Mask-Guided Discovery of Semantic Manifolds in Generative Models

Mengyu Yang
my.yang@mail.utoronto.ca
BMO Lab for Creative Research
University of Toronto

David Rokeby
david.rokeby@utoronto.ca
BMO Lab for Creative Research
University of Toronto

Xavier Snelgrove
xavier@cs.toronto.edu
BMO Lab for Creative Research
University of Toronto

1 Introduction

Advances in the realm of Generative Adversarial Networks (GANs) [5] have led to architectures capable of producing amazingly realistic images such as StyleGAN2 [7] which, when trained on the FFHQ dataset [6], generates images of human faces from random vectors in a lower-dimensional latent space. Unfortunately, this space is entangled – translating a latent vector along its axes does not correspond to a meaningful transformation in the output space (e.g., smiling mouth, squinting eyes). The model behaves as a black box providing neither control over its output nor insight into the structures it has learned from the data.

However, the smoothness of the mappings from latents to faces plus empirical evidence [9, 4] suggest that manifolds of meaningful transformations are in fact hidden inside the latent space but obscured by not being axis-aligned or even linear. Travelling along these manifolds would provide puppetry-like abilities to manipulate faces while studying their geometry would provide insight into the nature of the face variations present in the dataset – revealing and quantifying the degrees-of-freedom of eyes, mouths, *etc*.

We present a method to explore the manifolds of changes of spatially localized regions of the face. Our method discovers smoothly varying sequences of latent vectors along these manifolds suitable for creating animations. Unlike existing disentanglement methods that either require labelled data [9, 11] or explicitly alter internal model parameters [1, 2], our method is an optimization-based approach guided by a custom loss function and manually defined region of change.

2 Method

We design functions defined on the images generated by our pre-trained model (we will continue to work with the example of StyleGAN2 trained on FFHQ). The desired property of these functions is that they are at their minimum when only the target region of the face (for instance, the mouth) has changed. We then use standard optimization techniques to discover smoothly varying paths through the latent space that lie on the manifold.

We start with a user-provided initial generated image $\vec{x}^* = G(\vec{z}^*)$, where G is the generator network and $\vec{z}^*$ some latent vector (note that in this work we use StyleGAN2's higher-dimensional intermediate latent space $\vec{z}^* \in \mathcal{W}+$, refer to [6] for details). We then define a rectangular mask region M over the image, for instance around the mouth, and define $\vec{x}^*_M$ as the image formed by cropping $\vec{x}^*$ to M, and $\vec{x}^*_{\overline{M}}$ as its complement (i.e. the rest of the image). We seek a manifold containing images which have primarily changed in the mouth region M but not in the rest of the image $\overline{M}$. We can define this manifold as minima of the function

$$\mathcal{L}_X(\vec{x}^*, \vec{x}_i; M) = |D(\vec{x}^*_M, \vec{x}_{i,M}) - c| + D(\vec{x}^*_{\overline{M}}, \vec{x}_{i,\overline{M}}) \tag{1}$$

