Discovering Interpretable Directions in the Semantic Latent Space of Diffusion Models

René Haas[1]
Inbar Huberman-Spiegelglas[2] • Rotem Mulayoff[2] • Stella Graßhof[1] • Sami S. Brandt[1] • Tomer Michaeli[2]
[1] IT University of Copenhagen, Denmark • [2] Technion, Israel

Background

Problem:
The latent space of diffusion models is not yet well understood.

Previous work: Kwon et al. [14]
- Introduces *h*-space as semantic latent space
- Semantic directions are found using CLIP

Contributions

We propose 2 unsupervised and 3 supervised approaches for intuitive **semantically disentangled image editing without**:
- CLIP guidance
- Changes in diffusion model architecture and fine-tuning

Semantic Image Editing

Unsupervised

Global edits by PCA

Incremental PCA on bottleneck activation from images generally yields global semantic directions.

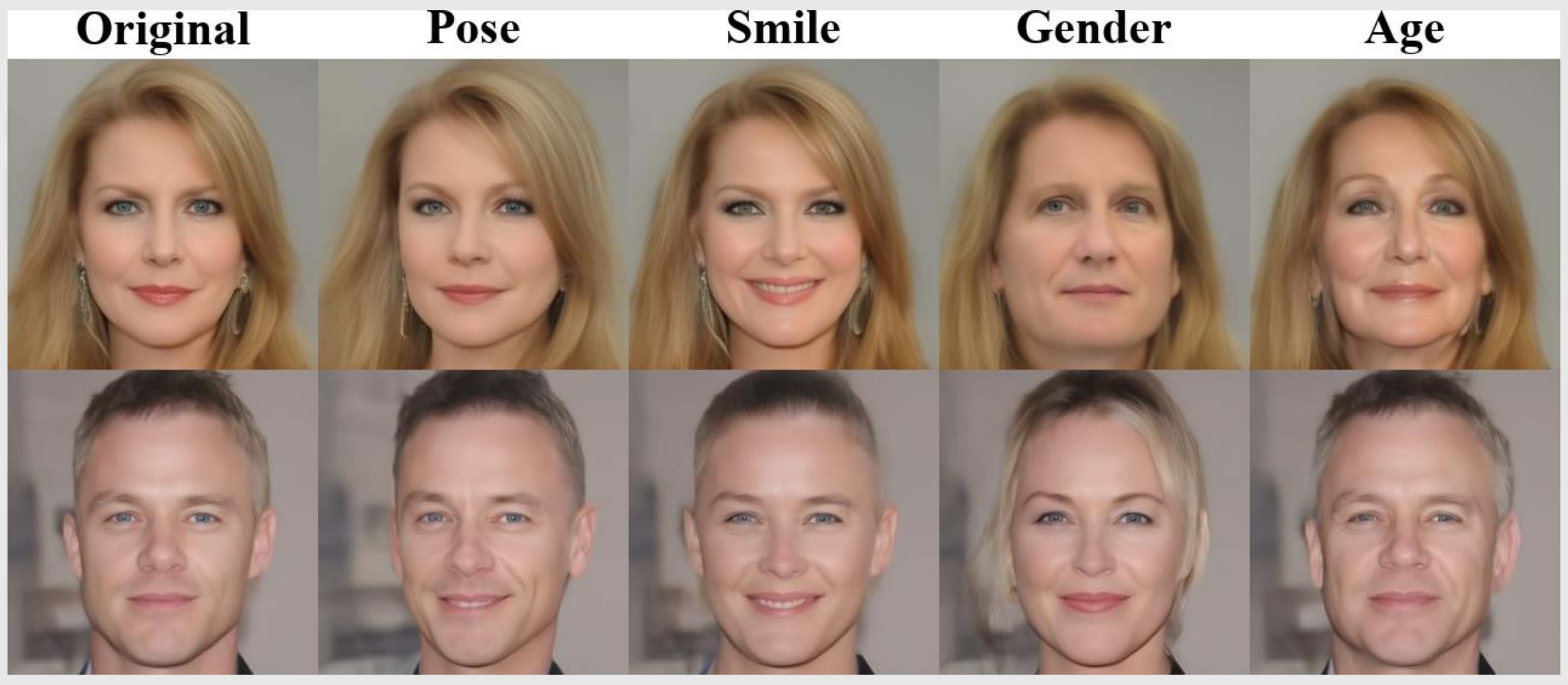


Image-specific edits

- Directions in *h*-space that change $\epsilon_t^\theta(\mathbf{x}_t)$ the most
- Eigenvectors of Jacobian of denoiser

$$\mathbf{J}_t \triangleq \frac{\partial \epsilon_t^\theta(\mathbf{x}_t, \mathbf{h}_t)}{\partial \mathbf{h}_t} = \mathbf{U}_t \mathbf{\Sigma}_t \mathbf{V}_t^{\mathrm{T}}$$

