Navigating the GAN Parameter Space for Semantic Image Editing

Anton Cherepkov, Andrey Voynov, Artem Babenko

Coby Penso

What we'll see

Bar-Ilan University

1. Introduction
2. Optimization-based approach
3. Spectrum-based approach
4. Hybrid
5. Experiments

Different Uses of GANs

Bar-Ilan University

Super-Resolution

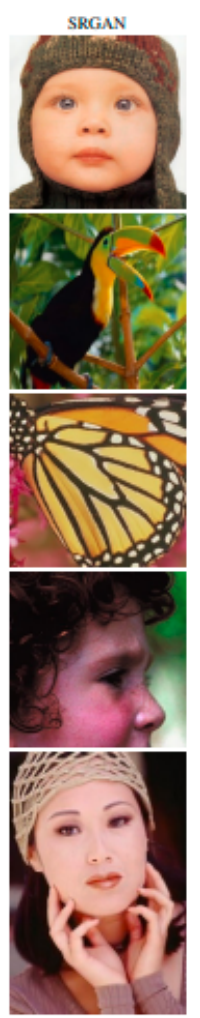
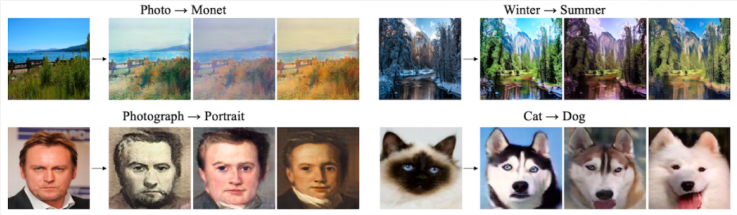
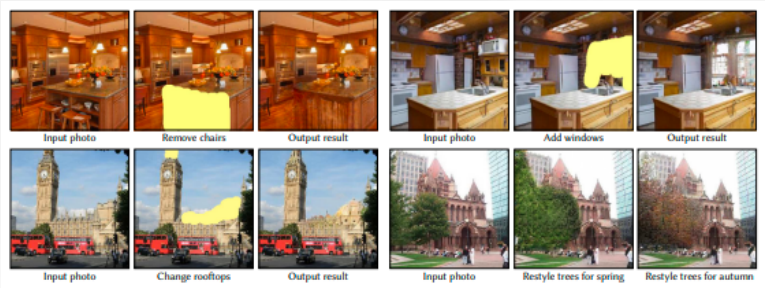