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

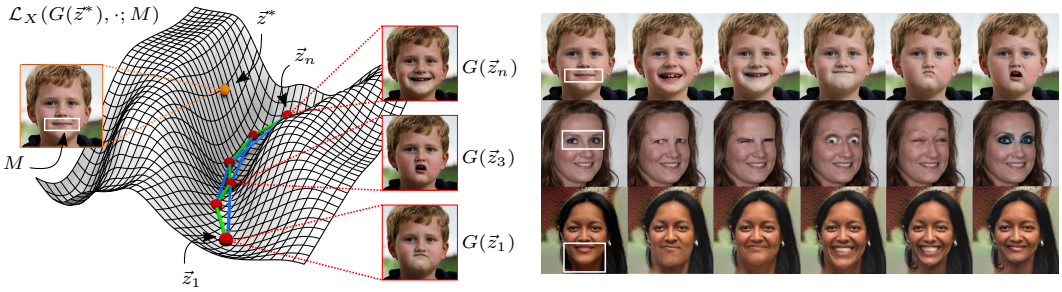

Figure 1: **Left**: A choice of seed vector $\vec{z}^*$, shown as an orange dot, and mask region M creates a function $\mathcal{L}_X$ illustrated here on the latent space of the generative model. The optimization seeks a set of vectors $\{\vec{z}_i\}$, shown as red dots, that lie along the minima of this landscape. They are encouraged to be evenly spaced along a path with minimal curvature by the spring loss $\mathcal{L}_{spring}$, where springs of order $k = 1, 2$ are shown as green and blue connectors respectively. **Right**: Various results of our algorithm. Each row consists of a different seed vector and mask region as shown in the first column. The other columns are selected images from the generated sequences $\{\vec{z}_i\}$. Note that the change in the images is well localized to the masked region. Also note that we use a large value of c to exaggerate the changes for clarity. Refer to the Appendix for many more results.

where $D(\cdot, \cdot)$ is a distance function between images. We have experimented with both L^2 pixel-wise distance and the LPIPS perceptual loss [12]. $\mathcal{L}_X$ satisfies our requirement as it is minimal when the target region has changed by a factor of c while the rest of the image remains unchanged.

In order to create smoothly varying animations that explore this manifold, we use a physically-inspired model of masses connected by springs. Take a matrix of n latent vectors $Z = [\vec{z}_1, \ldots, \vec{z}_n]^T \in \mathbb{R}^{n \times w}$ where w is the dimension of the latent space. The vectors are connected by springs of rest-length σ (an adjustable parameter) in series, encouraging each to be similar, but not too similar, to its neighbours. We further encourage the path to have minimal curvature by also adding higher-order "stiffener" springs connecting to vectors that are further apart. This system can be formalized as follows,

$$\mathcal{L}_{spring}(Z; k) = \sum_{i=1}^{n-k} \left(||\vec{z}_i - \vec{z}_{i+k}||_2 - k\,\sigma \right)^2 \tag{2}$$

For our experiments, we include $k = 1, 2$. Putting everything together, given a reference latent vector $\vec{z}^*$ and mask region M, we use the L-BFGS [8] algorithm to optimize and find $\tilde{Z}$, as seen in Equation 3, where α, β, γ are tuneable parameters controlling the importance of each term. The result of this optimization is visually represented in Figure 1, left.

$$\tilde{Z} = \arg\min_Z \; \alpha \sum_{i=1}^{n} \mathcal{L}_X(G(\vec{z}^*), G(\vec{z}_i); M) + \beta \mathcal{L}_{spring}(Z; 1) + \gamma \mathcal{L}_{spring}(Z; 2) \tag{3}$$

3 Results and discussion

Figure 1, right, shows some of our experimental results. It can be seen that changes to the face are all localized within the mask region while minimal change occurs outside. More importantly, we demonstrate that our method is generalizable to any mask region of choice as well as initial face (see Figures 2 to 10 in the Appendix for additional experiments). The spring constraints of our method are designed to generate smooth videos, please refer to our supplementary material to experience this qualitatively.

This work is a small contribution towards the larger vision of exploring and characterizing the semantic manifolds lurking in the latent spaces of generative models. Generalizing to different models, different dimensionalities of manifolds, and more controls than just rectangular masks are a small sampling of the natural extensions of this line of inquiry.

4 Ethical implications

The StyleGAN2 model we use is capable of generating realistic faces while also demonstrating a proficient understanding of how faces tend to vary in the dataset. Given these qualities, GANs can be used as a popular tool for modelling and promulgating what is considered to be "normal", which if used uncritically, could marginalize people labelled "abnormal" by these systems [3].

Furthermore, there has been much popular discussion about whether we are entering a post-truth contemporary era, where generative tools such as the one we present here have raised fears of hyper-realistic "deepfake" videos impersonating real people, poisoning the information ecology and further eroding trust in any consensus reality [10].

