# StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation

Zongze Wu, Dani Linchinski, and Eli Shechtman

Hebrew University, Adobe Research

*CVPR 2021 (Oral)*

*Jun 28, 2021*

*Davian Vision Seminar*

# Backgrounds Disentangled representations for style

## Architecture of StyleGAN/StyleGAN2

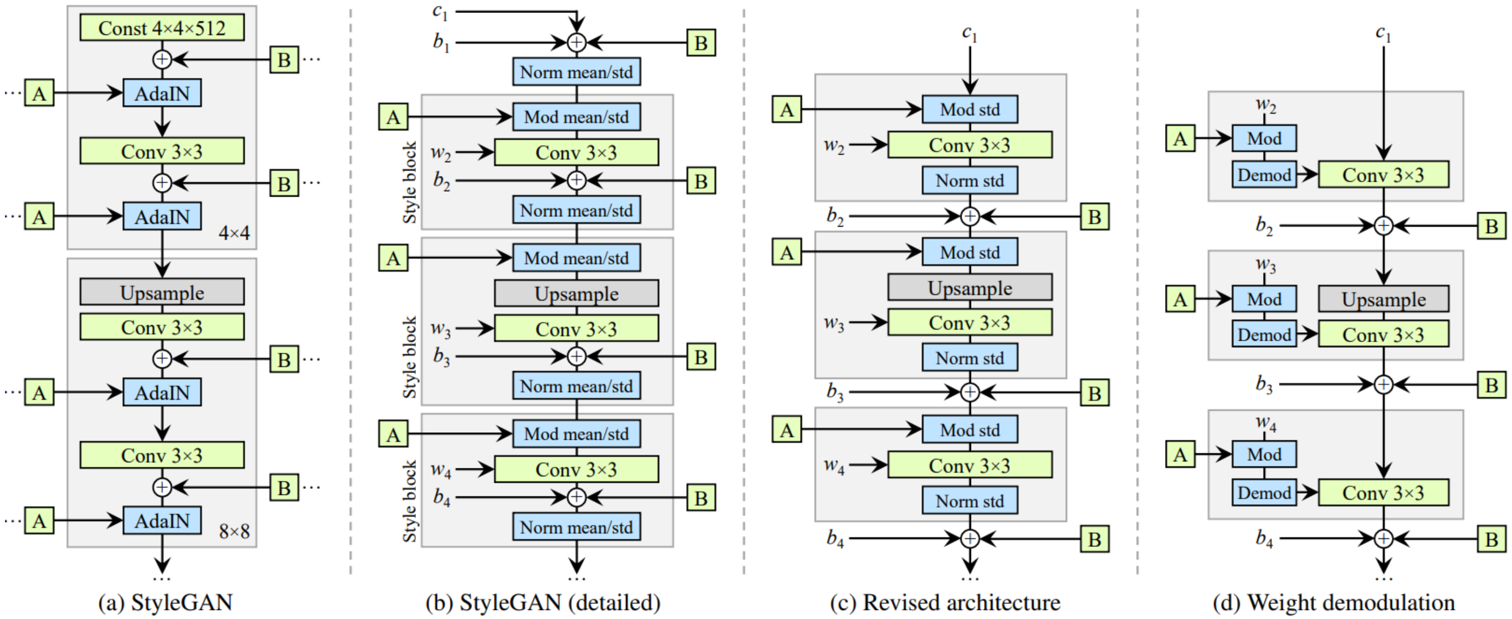
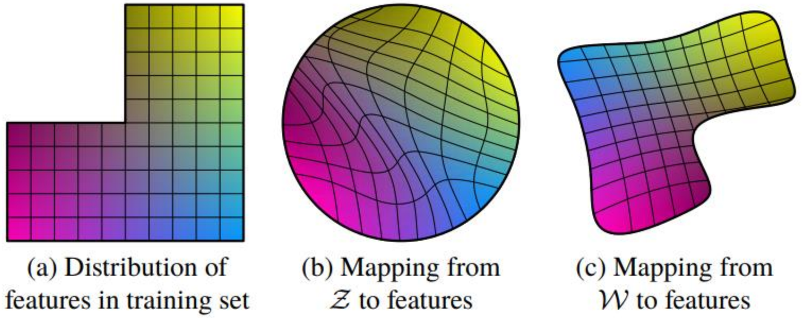