- Challenge: dimension of full bottleneck h-space.
- Solution: **power iteration** to circumvent the intractable computational cost

$$\mathbf{J}_t^{\mathrm{T}} \mathbf{J}_t \mathbf{v} = \frac{\partial}{\partial \mathbf{h}_t} \left\langle \epsilon_t^\theta(\mathbf{x}_t, \mathbf{h}_t), \mathbf{J}_t \mathbf{v} \right\rangle$$

$$\mathbf{J}_t \mathbf{v} = \frac{\partial}{\partial a} \epsilon_t^\theta(\mathbf{x}_t, \mathbf{h}_t + a\mathbf{v}) \Big|_{a=0} .$$

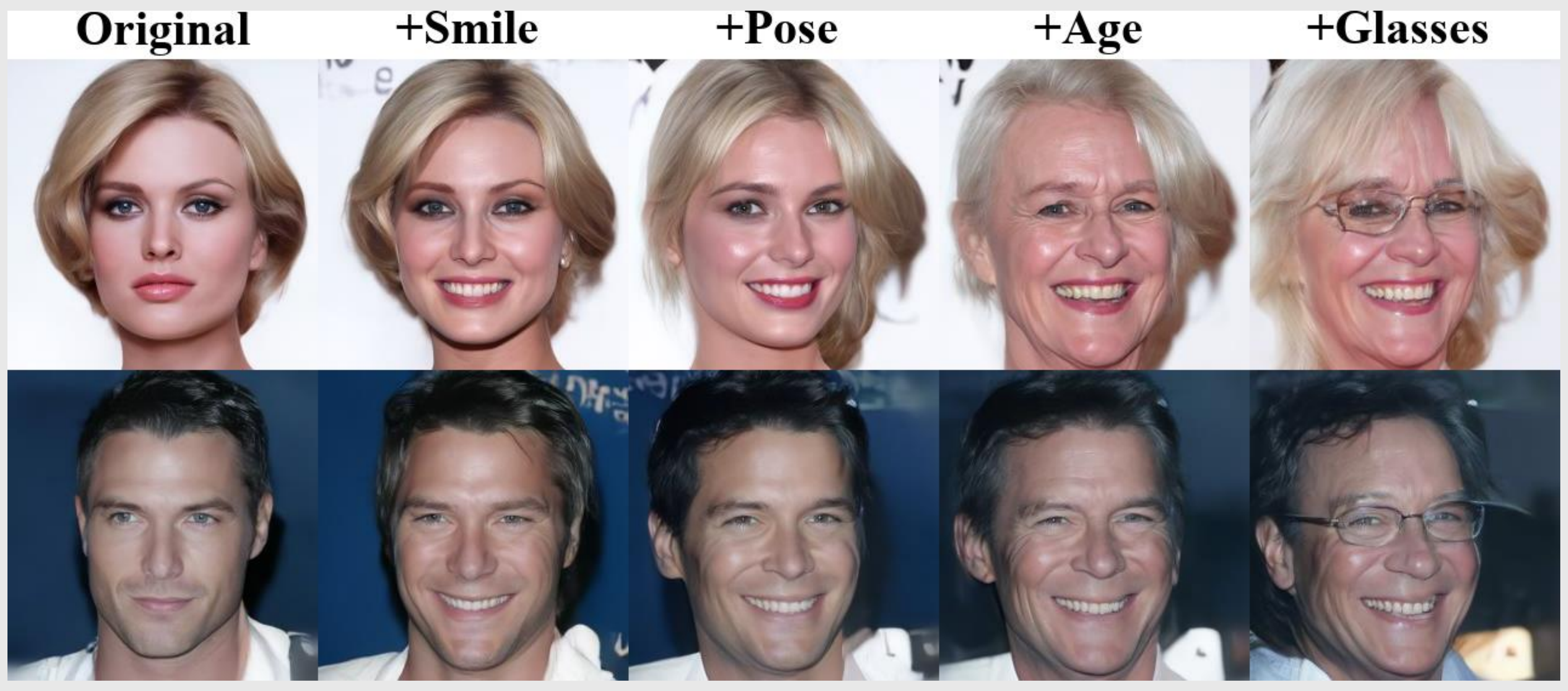
Supervised

Edits by examples

Semantic directions based on latent representation of image with (+) and without (-) binary attribute:

$$\mathbf{w} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{q}_i^+ - \mathbf{q}_i^- \right)$$

