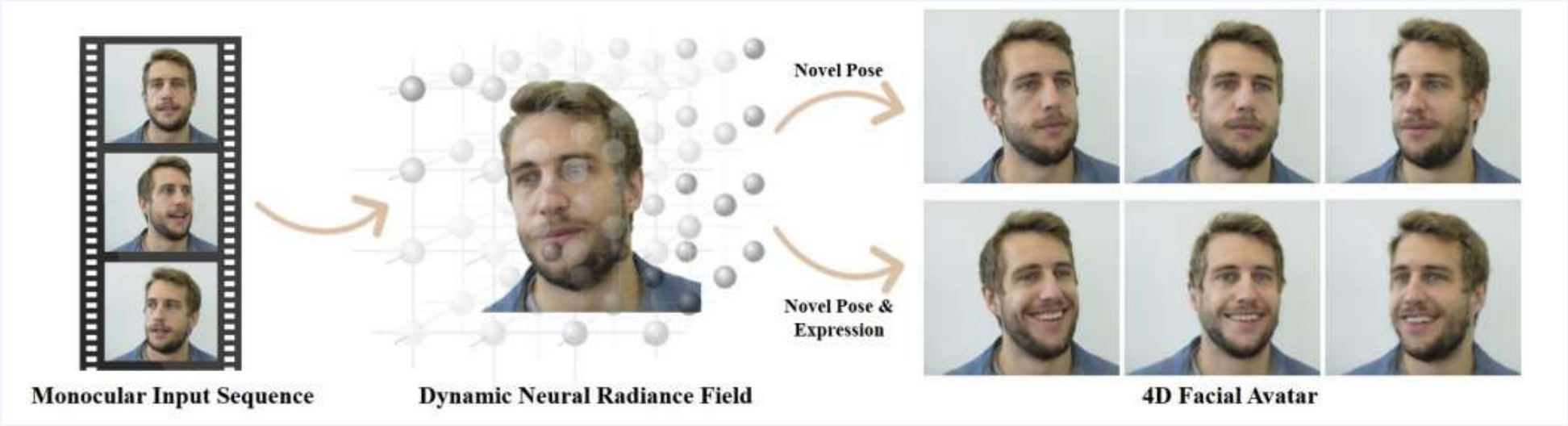
Part3 NerFACE

Presenter: 형준하

Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction

CVPR 2021 Oral

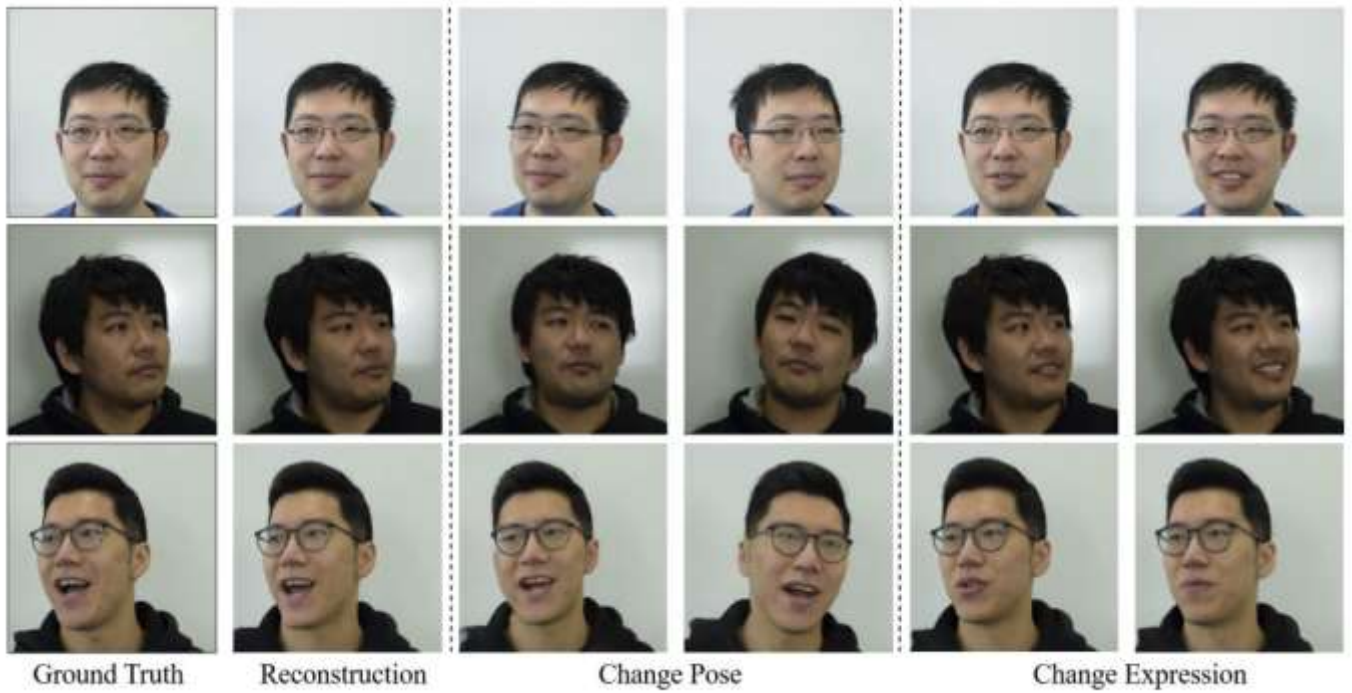
Guy Gafni[1] Justus Thies[1] Michael Zollhöfer[2] Matthias Nießner[1]
[1]Technical University of Munich [2]Facebook Reality Labs



