Rank in Style: A Ranking-based Approach to Find Interpretable Directions

Umut Kocasari[1*] Kerem Zaman[1*] Mert Tiftikci[1*] Enis Simsar[2] Pinar Yanardag[1]
Bogazici University[1] TUM[2]
{umut.kocasari, kerem.zaman, mert.tiftikci}@boun.edu.tr, enis.simsar@tum.de
yanardag.pinar@gmail.com

Figure 1. Interpretable directions *Blue, Floral, Sleeveless, Sequin* and *Leopard* discovered in Fashion model. Original image is shown on the left.

Abstract

Recent work such as StyleCLIP aims to harness the power of CLIP embeddings for controlled manipulations. Although these models are capable of manipulating images based on a text prompt, the success of the manipulation often depends on careful selection of the appropriate text for the desired manipulation. This limitation makes it particularly difficult to perform text-based manipulations in domains where the user lacks expertise, such as fashion. To address this problem, we propose a method for automatically determining the most successful and relevant text-based edits using a pre-trained StyleGAN model. Our approach consists of a novel mechanism that uses CLIP to guide beam-search decoding, and a ranking method that identifies the most relevant and successful edits based on a list of keywords. We also demonstrate the capabilities of our framework in several domains, including fashion.

*Equal contribution

1. Introduction

Popular GANs such as BigGAN [2] and StyleGAN [7] have gained popularity due to their high-quality and realistic image generation capabilities. However, the questions of what knowledge GANs encapsulate in latent space and how the latent representations can be used to manipulate images remain open. Early work uses simple approaches such as modifying the latent representation of images [15] as well as more complex approaches such as searching for directions and interpolating latent codes within pre-trained GANs such as StyleGAN to guide the underlying generation process. Recently proposed methods aim to improve the latent space structure of GANs in more principled ways [5, 6, 18, 20] using supervised or unsupervised approaches. Some recent works, such as Paint by Word [1] and StyleCLIP [14], leverage the joint latent space of CLIP and the generative capabilities of StyleGAN to manipulate real images with text prompts. However, the performance of the manipulation often relies on carefully crafted text prompts that are compatible with the desired manipulation, otherwise the results become unsatisfactory. This limitation makes it difficult to perform text-based manipulations for

users who lack knowledge in the area, such as *fashion*.

In this work, we present *Rank in Style*, a ranking-based approach to find interpretable directions in StyleGAN. In particular, we first rank a large list of keywords to find a variety of semantically meaningful manipulations based on their *editability* and *relevance*. We then extend our method with a novel CLIP-guided text generation strategy to generate a list of domain-specific keywords from a set of generated images, thus removing the dependency on the keyword collection process. Our contributions are as follows:

- We present a novel ranking-based approach to finding interpretable directions in StyleGAN. This approach is particularly useful for scenarios where domain knowledge is limited to generate reasonable text prompts.
- We propose a novel CLIP-guided beam search strategy to generate a list of keywords that can be used to perform domain-specific edits.

2. Related Work

InterfaceGAN [18] uses labeled data such as *gender*, *age*, and *facial expression* to train Support Vector Machines (SVMs) [13]. GANSpace [5] uses the outputs of the intermediate layers of StyleGAN and BigGAN models to find meaningful manipulation directions such as *transformation*, *rotation*, and *augmentation* or high-level attributes such as *gender*, *hair color*, or *age*. They applied principal component analysis (PCA) [21] to the outputs of the intermediate layer of randomly selected latent vectors, using principal components as latent directions. [19] uses dominant eigenvectors of the intermediate weight matrix as latent directions.

Another line of research focuses on finding latent directions using text prompts. [3] proposes an encoder-decoder architecture capable of generating images based on feature representations of images and text. However, they require a labeled dataset consisting of images from the same domain. StackGAN [24] consists of two submodules, the first of which generates a low quality image aligned to the shape and color defined by the text description, and the second of which translates the output of the first stage into a high quality image considering the target visual features. ManiGAN [11] proposes a multi-level architecture capable of learning the association between text segments and visual attributes with a co-attention module. TAGAN [12] uses multiple local word-level discriminators associated with specific visual attributes, and is used to disentangle visual features based on a given text within a single domain. TediGAN [23] trains parallel encoders for text and image that are able to map given input to the latent space of StyleGAN based on visual-linguistic similarity. They are able to manipulate a given image with text instructions by mixing the embeddings generated by the trained encoders. However, these methods require carefully selected text-prompts in order to perform desired edits.