Figure 2. We redesign the architecture of the StyleGAN synthesis network. (a) The original StyleGAN, where $\boxed{A}$ denotes a learned affine transform from $\mathcal{W}$ that produces a style and $\boxed{B}$ is a noise broadcast operation. (b) The same diagram with full detail. Here we have broken the AdaIN to explicit normalization followed by modulation, both operating on the mean and standard deviation per feature map. We have also annotated the learned weights ($w$), biases ($b$), and constant input ($c$), and redrawn the gray boxes so that one style is active per box. The activation function (leaky ReLU) is always applied right after adding the bias. (c) We make several changes to the original architecture that are justified in the main text. We remove some redundant operations at the beginning, move the addition of $b$ and $\boxed{B}$ to be outside active area of a style, and adjust only the standard deviation per feature map. (d) The revised architecture enables us to replace instance normalization with a "demodulation" operation, which we apply to the weights associated with each convolution layer.

## Disentangled style representation



(a) Distribution of features in training set

(b) Mapping from $\mathcal{Z}$ to features

(c) Mapping from $\mathcal{W}$ to features

Figure 6. Illustrative example with two factors of variation (image features, e.g., masculinity and hair length). (a) An example training set where some combination (e.g., long haired males) is missing. (b) This forces the mapping from $\mathcal{Z}$ to image features to become curved so that the forbidden combination disappears in $\mathcal{Z}$ to prevent the sampling of invalid combinations. (c) The learned mapping from $\mathcal{Z}$ to $\mathcal{W}$ is able to "undo" much of the warping.

# Backgrounds Limitations of existing debiasing approaches.

**Control over synthesized outputs**

- Recent studies [1,2,3,4,5,6] have tackled this image manipulation in the generative tasks.

- However, they 1) *require annotated data,* 2) *pretrained classifier,* or *3) a large number of paired examples.*

- Furthermore, the individual controls of these methods are typically entangled and are often non-local.

**Control by utilizing the disentangled representations**

- It is important to understand how to find the disentangled controls for synthesizing the images.

# Contributions

- This paper first proposes to apply the quantitative evaluation on the *disentanglement* across latent spaces.

- Based on this comparisons, the authors propose to detect *StyleSpace* channels that control the appearance of local semantic regions in the image.

- Next, they identify the style channels that control a specific target attribute, which leads the manipulation of images in a *disentangled* manner.

- This paper proposes Attribute Discovery (AD) for measuring how manipulating a target attribute affects other attributes.

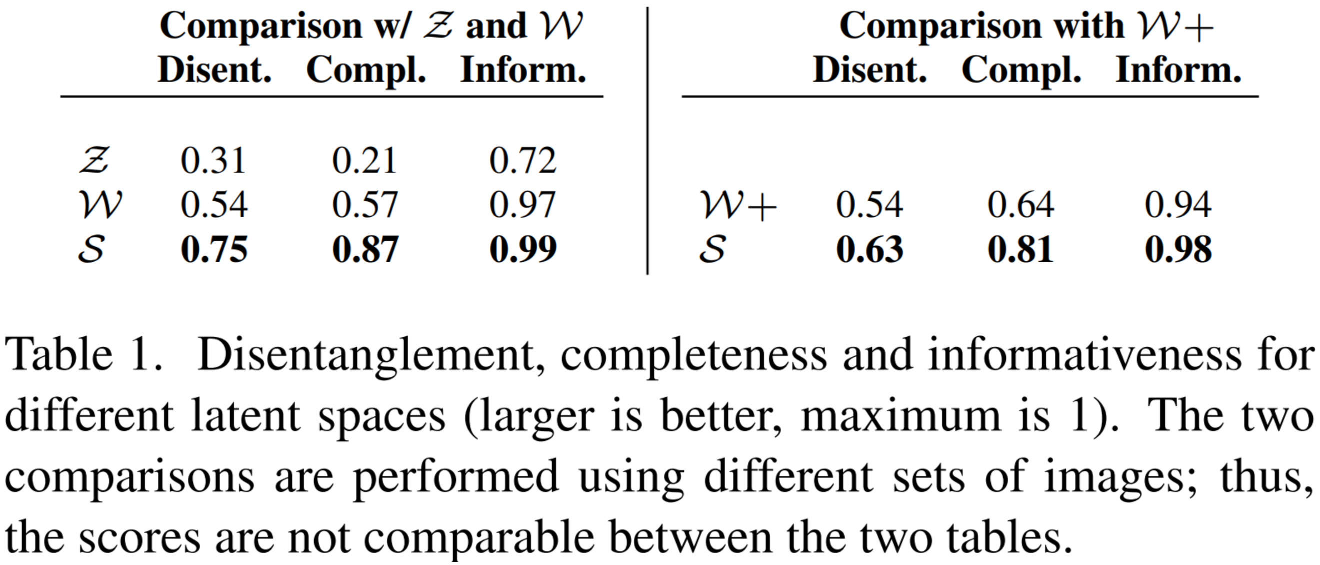
- Qualitative as well as quantitative results demonstrate that this paper presents a plausible disentangled image manipulation performance against existing baselines on the real-world datasets.