Introduction

Task : Reconstruct 4D models of human head (4D avatars)

- Augmented reality(AR), Virtual reality (VR), visual dubbing ..
- Representing human head with explicit representation is difficult
  - Albedo, reflectance
  - Complex geometry of hair



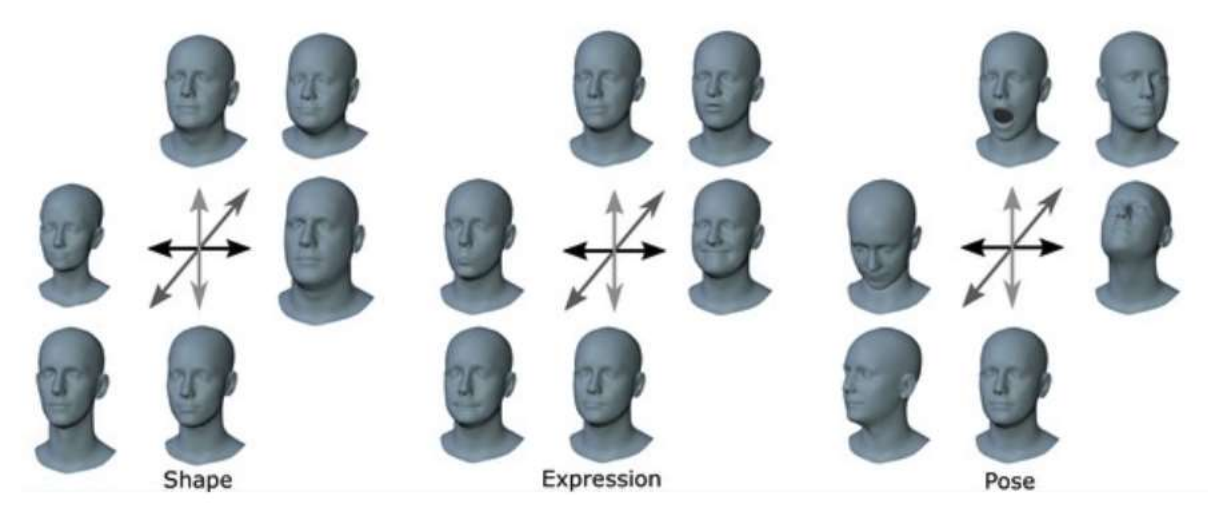
Introduction

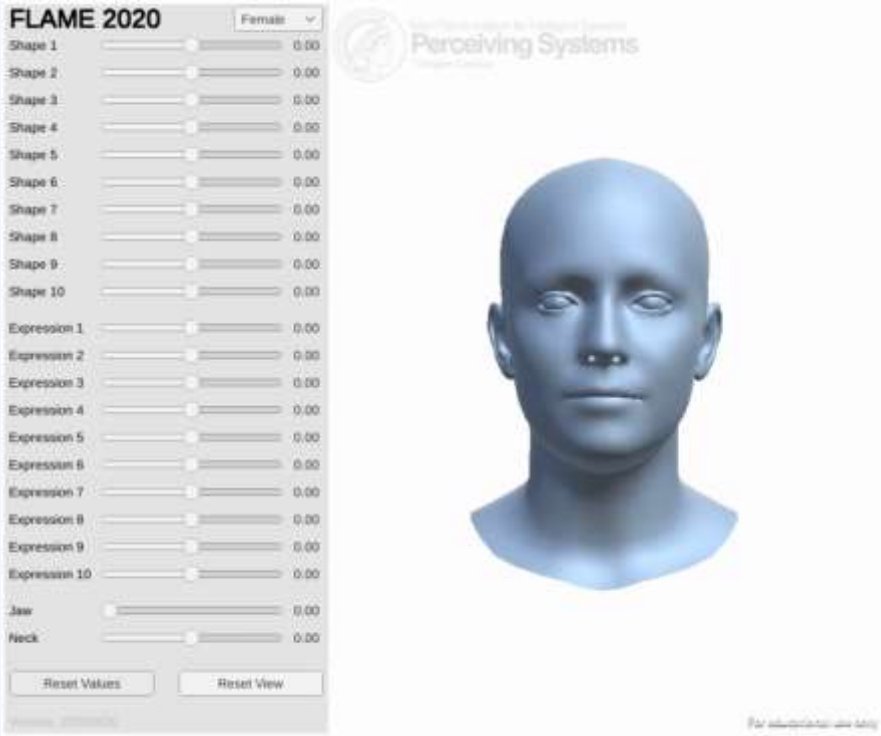
Contribution

- Dynamic Neural Radiance Fields to represent 4D facial avatars based on a low dimensional morphable model.
- An efficient end-to-end learnable approach that uses a single camera to reconstruct such a radiance field.

3D Morphable Model

- Low dimensional representation of 3D human head in 3D mesh
  - Face warehouse (FaceWarehouse: a 3D Facial Expression Database for Visual Computing)
  - Basel Face Model
  - FLAME (Learning a model of facial shape and expression from 4D scans)
- Control identity, expression, pose ..