Image-To-Image transfer



Texture transfer

Inpainting

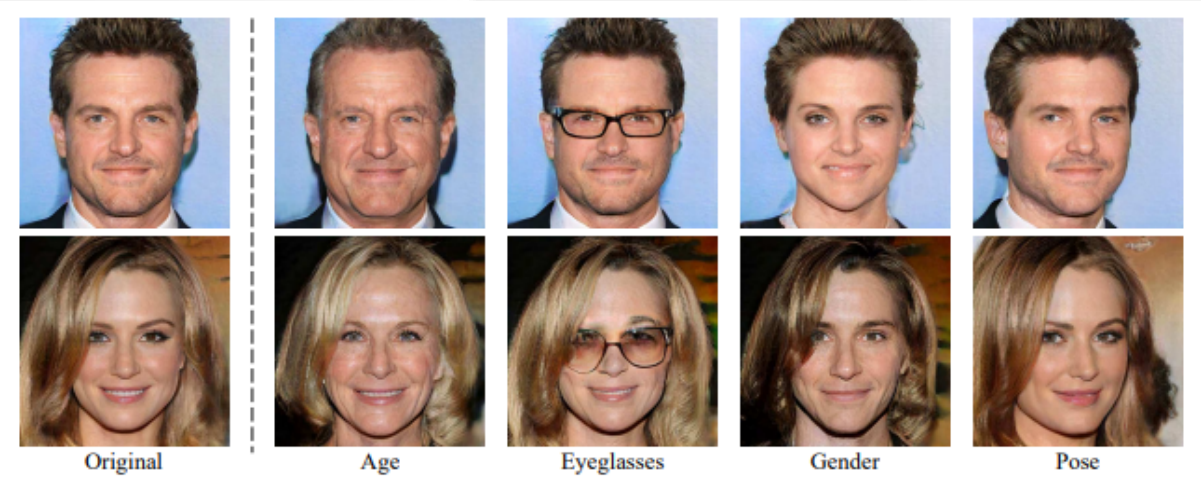


Colorization

Latent manipulations in GANs

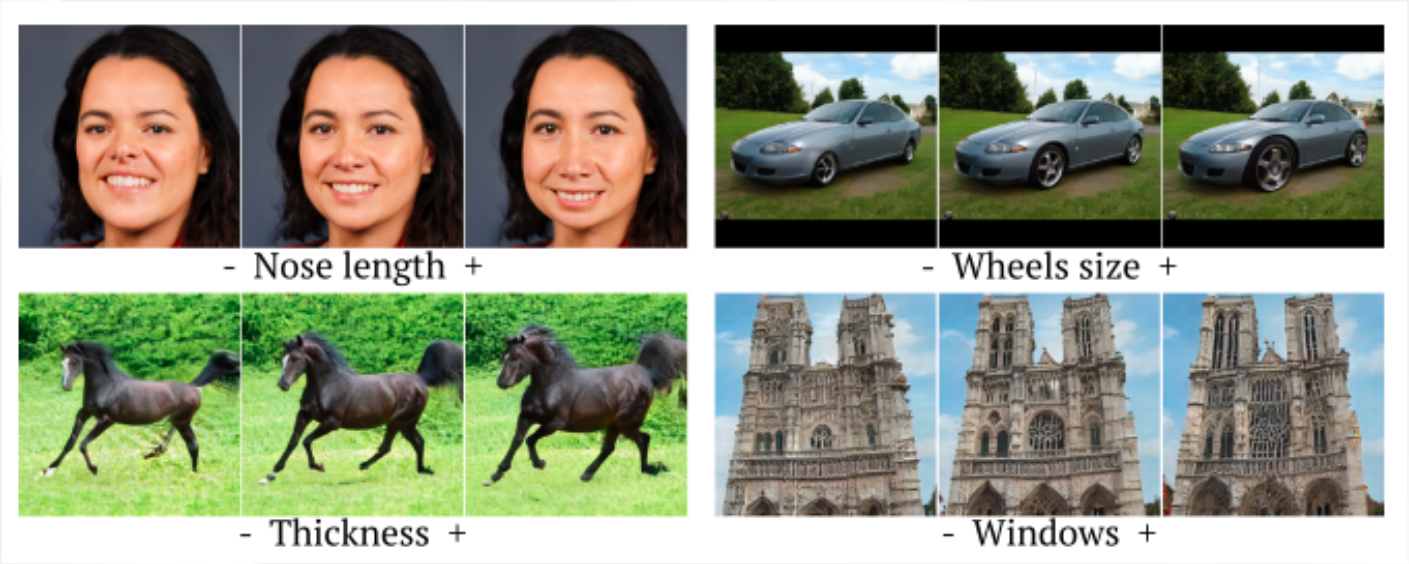
GAN latent spaces are endowed with human-interpretable vector space arithmetic.
Several approaches for latent manipulations for image control:

- Explicit human supervision to identify interpretable directions in l.s.
- Solve optimization problem in latent space. Done via Classifiers pretrained on specific attributes - for pseudo labels.
- Compute the spectrum of the Hessian for LPIPS model with respect to l.s.



Introduction

Image Synthesis using GANs alows us to create realistic images.
But how can we control the properties of the generated image besides being realistic?



In this paper, we'll dive into a method for Semantic Image Editing by Navigating the GAN Parameter Space.

The General Idea

Setting:

- Sample $z \sim \mathcal{N}(0, \mathbb{I}) \in \mathbb{R}^d$
- Image space $I = \mathbb{R}^{WxHx3}$
- $G : z \to I$
- Pretrained GAN generator G_θ with parameters $\theta \in \Theta$

Goal:
Learn $\xi_1, ..., \xi_K \in \Theta$ such that changes along these vectors effectively performs continious semantic editing operations.

$$G_\theta(z) \to G_{\theta + t\xi_k}(z), \;\; k = 1, ..., K, \;\; \forall z \sim \mathcal{N}(0, \mathbb{I})$$

- $\xi_1, ..., \xi_K \in \Theta$ - Correspond to interpretable visual effects
- $t \in [-T, T]$ - shift magnitude - degree of visual effect