Perhaps more subtly, our method and others like it can create very physically plausible videos of faces changing in "unnatural" ways, such as shifting bone structure, smoothly varying a face from one identity to another. If such videos become commonplace in our culture, might this contribute to a reconfiguring of our traditional conception of separate, fixed, and individual identities towards fluid, overlapping, and changeable ones? The consequences of such a fundamental shift, be they negative, positive, or neutral are difficult to anticipate but worthy of consideration.

References

[1] Yazeed Alharbi and Peter Wonka. "Disentangled Image Generation Through Structured Noise Injection". In: *arXiv:2004.12411 [cs]* (May 2020). arXiv: 2004.12411. URL: http://arxiv.org/abs/2004.12411.

[2] Terence Broad, Frederic Fol Leymarie, and Mick Grierson. "Network Bending: Manipulating The Inner Representations of Deep Generative Models". en. In: (May 2020). URL: https://arxiv.org/abs/2005.12420v1.

[3] Kate Crawford. *The Trouble with Bias*. Conference on Neural Information Processing Systems, invited speaker, 2017.

[4] Antonia Creswell and Anil A. Bharath. "Inverting The Generator Of A Generative Adversarial Network (II)". In: *arXiv:1802.05701 [cs]* (Feb. 2018). arXiv: 1802.05701. URL: http://arxiv.org/abs/1802.05701.

[5] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[6] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *arXiv:1812.04948 [cs, stat]* (Mar. 2019). arXiv: 1812.04948. URL: http://arxiv.org/abs/1812.04948.

[7] Tero Karras et al. "Analyzing and Improving the Image Quality of StyleGAN". In: *arXiv:1912.04958 [cs, eess, stat]* (Mar. 2020). arXiv: 1912.04958. URL: http://arxiv.org/abs/1912.04958.

[8] Dong C. Liu and Jorge Nocedal. "On the limited memory BFGS method for large scale optimization". en. In: *Mathematical Programming* 45.1 (Aug. 1989), pp. 503–528. ISSN: 1436-4646. DOI: 10.1007/BF01589116. URL: https://doi.org/10.1007/BF01589116.

[9] Yujun Shen et al. "Interpreting the Latent Space of GANs for Semantic Face Editing". In: *arXiv:1907.10786 [cs]* (Mar. 2020). arXiv: 1907.10786. URL: http://arxiv.org/abs/1907.10786.

[10] James Vincent. *Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news*. en. Apr. 2018. URL: https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed.

[11] Yi Wei et al. "MagGAN: High-Resolution Face Attribute Editing with Mask-Guided Generative Adversarial Network". In: *arXiv:2010.01424 [cs]* (Oct. 2020). arXiv: 2010.01424. URL: http://arxiv.org/abs/2010.01424.

[12] Richard Zhang et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *arXiv:1801.03924 [cs]* (Apr. 2018). arXiv: 1801.03924. URL: http://arxiv.org/abs/1801.03924.

5 Appendix

Below are some additional figures of experiments with different faces and masks. For all figures below, the first column is the reference image $\vec{x}^*$ with mask region shown. Unless its value is explicitly stated, we use a large value of c to exaggerate the changes for clarity.

We encourage readers to view our video animations, from which stills were taken to create these figures below. An important aspect of our method is that it creates smooth animations while exploring the manifolds. As a result, the video animations convey much more visual information, whereas some of that is lost with still figures. Refer to supplementary materials for the animations.

(a) $c = 0.15$

(b) $c = 0.25$

(c) $c = 0.35$

(d) $c = 0.45$

Figure 2: Each row represents an experiment with a different offset c value on the same face and mask region.

Figure 3: Experiment with a single offset c value with mask region around the eyes.

Figure 4: Experiment with a single offset c value with mask region around the right half of the face.

Figure 5: Experiment with a single offset c value with mask region around everything except the eye region (i.e., eye region remains unchanged).

Figure 6: Experiment with a single offset c value with mask region around the right half of the face.

Figure 7: Experiment with a single offset c value with mask region around the mouth.

Figure 8: Experiment with a single offset c value with mask region around the mouth and chin.

Figure 9: Experiment with a single offset c value with mask region around the right half of the face.

Figure 10: Experiment with a single offset c value with mask region around the mouth.