# Motivations

Evaluation on the disentanglement of latent spaces

As Eastwood *et al.*[7] proposed, the authors present the DCI metrics on four different latent spaces, i.e., $Z, W, S, and\ W+$.

- **Disentanglement** measures the degree to which each latent dimension captures at most one attribute.
- **Completeness** measures the degree to which each attribute is controlled by at most one latent dimension.
- **Informativeness** measures the classification accuracy of the attributes, given the latent representation.



| | **Comparison w/ $\mathcal{Z}$ and $\mathcal{W}$** | | |
|---|---|---|---|
| | **Disent.** | **Compl.** | **Inform.** |
| $\mathcal{Z}$ | 0.31 | 0.21 | 0.72 |
| $\mathcal{W}$ | 0.54 | 0.57 | 0.97 |
| $\mathcal{S}$ | **0.75** | **0.87** | **0.99** |

| | **Comparison with $\mathcal{W}+$** | | |
|---|---|---|---|
| | **Disent.** | **Compl.** | **Inform.** |
| $\mathcal{W}+$ | 0.54 | 0.64 | 0.94 |
| $\mathcal{S}$ | **0.63** | **0.81** | **0.98** |

Table 1. Disentanglement, completeness and informativeness for different latent spaces (larger is better, maximum is 1). The two comparisons are performed using different sets of images; thus, the scores are not comparable between the two tables.

# Proposed Method

Detecting locally-active style channels

Detection StyleSpace channels that control visual appearance of local semantic regions.

1) First, examine the gradient maps of generated images w.r.t different channels.
2) Then, measure the overlaps with specific semantic regions.

$$OC^s_{u,c} = \frac{|(G^s_u > t^s_{u,c}) \cap M^s_c|}{|M^s_c|^d},$$

$$c^*_{s,u} = \arg\max OC^s_{u,c}.$$

- $s \in S$ : a style code
- $u$ : channel index
- $c$ : semantic category
- $G^s_u$ : gradient map of generated image w.r.t each channel of $s$
- $t^s_u$ : threshold to make $(G^s_u > t^s_u)$ have the same size as $M^s_c$
- $M^s_c$ : semantic regions for class $c$
- $d$ : correction factor

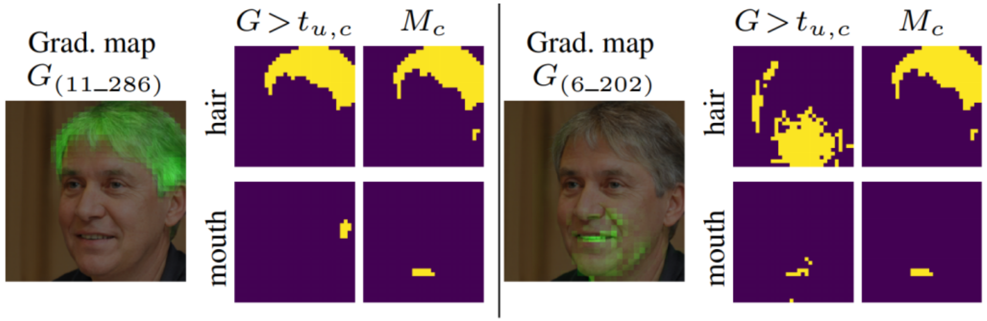


Figure 2. A gradient map with respect to each style channel $u$, e.g., (11_286), channel 286 of generator level 11, is thresholded against a category-specific threshold, chosen such that the resulting mask has the same size as the semantic mask $M_c$. The gradient mask of (11_286) has large overlap with the mask for hair, and no overlap with the mouth, while that of (6_202) has large overlap with the mask for mouth and almost none with the hair.