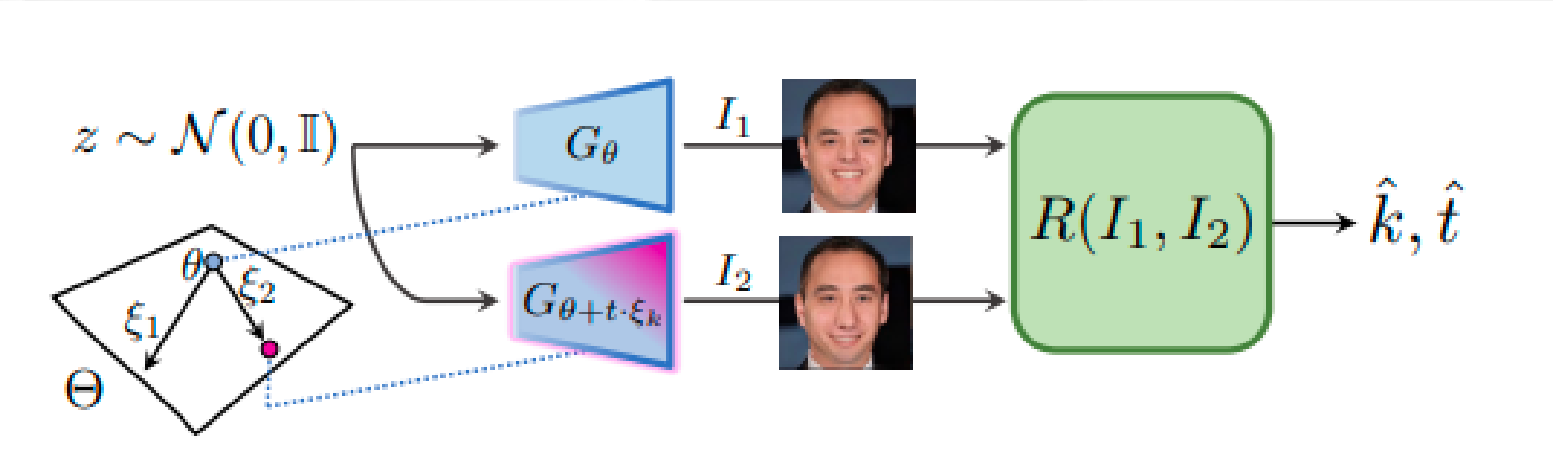
Optimization-based approach

Intuitively, interpretable directions are the ones that are easy to distinguish from each other, observing only the results of the corresponding image manipulations.

Two learnable modules:

- Direction matrix $\Xi = [\xi_1, ..., \xi_K] \in \mathbb{R}^{dim(\Theta)xK}$
- Reconstructor R - get pairs $\{G_\theta(z), G_{\theta+t\xi_k}(z)\}$ and predicts k, t

$$R : (I_1, I_2) \rightarrow (1, ..., K, \mathcal{R})$$



Learning Procedure

Learning is performed by minimizing the expected reconstructor's prediction error:

$$\min_{\{\xi_1,\xi_2,\ldots,\xi_K\},R} \mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}),\, k\sim\mathcal{U}\{1,K\},\, t\sim\mathcal{U}[-T,T]} \left[L_{cl}(k,\hat{k}) + \lambda L_r(t,\hat{t}) \right]$$

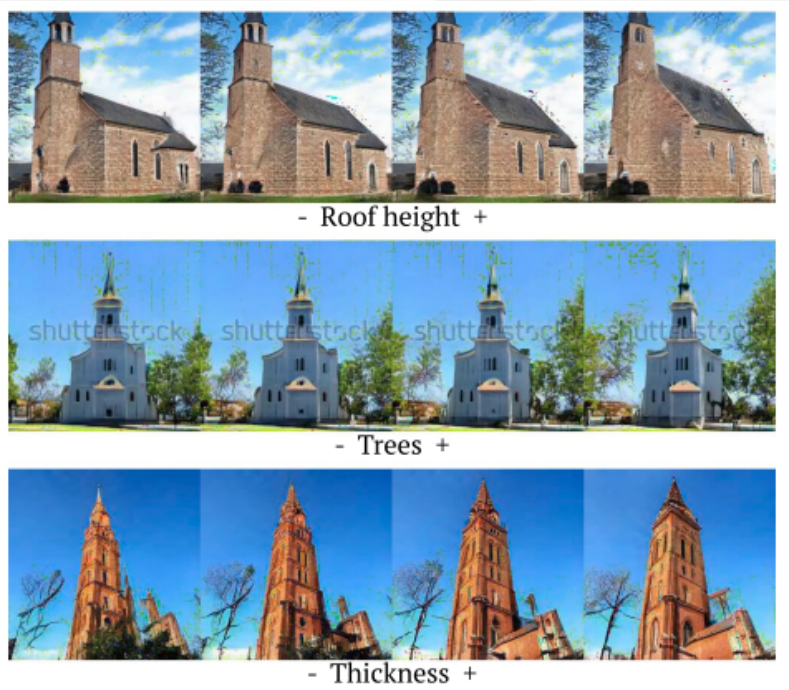
where $\hat{k}$ and $\hat{t}$ denote the reconstructor's output:

$$(\hat{k};\hat{t}) = R(G_\theta(z); G_{\theta+t\xi_k}(z))$$

- For classification objective term $L_{cl}(\cdot,\cdot)$ - Cross Entropy
- For the regression term $L_r(\cdot,\cdot)$ - Mean Absolute Error

Objectives Purpose

- **Classification objective**
  Optimized jointly, the minimization process seeks to obtain such directions that correspond to image transformations that are easier to distinguish from each other, the make the classification problem of R simpler.
- **Regression objective**
  force shifts along discovered directions to have the continuous effect, thereby preventing “abrupt” transformations

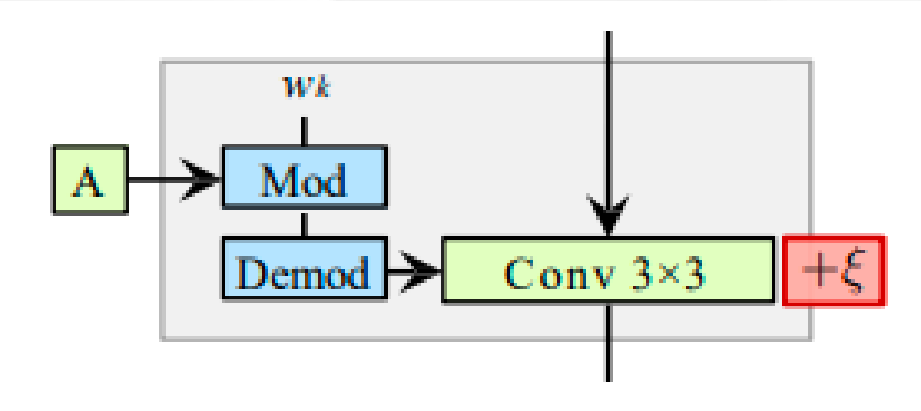


Reducing the dimensionality of the optimization space

SOTA generators (e.g StyleGAN2) typically have millilons of parameters θ.