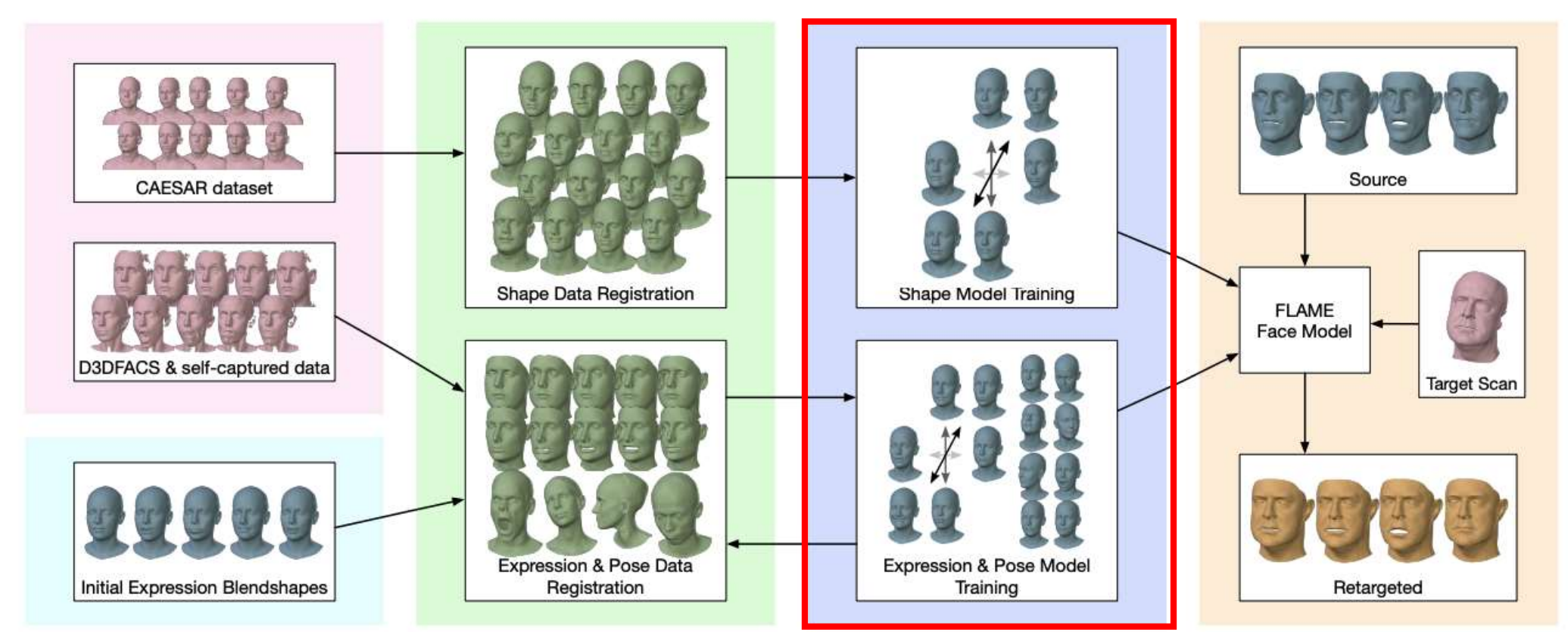


3D Morphable Model

Overall pipeline of building 3DMM

$$B_E(\vec{\psi};\mathcal{E}) = \sum_{n=1}^{|\vec{\psi}|} \vec{\psi}_n \mathbf{E}_n, \quad (5)$$

where $\vec{\psi} = [\psi_1, \cdots, \psi_{|\vec{\psi}|}]^T$ denotes the expression coefficients, and $\mathcal{E} = [\mathbf{E}_1, \cdots, \mathbf{E}_{|\vec{\psi}|}] \in \mathbb{R}^{3N \times |\vec{\psi}|}$ denotes the orthonormal expression

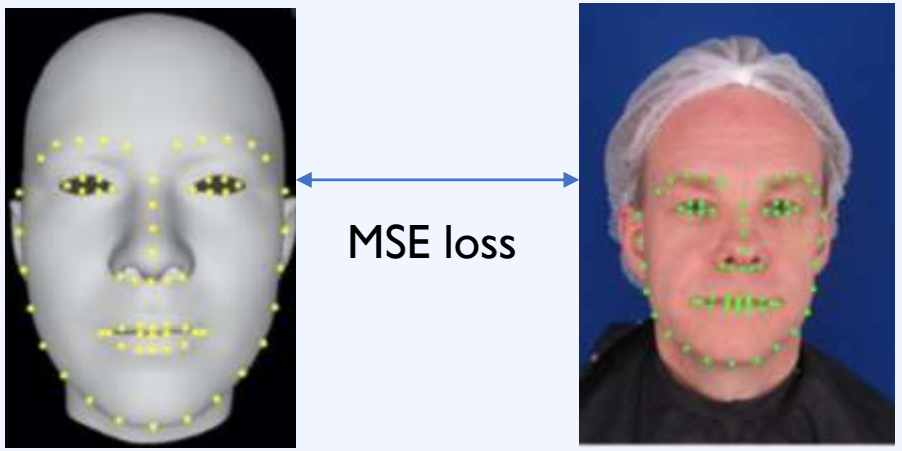


Find 3DMM basis vectors and coefficients

3D Morphable Model

- 2D Image to 3DMM(coefficients) : Face reconstruction, retargeting

- Optimization based reconstruction(e.g. Face2face)