# Experiments

## Detecting locally-active style channels

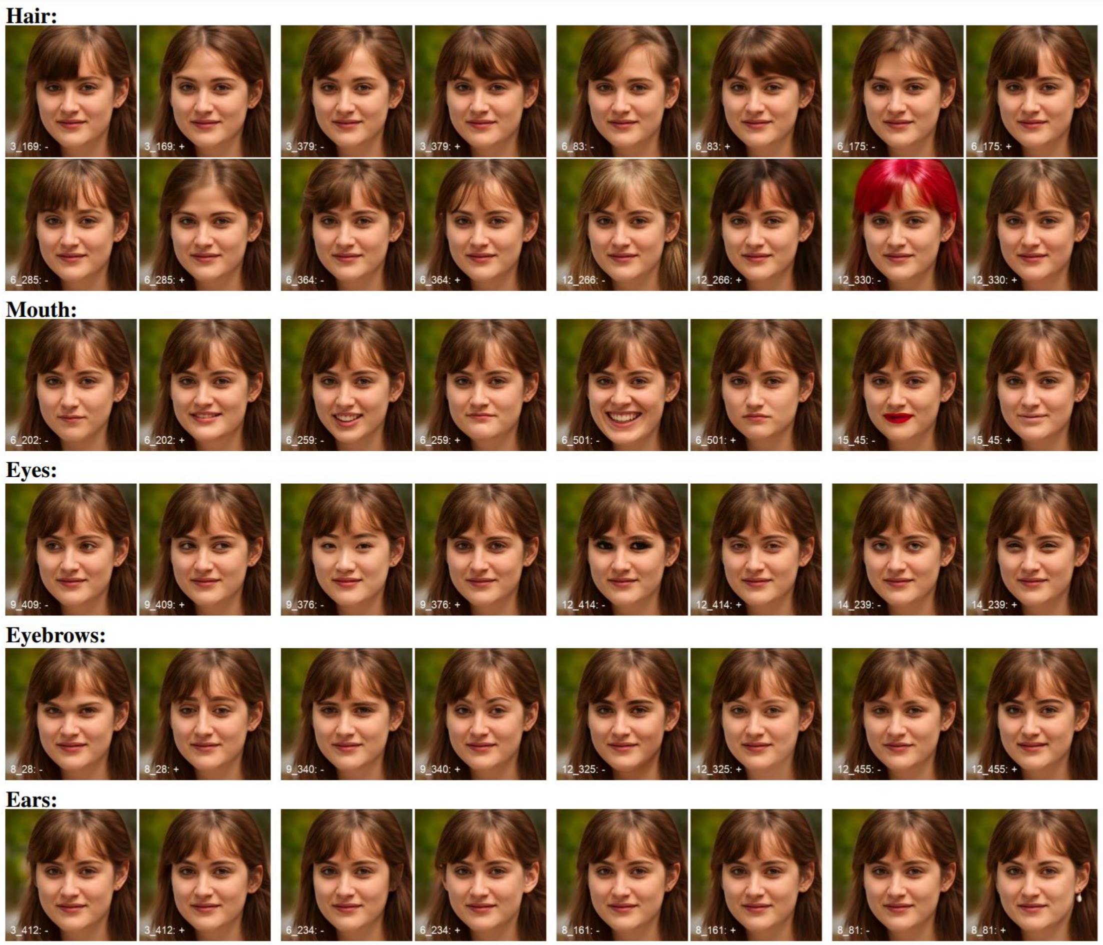


Figure 3. Examples of manipulations, each controlled by a single style channel. Each pair of images shows the result of manipulation by decreasing (-) and increasing (+) the value of the style parameter (the original image is omitted). The layer index, channel index, and the direction of change is overlayed in the bottom left corner.