3. Methodology

We propose a method that finds the most successful text-based edits from a list of keywords by considering the attributes *relevance* and *editability*. The list of keywords can be provided manually, or can be automatically determined using the proposed CLIP-guided beam search in Section 3.2.

3.1. Keyword Ranking

Our method uses the Stylespace $\mathcal{S}$ of StyleGAN2 [8] which includes channel-wise style parameters, $\mathbf{s} \in \mathcal{S}$, due to its disentangled nature [22]. Our objective function consists of two distinct parts. First, we consider the *relevance* of a given keyword. From a set of given keywords, the successful texts should be relevant to the images generated by the GAN model. We compute the *relevance* as the similarity between generated images and keywords in the CLIP embedding space. The *relevance* $\mathcal{V}_R$ is defined as:

$$\mathcal{V}_R = \mathcal{S}_{CLIP}(G(s), t) \tag{1}$$

where s is the style code that generates the original input images, G is the pre-trained StyleGAN2 generator, t is the text prompt, $\mathcal{S}_{CLIP}$ is the cosine similarity between CLIP embeddings. *Relevance* of each keyword is normalized between 0 and 1 using the *relevance* value of all keywords.

The relevance is not enough to assess whether a successful manipulation can be performed using that keyword. For instance, *face* keyword in the FFHQ [7] domain is associated with almost all images generated by GAN, but it is not a suitable keyword to manipulate the images. Therefore, we should measure the increase in the similarity score between the text prompt and the images before and after the editing process. The magnitude of the increase in similarity should be high compared to the magnitude of the change. The *editability* $\mathcal{V}_E$ is defined as:

$$\mathcal{V}_E = \frac{\mathcal{S}_{\text{CLIP}}(G(s+\alpha), t) - \mathcal{S}_{\text{CLIP}}(G(s), t)}{L_2(\text{CLIP}(G(s+\alpha)) - \text{CLIP}(G(s)))} \tag{2}$$

where α is the degree of perturbation, CLIP is the model to produce embeddings for images and text, and L_2 is the L_2 norm. Finally, we rank the keywords for each channel by the value of $\mathcal{V}_R\mathcal{V}_E$. Larger values indicate a better match between the keyword and the channel. Each channel is assigned to the keyword with the highest $\mathcal{V}_R\mathcal{V}_E$ score. This is due to the fact that each channel in the style space is responsible for a particular disentangled attribute [22]. To find the most successful channel-keyword relationships, we

consider the difference in ranking score between the first and j^{th} keywords for a channel. We then use the channel, keyword and score information to perform edits.

3.2. CLIP-Guided Keyword Generation

Since collecting keywords from a predefined list is not practical, we use a CLIP-guided text decoding strategy to generate a set of keywords. To this end, we propose a modified beam-search formulation in which the beam scores are updated with respect to the similarity score between a generated image and the previously generated sequence for each time step.

We use a similar notation as in [4], where β_t represents the score of the token t for a candidate beam. Each candidate receives a token extension $w' \in V$, where V represents the vocabulary and each token has the likelihood of $\log p(w'|w_{<t})$. In the greedy approach, the most likely B beams are selected for the next step. We add the CLIP similarity score between the candidate text and the provided image, which is coming from the target domain, to the probability scores of the top-K most likely tokens by model prediction:

$$\beta'_{t+1} = \beta_t + \log p(w'|w_{<t}) + \eta \, \text{CLIP}(x, w_{<t} \oplus w') \quad (3)$$

where $w' \in V_{\text{top-K}}, V_{\text{top-K}} \subset V$ is the candidate token and, $\oplus$ operator stands for string concatenation.

We apply this generation process to N given images up to T steps to curate a corpus containing outputs from all steps. The reason we include the outputs from intermediate steps is that we want to collect a diverse set of descriptions without having to process a large number of images. After text generation, we apply TF-IDF based post-processing to extract the final keywords, as the generated texts can be noisy and contain irrelevant words. By accepting each beam output as a separate document, we calculate TF-IDF scores of all bigrams and trigrams from the generated text corpus. In the final step, we select the top ranking n-grams and remove punctuation and stop words to obtain a unique list of keywords.