Projected 3D vertices to 2D

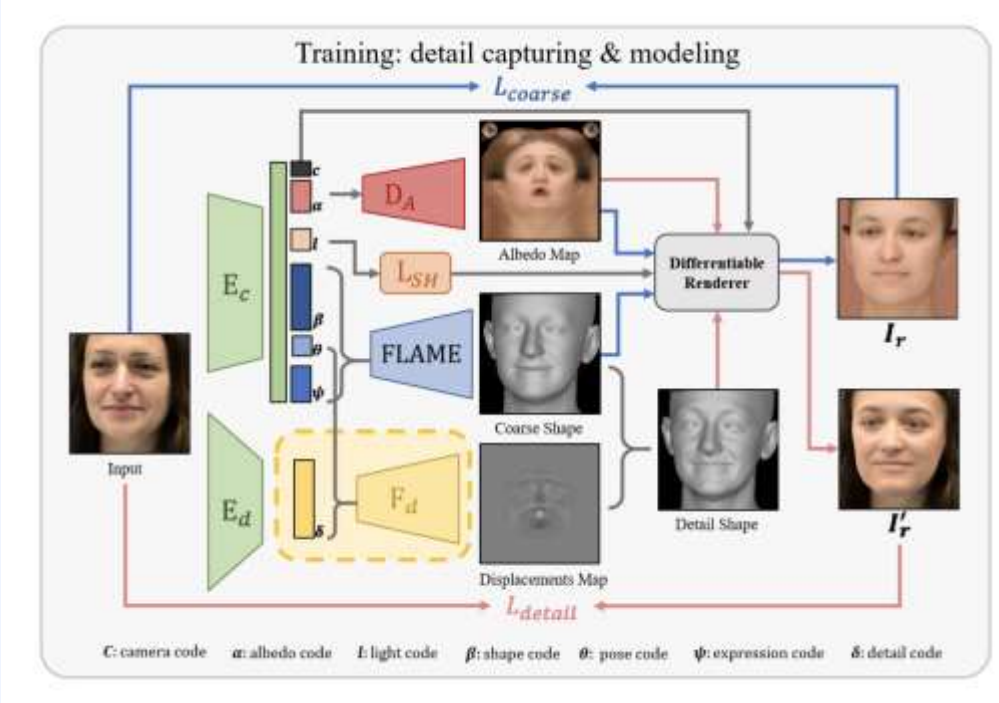
Predicted 2D landmarks with pre-trained landmark detectors

- Accurate, necessary for tracking videos
- Takes long time ☹

3D Morphable Model

- 2D Image to 3DMM(coefficients) : Face reconstruction, retargeting

- Reconstruction w/o optimization(e.g. DECA)