Problem:
Infeasible to learn from $\Xi \in \mathbb{R}^{dim(\Theta)xK}$ matrix.

Solution - part 1:
Minimize only the shifts $\xi_1, ..., \xi_K$ applied to a particular generator's layer (with all other fixed).



Solution - part 2:

Compute SVD decomposition of the chosen convolutional layer, flattened to a 2D matrix.
Then, optimize $\xi_1, ..., \xi_K$ applied only to the singular values of the diagonal matrix.

$$SVD = U \cdot diag(\sigma_1, ..., \sigma_n) \cdot V$$

Apply additive shift $\xi^{(1)}, ..., \xi^{(n)}$), i.e

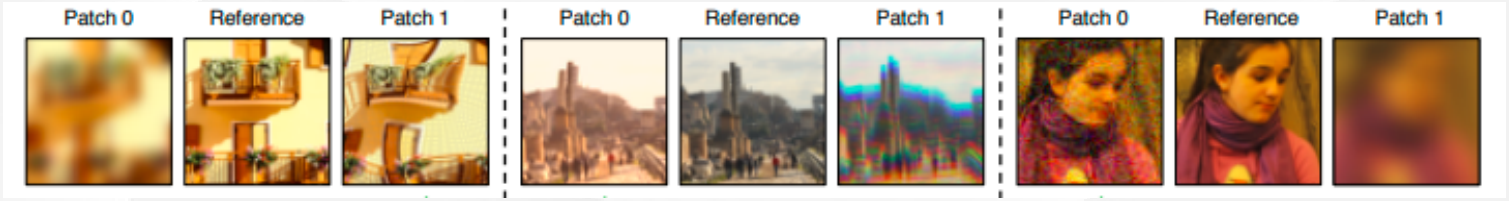
$$diag(\sigma_1, ..., \sigma_n) \rightarrow diag(\sigma_1 + \xi^{(1)}, ..., \sigma_n + \xi^{(n)})$$

Also normalize ξ to a unit length to avoid parameter explosion.

Spectrum-based approach

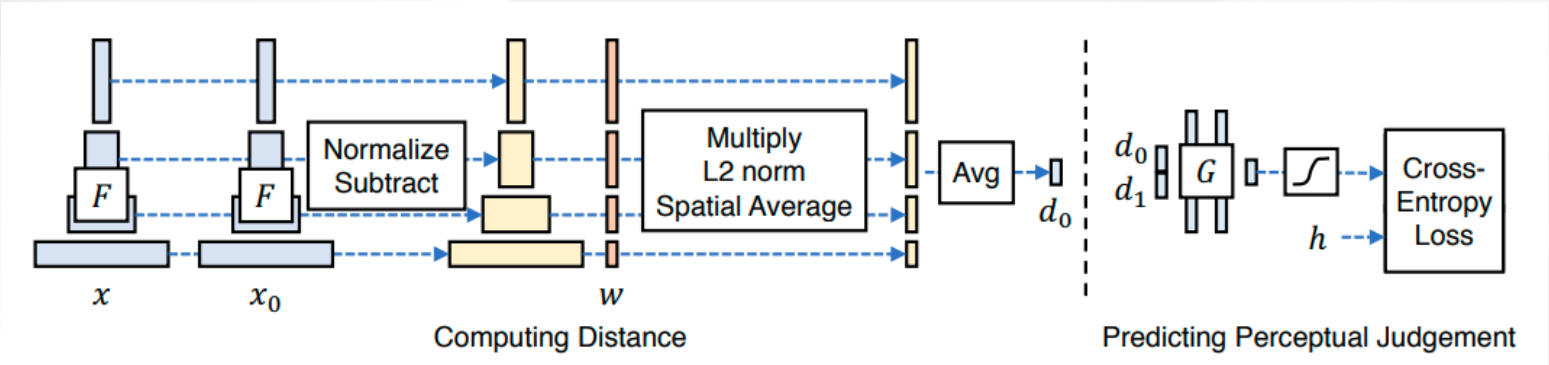
- Inspired and based heavily on the paper "THE GEOMETRY OF DEEP GENERATIVE IMAGE MODELS AND ITS APPLICATIONS - Binxu Wang, Carlos R. Ponce" (ICLR 21).
- The approach is based on the observation that Interpretable directions in the latent space should correspond to as large perceptual changes of the generated images as possible.
- **Method in nutshell:**
  Compute top-k eigenvectors (corresponding to the largest eigenvalues) of the Hessian of LPIPS model, computed with respect to the generator's parameters.