4. Experiments

We evaluate the proposed method for detecting semantically meaningful directions using StyleGAN2 models trained on different datasets. We apply the proposed model to StyleGAN2 on a wide range of datasets, including human faces (FFHQ) [7] and Fashion. The fashion model is a StyleGAN2 model trained on a custom dataset collected from high-end fashion websites[1]. We also apply the proposed model with the automatically generated keywords using our CLIP-guided beam search strategy. We then perform several qualitative experiments to demonstrate the effectiveness of our approach. Next, we discuss our experimental setup and then present the results for StyleGAN2 models.

[1]Farfetch: https://www.farfetch.com

4.1. Experimental Setup

For StyleGAN2 experiments, we use *truncation* = 0.7, and use the PyTorch framework and two NVIDIA Titan RTX GPUs. For the CLIP-guided beam search strategy, we set beam size to 15, the maximum sequence length to six more than the number of tokens in the given text prompt, $\eta = 1.5$ and $K = 10000$. To collect all the texts, we generate 100 images for each domain. We use DistillGPT2 [16, 17] as the language model.

4.2. Qualitative Experiments

We use the keyword-based approach for the fashion model, where the list of 120 keywords was collected from the clothing categories of Farfetch. Figure 1 shows the top ranked text-based directions obtained from set of keywords provided. However, since it is not always possible to compile a list of keywords for a given model, we generate descriptive captions for each domain using our CLIP-guided beam search strategy and apply the post-processing and reranking steps to obtain the final list of keywords. Then we apply manipulations for the directions found for the generated descriptions. For the fashion domain, Figure 3 shows the manipulations for selected keywords such as *orange*, *black cotton dress* among the top-20 keywords returned by our method.

Figure 2 shows the comparison between our method and other supervised and unsupervised methods. We compare with StyleCLIP [14] and StyleMC [9] as supervised techniques. As can be seen from the figure, our method produces comparable edits to these. We compare with GANSpace [5] and SeFA [19] as unsupervised techniques. They make more entangled edits and change the identity of the original image, while the edits of our method are disentangled.

4.3. Human Evaluation

We performed a user study with $n = 25$ to compare our method with popular unsupervised latent direction finding methods, namely GANSpace [5] and SeFa [19], in terms of disentanglement and semantically meaningfulness of the discovered directions. First, we present the original input image and the top 10 directions discovered by each method, and ask them to assign a score between 1 and 5 for the disentanglement of the manipulations. Then we ask them to assign a score between 1 and 5 for the semantically meaningfulness of the manipulations. In both cases, higher score indicates better results. Table 1 shows that our method achieves better results in terms of disentanglement