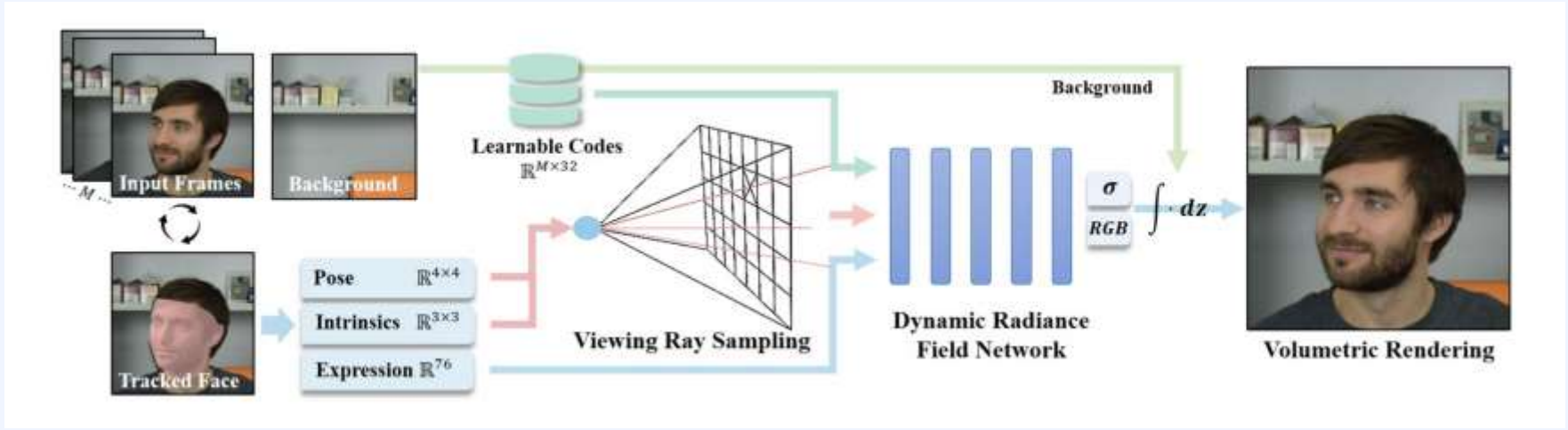


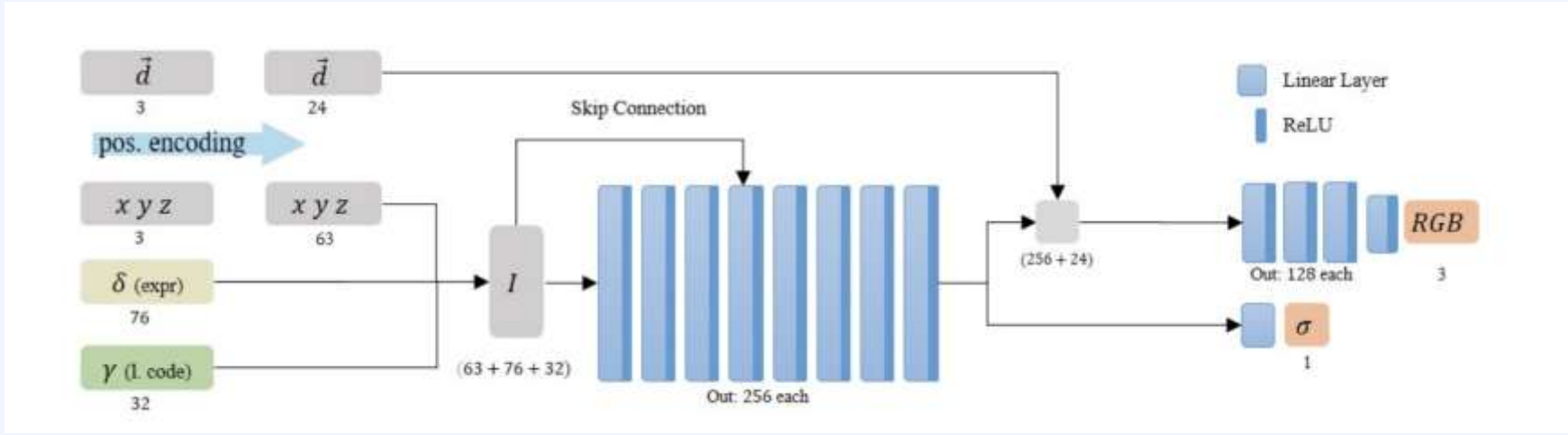
- Fast, able to capture further details
- Can be less accurate, poor tracking quality for videos
- => Initialize with DECA, and further optimize