LPIPS model



- Berkeley-Adobe Perceptual Patch Similarity (BAPPS) Dataset
- **Learned Perceptual Image Patch Similarity**

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$



Spectrum-based approach - In more details

Given $d(\cdot,\cdot)$ - LPIPS model

Consider $\mathbb{E}_z d^2(G_\theta(z), G_{\theta+\alpha}(z)))$

$$\mathbb{E}_z d^2(G_\theta(z), G_{\theta+\delta\alpha}(z))) =$$

$$\mathbb{E}_z d^2(G_\theta(z), G_\theta(z))) + \frac{\partial d^2(G_\theta(z), G_{\theta+\alpha})}{\partial \alpha}|_{\alpha=0} \cdot \delta\alpha +$$

$$\delta\alpha^T \cdot \frac{\partial d^2(G_\theta(z), G_{\theta+\alpha}(z))}{\partial \alpha^2}|_{\alpha=0} \cdot \delta\alpha + \hat{o}(||\delta\alpha||_2^2)$$

The first two terms are equal to zero since d^2 achieves its global minimum at $\alpha = 0$. Thus, we focus on the eighenvectors of the Hessian.

Spectrum-based approach

The method comes down to finding the eigenvectors of the Hessian of the LPIPS model with respect to the generator parameters.

$$H = \delta\alpha^T \cdot \frac{\partial d^2(G_\theta(z), G_{\theta+\alpha}(z))}{\partial \alpha^2}|_{\alpha=0}$$

For efficient computation, define $g(a) = \frac{\partial \mathbb{E}_z d^2(G_\theta(z), G_{\theta+\alpha}(z))}{\partial \alpha}|_{\alpha=0}$.
Sample $v \sim \mathcal{N}(0, \mathcal{I})$ and iterativly update:

$$v \rightarrow \frac{g(\epsilon v) - g(-\epsilon v)}{2\epsilon||v||}; \quad \epsilon = 0.1$$

This process convergence to the leading eigenvector of H. Repeat s.t in k-th step restrict α to k-1 found eigenvectors orthogonal complementary.

Hybrid

Combine optimization-based and spectrum-based approaches

- First, compute the top-k eigenvectors of the LPIPS Hessian $v_1, \ldots, v_k$
- Then solve the optimization problem considering only the shifts ξ applied to $v_1, \ldots, v_k$

Informally, optimization based approach that operates in the parameter subspace that captures the maximal perceptual differences in the generated images.

Inspecting directions

These K directions are then inspected manually by observing the image sequences for several latent codes z

$$\{G_{\theta+t\xi_k}(z) \mid t[T, T]\}$$

Since K typically small (e.g 64), this procedure takes only several minutes for a single person.

Experiments Settings

- StyleGAN2 generators
- FFHQ, LSUN-Cars, LSUN-Horse, LSUN-Church datasets.
- Optimization based -
  ResNet18 for Reconstructor, downscale input to 256x256
  K=64, $\lambda = 2.5 \cdot 10^{-3}$, T=3500, batch-size=32
- Spectrum based -
  K=64, mini-batch=512, 10 iterations for each eigenvector.
- Hybrid -
  same as optimization based but T=80, batch-size=16

