paper link:

Unsupervised Disentanglement of Linear-Encoded Facial Semantics

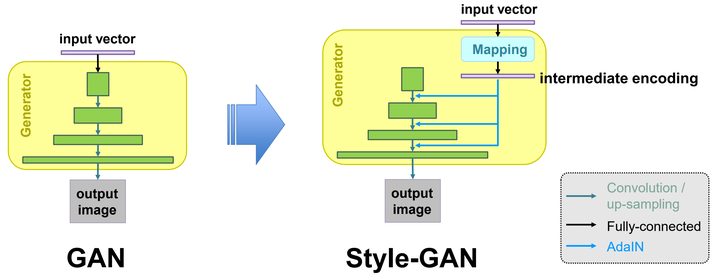
Yutong Zheng, Yu-Kai Huang, Ran Tao, Zhiqiang Shen, and Marios Savvides

{yutongzh,yukaih2,rant,zhiqians,marioss}@andrew.cmu.edu

CVPR VIRTUAL JUNE 19-25

Background

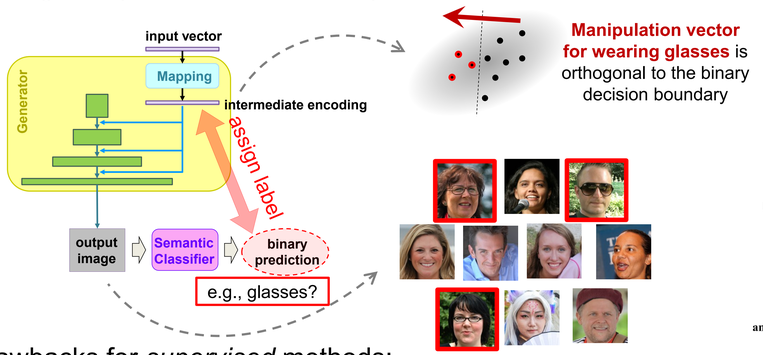
Style-GAN makes representations linear-encoded:



Intermediate encoding with AdaIN operation provides more flexibility to the representations, making linear encoding possible.

Previous Work: Manipulation Vectors

Finding manipulation vectors with pretrained semantic classifiers:

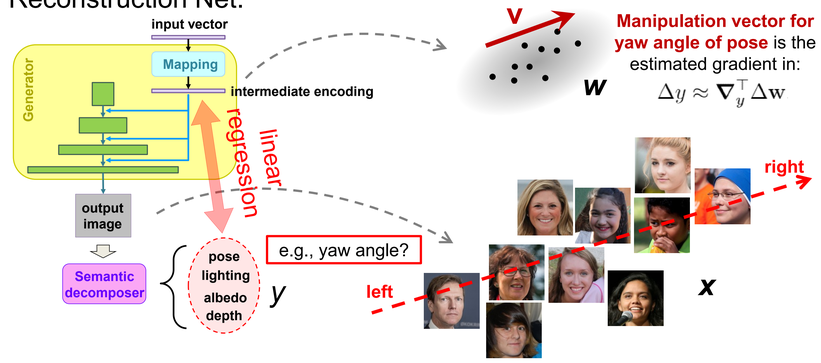


Drawbacks for *supervised* methods:

- Requires pre-training **semantic classifiers** with labeled data.
- Semantics are artificially defined, resulting in potential redundancy and incompleteness.

The Method: Going Unsupervised

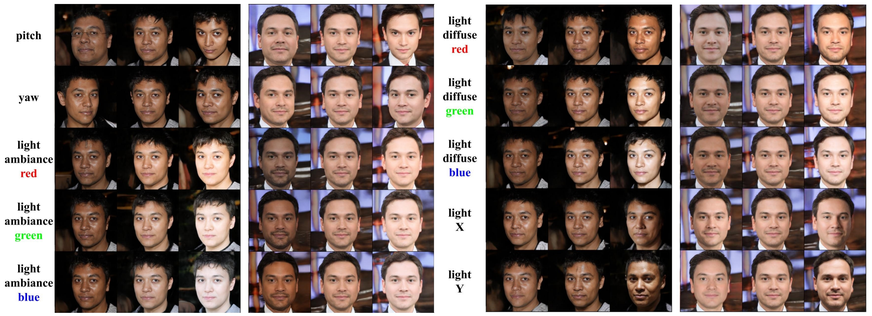
Setup: Unsupervised Pre-training of **Style-GAN generator** and **semantic decomposer** from the 3D Deformable Face Reconstruction Net.



Gradient (v) for any scaler semantic components (y): sample a bunch of $\boldsymbol{w}$ and output y, then optimize the objective:

$$\min_{\mathbf{v}} \quad \|\Delta Y - \Delta \mathbf{W}\mathbf{v}\|_2^2$$

Manipulate $\boldsymbol{v}$ associated with interpretable scaler semantic components: pose and lighting:

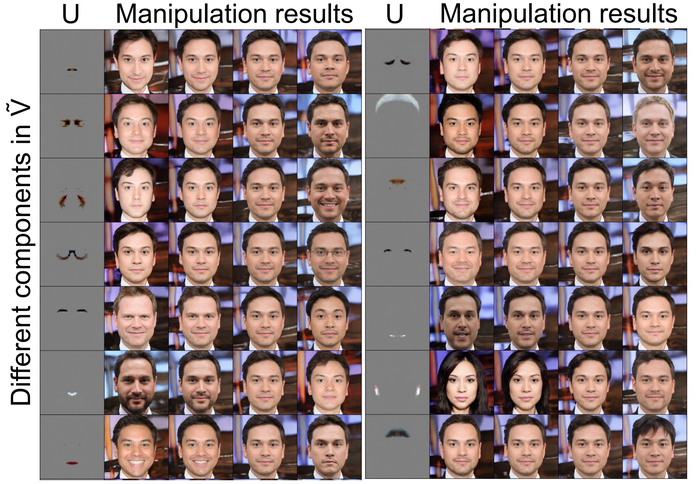


The Method: Localized Representations

Gradient ($\tilde{v}$) for pixel-level semantic components: combine $\boldsymbol{v}$ in albedo and depth from previous step as a Jacobian J_v^*, then estimate local facial variations by optimizing:

$$\min_{\mathbf{U},\hat{\mathbf{V}}} \quad \left\|\mathbf{J}_{\mathbf{v}}^* - \mathbf{U}\hat{\mathbf{V}}^\top\right\|_F^2 + \alpha\|\mathbf{U}\|_1 + \beta\sum_{i\neq j}(\hat{\mathbf{v}}_i^\top\hat{\mathbf{v}}_j)^2 \quad \text{s.t. } \|\hat{\mathbf{v}}_p\|_2 = 1,$$

sparsity orthogonality



Cosine distances of $\tilde{\boldsymbol{v}}$ comply with the contextual relations of local semantics:

Carnegie Mellon University
CyLab Biometrics Center