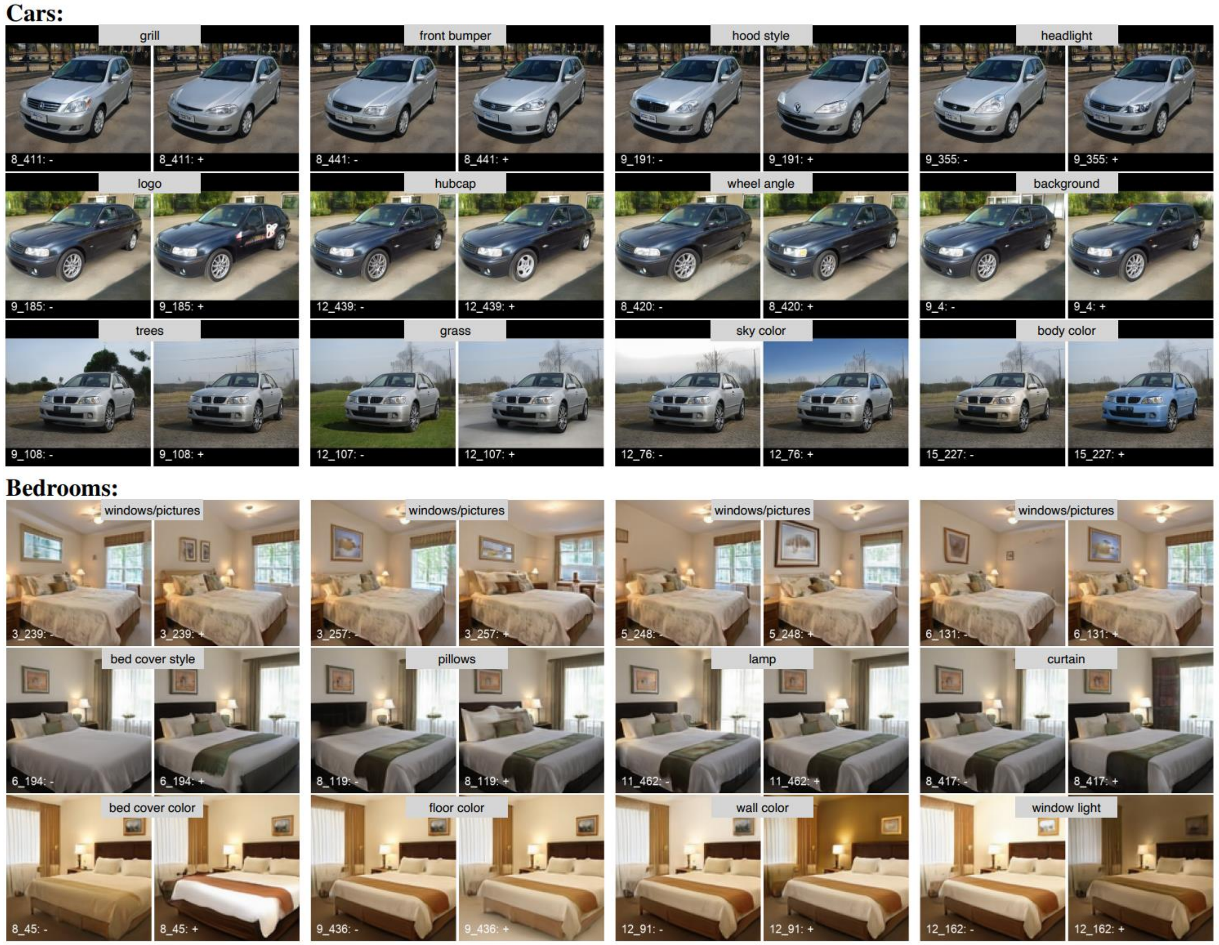


Figure 4. Examples of manipulations, each controlled by a single style dimension. Each pair of images shows the result of manipulation by decreasing (-) and increasing (+) the value of the style parameter (the original image is omitted). The layer index, channel index, and the direction of change is overlayed in the bottom left corner.

# Proposed Method Detecting attribute-specific channels

Detection StyleSpace channels that control a specific *target attribute*.

1) First, it is required to collect only 10-30 positive exemplars which include target attributes.
2) Detect the channels which most deviate from the statistics of overall images.

Let $\mu^p$ and $\sigma^p$ denote the mean and std of the style vectors over the generated distribution. Give the style vector $s^e$ of a specific positive example, we compute

$$\delta^e = \frac{s^e - \mu^p}{\sigma^p},$$

which denotes a normalized difference from the population mean.