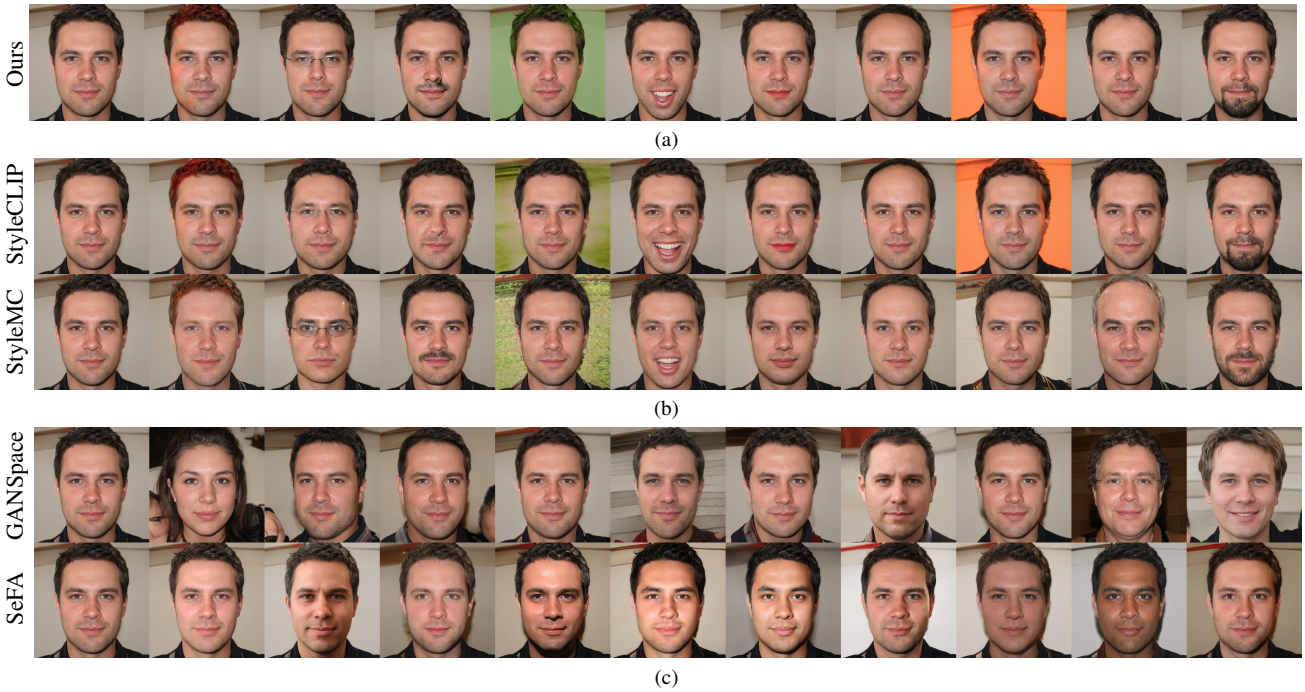

Figure 2. (a) shows the manipulations for our method. The text prompts are *red hair showing, woman wearing glasses, white mustache, green background, mouth open, red lipstick, forehead, orange, characteristic gray hairline, mature beard* respectively. (b) shows results for global direction method of StyleCLIP [14] and StyleMC [9]. (c) shows results for GANSpace [5] and SeFA [19].

Figure 3. Manipulations among the top-20 keywords in Fashion model using the CLIP-guided keyword generation method. From left to right: *orange* ($\alpha = 50$)*, linen brownish* ($\alpha = 60$)*, black cotton dress* ($\alpha = 60$)*, blue dress* ($\alpha = 30$)*, linen white* ($\alpha = 25$)*.*

and semantically meaningful manipulations.

5. Social Impact and Limitation

Our method poses concerns in terms of misuse as in any image synthesis and manipulation method as discussed in [10]. Moreover, the CLIP-guided beam search strategy can be prone to biases caused by GPT and CLIP models.

Method	Disentanglement	Semantically M.
SeFa	2.92 ± 1.31	2.88 ± 1.32
GANSpace	2.80 ± 1.43	2.96 ± 1.33
Ours	3.60 ± 1.31	3.08 ± 1.45

Table 1. Results for human evaluation experiment with mean and std values. Higher score is better.

6. Conclusion

We present a novel rank-based approach to image generation and manipulation by finding interpretable directions in StyleGAN. Unlike previous work that requires prompt engineering to perform text-based image manipulation, our method is able to discover interpretable edits among the large number of domain-specific keywords. Moreover, we eliminate the need to collect a large number of keywords for a given domain by introducing a CLIP-guided beam search strategy. We compared our method with various supervised and unsupervised latent manipulation methods in terms of disentanglement and semantically meaningful changes and obtained favourable results.

References

[1] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 1

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096, 2018. 1

[3] H. Dong, Simiao Yu, Chao Wu, and Y. Guo. Semantic image synthesis via adversarial learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5707–5715, 2017. 2

[4] Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada, July 2017. Association for Computational Linguistics. 3

[5] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls, 2020. 1, 2, 3, 4

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

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 1, 2, 3

[8] 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

[9] 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. 3, 4

[10] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 4

[11] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Manigan: Text-guided image manipulation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7877–7886, 2020. 2

[12] Seonghyeon Nam, Yunji Kim, and S. Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *NeurIPS*, 2018. 2

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

[14] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1, 3, 4

[15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[17] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. 3

[18] 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

[19] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020. 2, 3, 4

[20] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space, 2020. 1

[21] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 2

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

[23] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse image generation and manipulation. *ArXiv*, abs/2012.03308, 2020. 2

[24] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916, 2017. 2