StyleRig : Rigging StyleGAN for 3D Control over Portrait Images

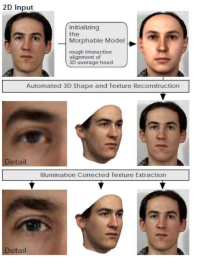
Tewari, Ayush, et al.

Presenter: Sejik Park

Contents

Introduction

- Improve StyleGAN controllability
  - control with 3 dimensional morphable face models (3DMMs)
    - ex. expressions, pose, illumination
  - 3DMMs
    - using 3D laser scan data
    - multivariate normal distribution
    - subregions (eyes, nose, mouth, surrounding)
    - markedness regression : neutral - expression
    - illumination-corrected texture extraction



- Blanz, V., & Vetter, T. (1999, July). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 187-194).

Model

- RigNet
  - Improve controllability by latent code modification
  - Model
    StyleGAN + differentiable face reconstruction (DFR) + RigNet
  - Train
    Pretrained StyleGAN + pretrained DFR + RigNet
    latent reconstruction loss + two-way cycle consistency losses
  - Data
    combining up 5 latent codes by StyleGAN

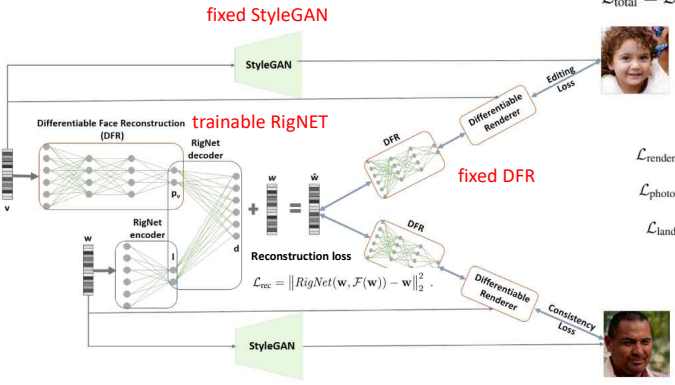
Model

- RigNet

$\mathbf{I_w}$ Image correspond to latent code w

$\mathbf{p_w}$ Semantic control parameters (input of differentiable renderer)

$\mathbf{L_{I_w}}$ Landmarks



RigNET loss

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{edit} + \mathcal{L}_{consist} \ .$$

$$\mathcal{L}_{edit} = \mathcal{L}_{render}(\mathbf{I_v}, \mathbf{p}_{edit}) \ .$$

$$\mathcal{L}_{render}(\mathbf{I_w}, \mathbf{p}) = \mathcal{L}_{photo}(\mathbf{I_w}, \mathbf{p}) + \lambda_{land}\mathcal{L}_{land}(\mathbf{I_w}, \mathbf{p}) \ .$$

$$\mathcal{L}_{photo}(\mathbf{I_w}, \mathbf{p}) = \left\|\mathbf{M} \odot (\mathbf{I_w} - \mathcal{R}(\mathbf{p}))\right\|_2^2 \ .$$

$$\mathcal{L}_{land}(\mathbf{I_w}, \mathbf{p}) = \left\|\mathbf{L}_{\mathbf{I_w}} - \mathbf{L_M}\right\|_2^2 \ ,$$

$$\mathcal{L}_{consist} = \mathcal{L}_{render}(\mathbf{I_w}, \mathbf{p}_{consist}) \ .$$

Results

- Differentiable Face Reconstruction

Figure 3: Differentiable Face Reconstruction. Visualized are (image, reconstruction) pairs. The network however, only gets the latent vector corresponding to the images as input.

Results

- Distribution of latent at different resolution and face param
  - Data : Flickr_HQ dataset
    1. less data around Z-axis
    2. neutral or smiling/laughing faces

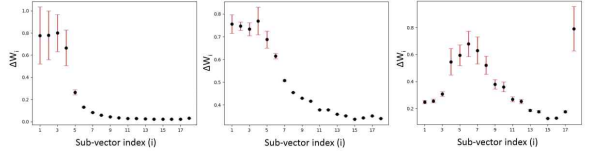


Figure 4: Change of latent vectors at different resolutions. Coarse vectors are responsible for rotation (left), medium for expressions (middle), medium and fine for illumination (right).

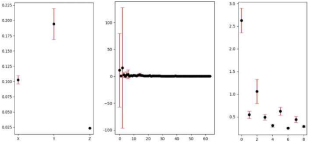


Figure 6: Distribution of face model parameters in the training data. x-axis shows the face model parameters for rotation, expression and illumination from left-right. y-axis shows the mean and variance of the parameters computed over $20k$ training samples.

Results

- Mixing source & target

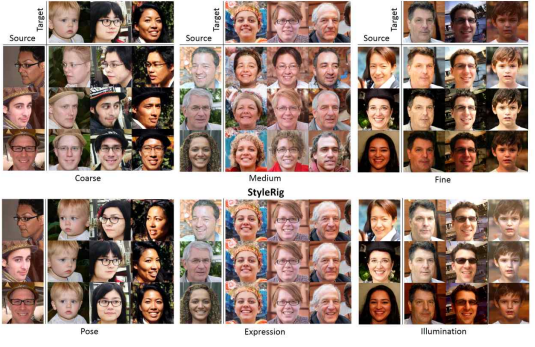


Figure 5: Mixing between source and target images generated by StyleGAN. For StyleGAN, the latent vectors of the source samples (rows) are copied to the target vectors (columns). StyleRig allows us to mix semantically meaningful parameters, i.e., head pose, expressions and scene illumination. These parameters can be copied over from the source to target images.

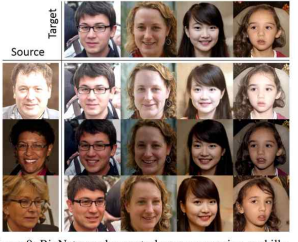
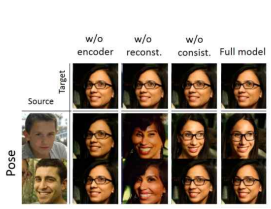
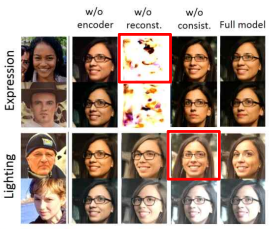


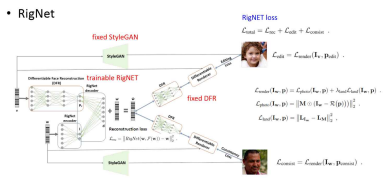
Figure 9: RigNet can also control pose, expression, and illumination parameters simultaneously. These parameters are transferred from source to target images, while the identity in the target images is preserved.

Results

- Baseline Comparison
  - Different Loss Function
    off reconstruction loss : non-face images
    off consistency : one change with other feature







Discussion

- Data
  - not able to exploit the full expressivity of the parametric face model
- Model
  - quality limitance : differentiable face reconstruction
    ex. no fine-scale detail
  - preserve parts : background, hair style
  - cf. train time (based Nvidia Volta GPU)
    - 24 hours to train StyleRig
    - more than 41 days to train StyleGAN

- End -