Then, calculate the $\mu^e$ and $\sigma^e$ of the differences $\delta^e$ over the exemplar set and the relevance of $u$ w.r.t target attribute as the ratio

$$\theta_u = \frac{|\mu_u^e|}{\sigma_u^e}.$$

# Proposed Method Detecting attribute-specific channels

- The authors first use a large number (1K) of positive examples to verity the proposed method and confirm that *16 out of 26 attributes are controlled by at least one single style channel*.

- In case of 10, 20, and 30 positive examples, top 5 channels detected from these small number of images *include the channels found from above verification with high accuracy*.

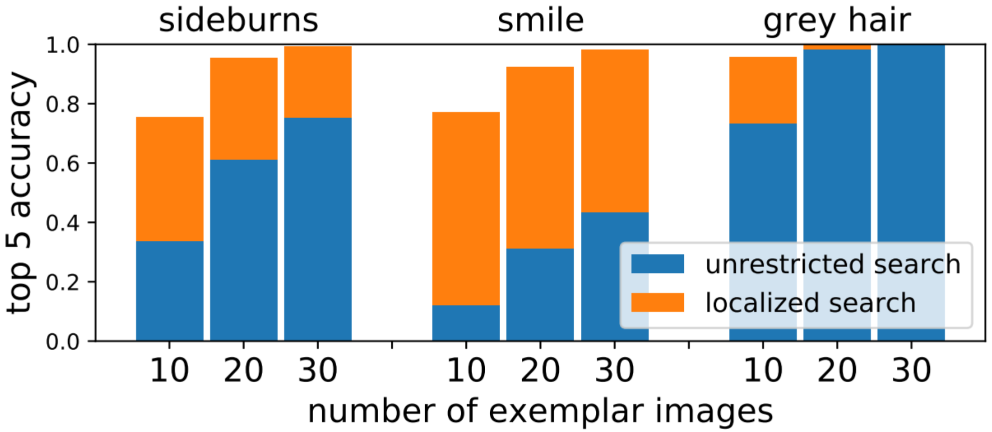


Figure 5. Top-5 detection accuracy for attribute-specific controls (for three target attributes) using 10, 20, or 30 positive examples.

