**A. Introduction**

- SOTA methods for neural face reenactment train generative models to learn disentangled embeddings for identity and facial pose using paired data.
- The main challenges are: a) realistic image generation, b) identity preservation and c) faithful facial pose transfer.
- We present a novel method for face reenactment leveraging the high quality generation of a pretrained StyleGAN2 and the disentangled properties of a 3D shape model.
- Our method is able to create realistic facial images, and also faithfully transfer the target head pose and expression.

**B. Preliminaries**

1. We finetune StyleGAN2 (trained on FFHQ) on VoxCeleb dataset, which is more diverse in terms of head poses and expressions compared to FFHQ dataset.

2. We use a 3D shape model [2] to extract the 3D facial model \( s \) and the facial pose parameter \( p \) defined as:
   \[ s = s_i + S_p p_i + S_e p_e, \quad p = [p_i, p_e] \]
   where \( p_i \), \( p_e \) are the identity and expression coefficients, and \( p_p \) the head orientation.

**C. Goal**

Learn the directions in the latent space of StyleGAN2 that control different facial attributes without altering the identity of the generated face.

We propose to associate a change \( \Delta p \) in the parameter space, with a change \( \Delta w \) in the intermediate latent space \( W^+ \).

**D. Our method**

- We train the matrix of directions \( A \), which takes as input the difference of facial pose parameters \( \Delta p \), and outputs a shift vector \( \Delta w \).
- The reenacted image is generated by shifting the source latent code using the predicted shift \( \Delta w \).

**E. Inference**

Given a source face and a target video:
1. We invert the source image to get the source latent code \( w_s \).
2. We finetune the generator to get a better reconstruction result [3].
3. We reenact the source face given a target pose.

**F. Quantitative Results**

<table>
<thead>
<tr>
<th>Method</th>
<th>CSIM</th>
<th>LPPR</th>
<th>PFD</th>
<th>FVD</th>
<th>NME</th>
<th>Pose</th>
<th>Exp.</th>
<th>CSIM</th>
<th>Pose</th>
<th>Exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>XFace [8]</td>
<td>0.70</td>
<td>0.33</td>
<td>30.5</td>
<td>35.5</td>
<td>12.5</td>
<td>1.1</td>
<td>0.98</td>
<td>0.80</td>
<td>1.5</td>
<td>1.3</td>
</tr>
<tr>
<td>FFoMM [5]</td>
<td>0.65</td>
<td>0.14</td>
<td>36.6</td>
<td>40.2</td>
<td>34.1</td>
<td>1.3</td>
<td>0.98</td>
<td>0.80</td>
<td>1.5</td>
<td>1.3</td>
</tr>
<tr>
<td>Face2Blender [6]</td>
<td>0.64</td>
<td>0.22</td>
<td>32.8</td>
<td>36.4</td>
<td>13.2</td>
<td>1.1</td>
<td>0.85</td>
<td>0.81</td>
<td>1.5</td>
<td>1.3</td>
</tr>
<tr>
<td>Neural-Head [7]</td>
<td>0.49</td>
<td>0.22</td>
<td>98.4</td>
<td>78.8</td>
<td>15.5</td>
<td>1.3</td>
<td>0.90</td>
<td>0.86</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>LSR [8]</td>
<td>0.59</td>
<td>0.13</td>
<td>45.7</td>
<td>40.4</td>
<td>17.8</td>
<td>1.0</td>
<td>0.75</td>
<td>0.50</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>PBR [9]</td>
<td>0.71</td>
<td>0.12</td>
<td>57.2</td>
<td>41.4</td>
<td>18.2</td>
<td>1.96</td>
<td>0.94</td>
<td>0.62</td>
<td>2.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Ours</td>
<td>0.66</td>
<td>0.11</td>
<td>35.0</td>
<td>34.5</td>
<td>14.4</td>
<td>1.1</td>
<td>0.68</td>
<td>0.63</td>
<td>1.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**G. Qualitative Results (I)**

Self reenactment: Source and target images have the same identity.

**G. Qualitative Results (II)**

Cross-subject reenactment: Source and target images have different identities.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>smile</td>
<td>yaw</td>
<td>pitch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Facial image editing: Only one facial attribute (yaw, pitch, smile etc.) is edited, without altering the identity and any other attribute of the source face (shown inside the red box).

---

[1] The et al., Designing an encoder for stylegan image manipulation. ACM TOG, 2021
[2] Feng et al., Learning an untangled detailed 3D face model from in-the-wild images. ACM TOG, 2021
[5] Feng et al., Learning an untangled detailed 3D face model from in-the-wild images. ACM TOG, 2021
[8] Calhoun et al., Fast-to-fine neural synthesis of high-resolution head avatars. ECCV, 2020
[9] Burkov et al., Neural head reenactment with latent pose descriptors. CVPR, 2020
[10] Menon et al., Learned spatial representations for few-shot talking head synthesis. ICCV, 2021