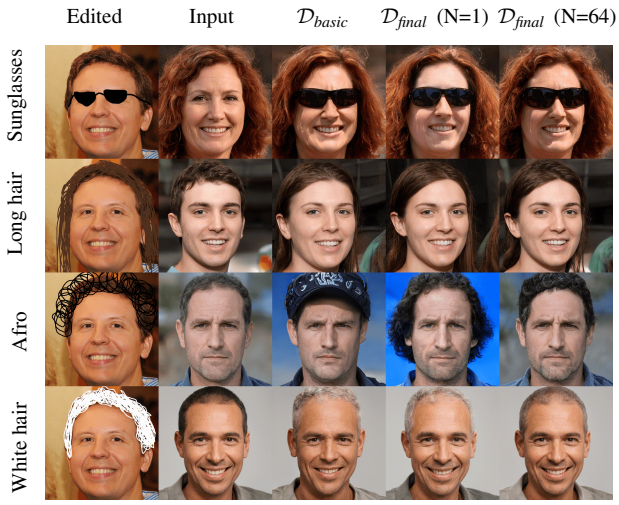

Figure 4. **Ablation study.** Images generated using the basic direction (D_{basic}) and final directions (D_{final}) using N=1 and N=64 images.

4. Experiments

4.1. Experimental Setup

For encoder training, we chose *batch size* of 4, *truncation*= 0.7 and trained the models for 2 to 3 days. We use an NVIDIA Titan RTX GPU. We selected k = 2 segmentation regions when performing the segmentation filtering method. We also randomly sample N=64 seeds to find the final direction. This step takes about 45 seconds per edit.

4.2. Ablation Study

In Figure 4 we have presented the results for different directions. As can be seen in Figure 4, the $\mathcal{D}_{\text{final}}$ preserves the features of the original image better than the $\mathcal{D}_{\text{basic}}$ while successfully manipulating the target attribute. This improvement can be observed for the *long hair* and *white hair* directions. $\mathcal{D}_{\text{basic}}$ manipulates the image towards the desired attribute but also changes the gender of the original image in the *long hair* direction and the identity of the input in the *white hair* direction. In addition, $\mathcal{D}_{\text{basic}}$ may lead to incorrect manipulations in the drawing region. For instance, the $\mathcal{D}_{\text{basic}}$ produces a hat instead of afro hair. Increasing the number of images N in the direction module further improves the quality of the direction, as can be seen from the last column.

4.3. Qualitative Results

Several directions on the FFHQ dataset can be seen in Figure 3. Various edits that change the hair, such as *dark hair*, *white hair*, *long hair*, *dark long hair*, *curly hair*, and *afro*, can be seen in the figure. Our method is also able to find diverse edits such as *sunglasses*, *smile*, *big ear*, *background removal* and complex edits controlled by multiple style channels such as *old*. *Smile* direction is able to capture the semantics of the attribute even when it is drawn on a sideways face. As can be seen in the last row, our method can also carry out edits on real images. We used the latent codes of real images provided by [20] in our experiments.

Figure 5. Manipulations identified in the fashion domain.

We also test our approach with other datasets that have a pre-trained StyleGAN2 network by training encoders for the LSUN Car, LSUN Church, and Fashion datasets. The fashion model is a StyleGAN2 model trained on a custom dataset collected from high-end fashion websites[1]. The directions *tree scenery* and *roof change* on LSUN Cars, *watermark removal* and *cloudy* on LSUN Church are shown in Figure 1. Several fashion manipulations can be seen in Figure 5 where our method is able to find the directions of *flowers, red color, v-neck, long sleeves* and *leopard*.

5. Conclusion

We have introduced a framework that uses latent space inversion to find manipulation directions. Unlike previous work that uses pre-trained classifiers or data annotations, or requires per-image optimization, we propose a one-shot framework to find manipulation directions in style space.

[1]http://farfetch.com

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020. 2

[2] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2, 3

[3] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020. 2

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[5] Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. Deepfaceediting: Deep face generation and editing with disentangled geometry and appearance control. *arXiv preprint arXiv:2105.08935*, 2021. 2

[6] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018. 2

[7] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019. 1, 2

[8] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 1, 2

[9] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019. 1, 2

[10] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1745–1753, 2019. 2

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 1, 2

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2

[13] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. Stylemc: Multi-channel based fast text-guided image generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 895–904, 2022. 2

[14] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 2, 3

[15] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *arXiv preprint arXiv:2111.03186*, 2021. 2

[16] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, Jing Liao, Bin Jiang, and Wei Liu. Deflocnet: Deep image editing via flexible low-level controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10765–10774, 2021. 2

[17] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[18] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006. 2

[19] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[20] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2, 4

[21] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 2

[22] Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep sketch-based face image editing. *arXiv preprint arXiv:1804.08972*, 2018. 2

[23] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Y. Azar, Stav Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *ArXiv*, abs/2008.00951, 2020. 2

[24] 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*, 2020. 1, 2

[25] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020. 2

[26] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 1

[27] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2

[28] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020. 1, 2

[29] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. *arXiv preprint arXiv:2011.12799*, 2020. 2

[30] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021. 2

[31] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, 129(5):1451–1466, 2021. 1

[32] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In *European Conference on Computer Vision*, pages 601–617. Springer, 2020. 2

[33] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1

[34] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 2

[35] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. *arXiv preprint arXiv:2104.00820*, 2021. 2

[36] Yu Zeng, Zhe Lin, and Vishal M Patel. Sketchedit: Mask-free local image manipulation with partial sketches. *arXiv preprint arXiv:2111.15078*, 2021. 2