# Experiments

Disentangled attribute manipulation

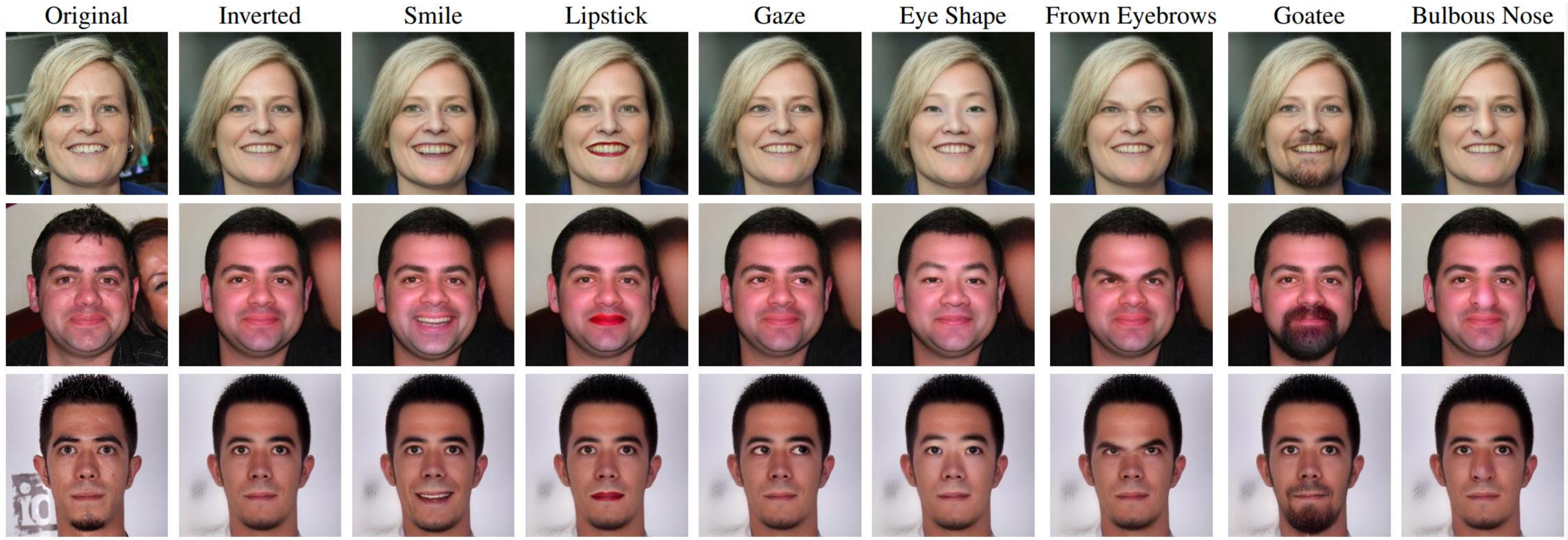


Figure 7. Manipulation of real images using encoder-based inversion. Original images are from FFHQ, and were not part of the encoder's training set. More results can be found in supplementary Figure 19.

# Experiments

Comparisons against existing baselines

### Qualitative comparisons

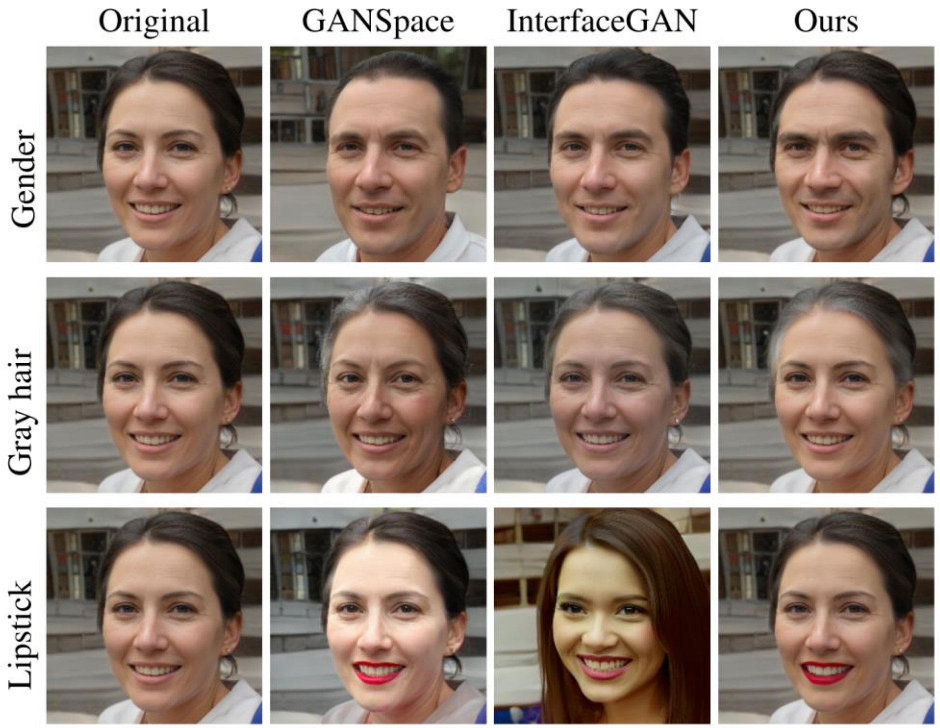


Figure 6. Comparison with state-of-the-art methods using the same amount of manipulation $\Delta l_t = 1.5\sigma(l_t)$.

### Attribute Dependency

- Images without the target attribute $t$ are manipulated towards $t$ by a certain amount measured by the changes of logits $\Delta l_t$ of a pretrained classifier.
- Next, measure the $\Delta l_i$ for other attributes $i \in A \backslash \mathrm{t}$, where $A$ is a set of all attributes.