Experiments

- Belly size +

- Moving legs +

- Head size +

- Roof height +

- Trees +

- Thickness +

- Glasses size +

- Eyes distance +

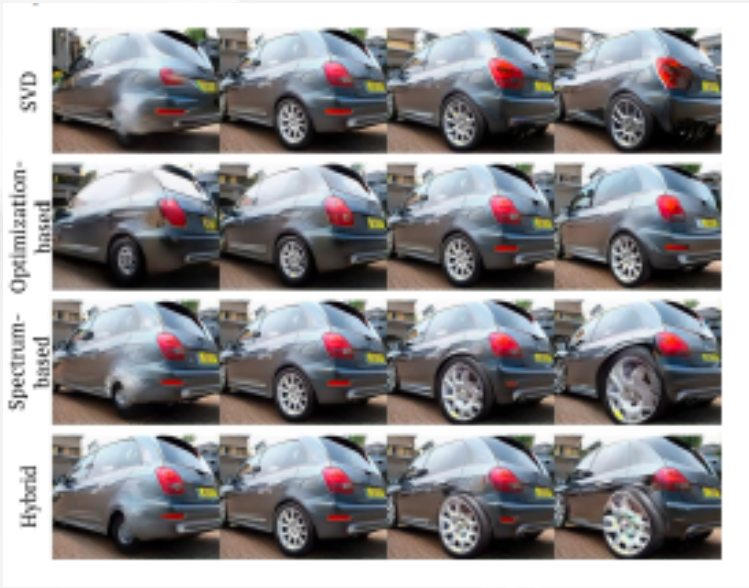
- Ears +

- Cabin size +

- Width +

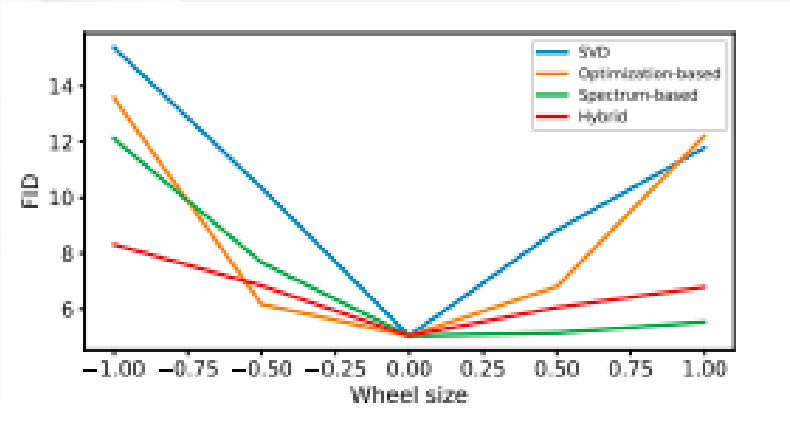
- Wheels rotation +

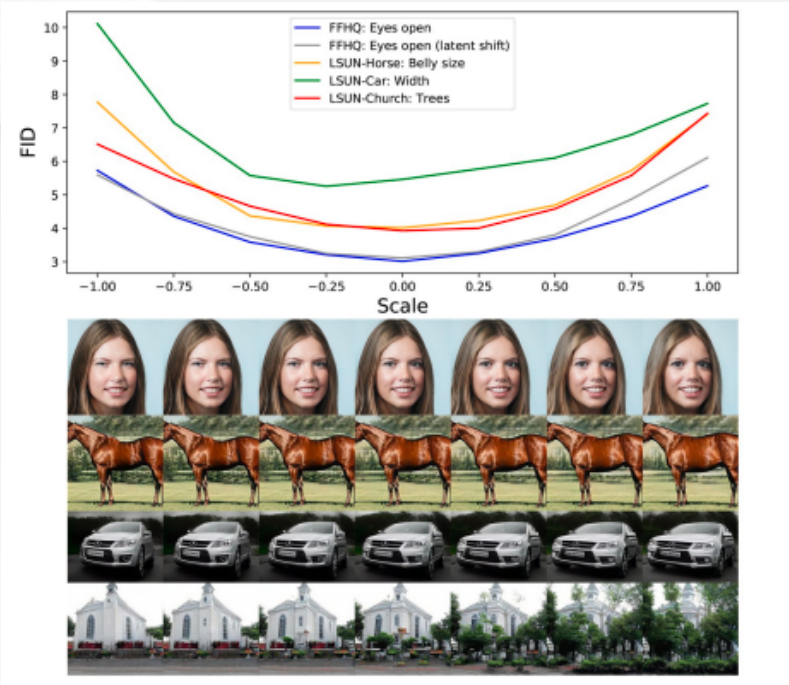
Comparing Methods



quantitative comparison

Consider a direction that corresponds to the "Wheel size" visual effect.
Plot FID curves by varying the shift magnitudes t. (FID computed with $5 \cdot 10^4$ real and generated images)