Sequential edits:

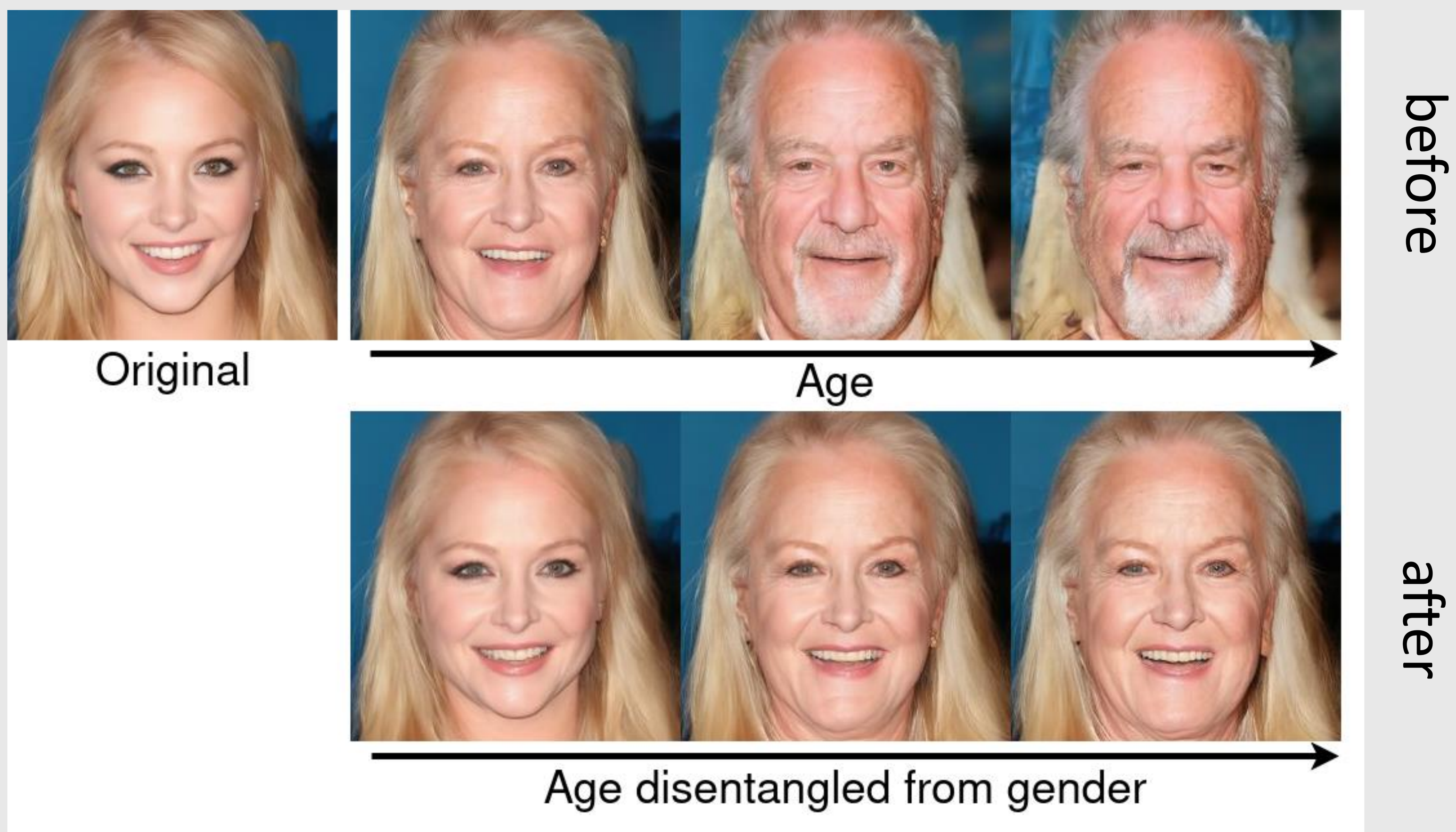


Classifier Annotation

Addition: use a pre-trained attribute classifier on the model generated images to gather positive and negative examples

Disentanglement

To remove effects of K directions $\mathbf{w}_k$ from $\mathbf{w}_0$, project $\mathbf{w}_0$ onto the orthogonal complement of $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_K]$. This yields an updated direction $\mathbf{w}_0$ disentangled from $\mathbf{w}_k$.



Project Website

https://github.com/renhaa/semantic-diffusion

IT UNIVERSITY OF COPENHAGEN

TECHNION Israel Institute of Technology