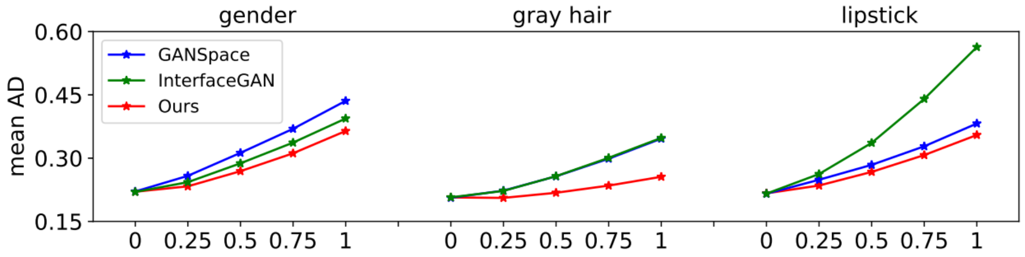


Figure 8. Mean-AD vs. the degree of target attribute manipulation $(\Delta l_t/\sigma(l_t))$. Lower mean-AD indicates better disentanglement.

# Additional Results

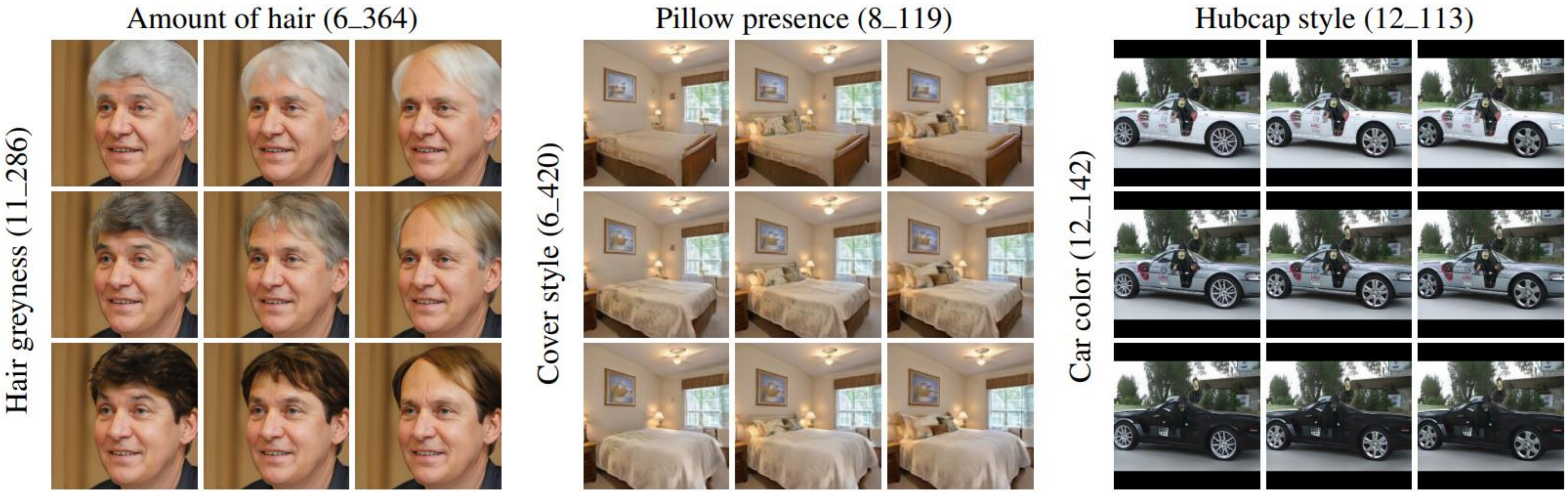


Figure 1. Disentanglement in style space, demonstrated using three different datasets. Each of the three groups above shows two manipulations that occur independently inside the same semantic region (hair, bed, and car, from left to right). The indices of the manipulated layer and channel are indicated in parentheses.

# References

[1] Mirza *et al.,* Conditional generative adversarial nets, Arxiv, 2014

[2] Goetschalckx *et al.,* GANalyze: toward visual definitions of cognitive image properties, ICCV, 2019

[3] Shen *et al.,* Interpreting the latent space of GANs for semantic face editing., CVPR, 2020

[4] Shen *et al.,* InterFaceGAN: interpreting the disentangled face representation learned by GANs., Arxiv, 2020

[5] Xu *et al.,* Generative hierarchical features from synthesizing images, Arxiv, 2020

[6] Jahanian *et al.,* On the “steerability” of generative adversarial networks, Arxiv, 2019

[7] Eastwood *et al.,* A framework for the quantitative evaluation of disentangled representations, ICLR, 2018

# Thank you