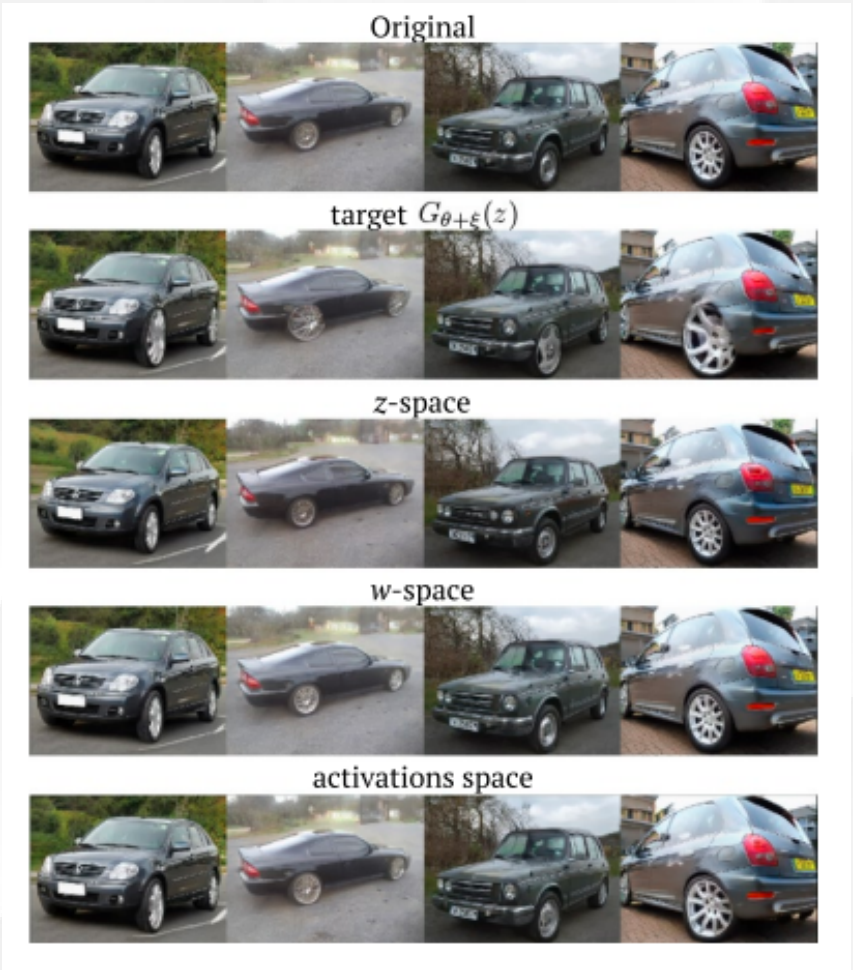


Latent transformations cannot produce these visual effects

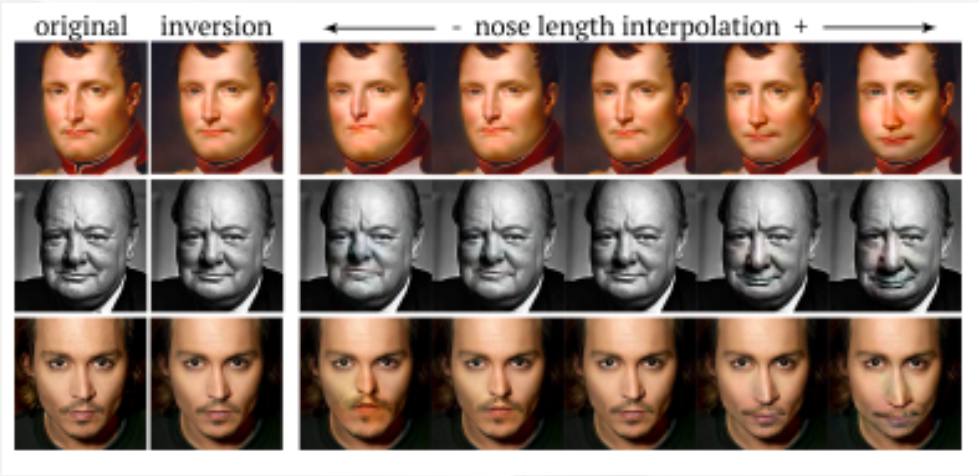
Bar-Ilan University

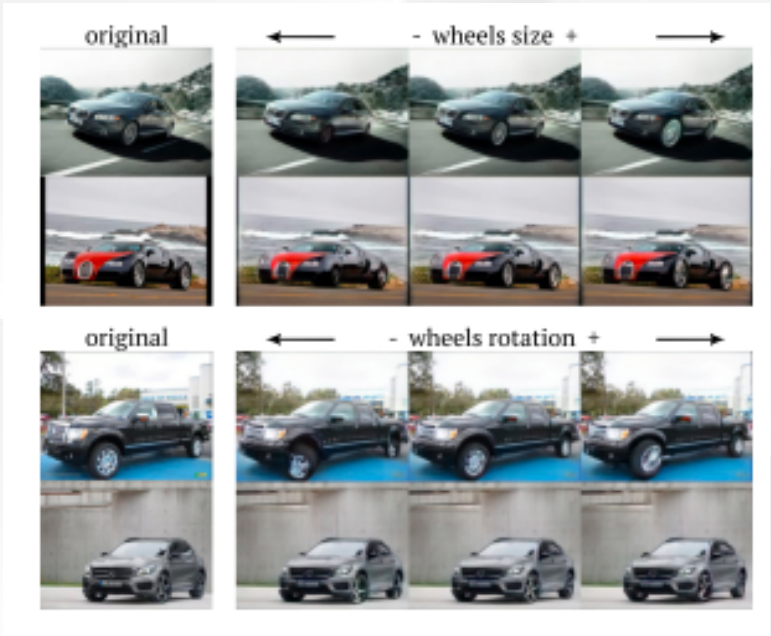
Given a shift ξ, search for latent shift h such that $G_\theta(z+h) = G_{\theta+\xi}(z)$.
Performing the following optimization:

$$\min_h \mathbb{E}_z ||G_\theta(z+h) - G_{\theta+\xi}(z)||_2^2$$



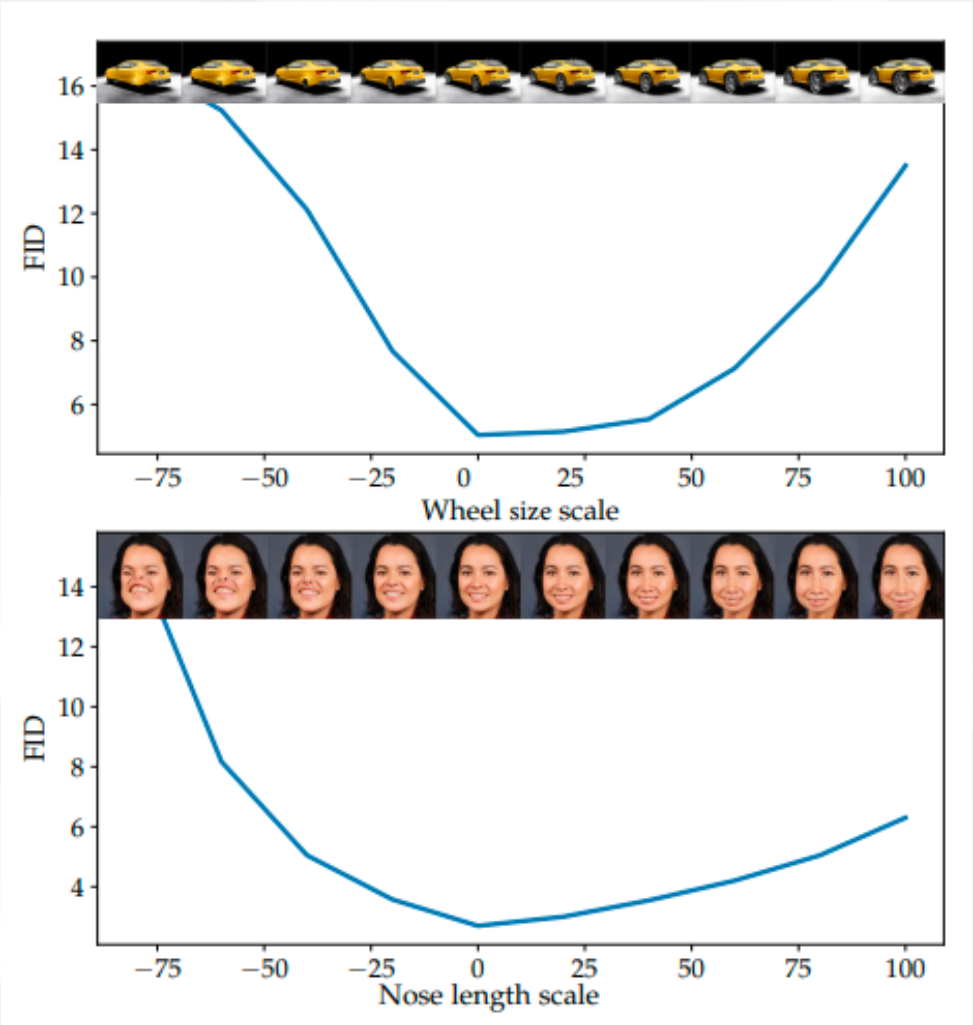
Editing real images





Maintaining realism

Even under extreme shift magnitudes, the manipulated samples have high visual quality.
Observation: FID plot is not symmetric.



Locality of visual effects

Compute per-pixel differences

$$||G_{\theta+t\xi}(z) - G_{\theta}(z)||_2^2$$

averaged over 1600 z-samples and 20 shift magnitues $t \in \mathcal{U}[-100, 100]$

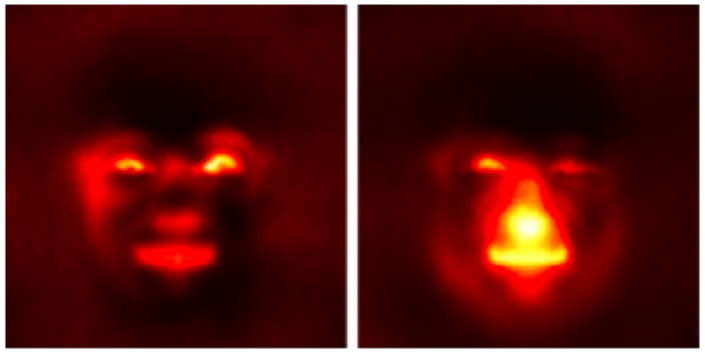


Figure 12: Averaged heatmaps of the pixel differences between the original and the edited images for the "Eyes distance" direction (*left*) and the "Nose length" direction (*right*).

Figure 13: *Left*: original image; *center*: a shift in the direction "Wheel rotation"; right: the squared distances between the original and edited image averaged over different shift magnitudes.

Questions?

Bibliography

- Navigating the GAN Parameter Space for Semantic Image Editing
  https://arxiv.org/pdf/2011.13786.pdf
- The Geometry of Deep Generative Image Models and its Applications
  https://arxiv.org/pdf/2101.06006.pdf
- The Unreasonable Effectiveness of Deep Features as a Perceptual Metric
  https://arxiv.org/pdf/1801.03924.pdf
- Unsupervised Discovery of Interpretable Directions in the GAN Latent Space
  https://arxiv.org/pdf/2002.03754.pdf
- Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network
  https://arxiv.org/pdf/1609.04802.pdf
- Semantic Photo Manipulation with a Generative Image Prior
  https://arxiv.org/pdf/2005.07727.pdf
- Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks
  https://arxiv.org/pdf/1604.04382.pdf
- Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation
  https://arxiv.org/pdf/2003.13659.pdf
- DRIT++: Diverse Image-to-Image Translation via Disentangled Representations
  https://arxiv.org/pdf/1905.01270.pdf