Learning an Animatable Detailed 3D Face Model from In-The-Wild Images(DECA)

Method

- Optimize with gradient descent on rendering loss

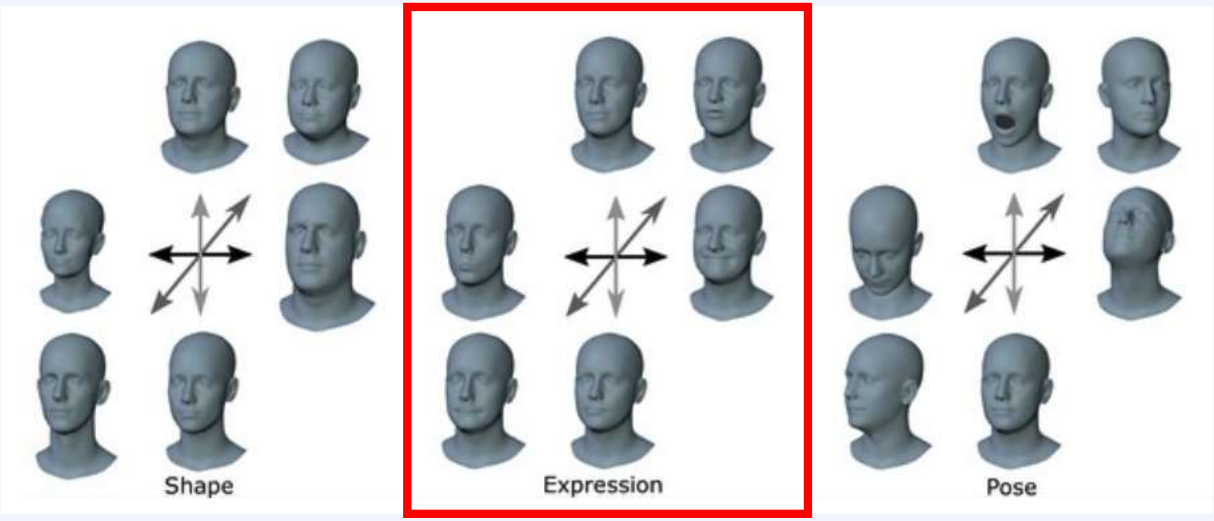




Method – Dynamics Conditioning

- Expression coefficients of delta-blendshape basis

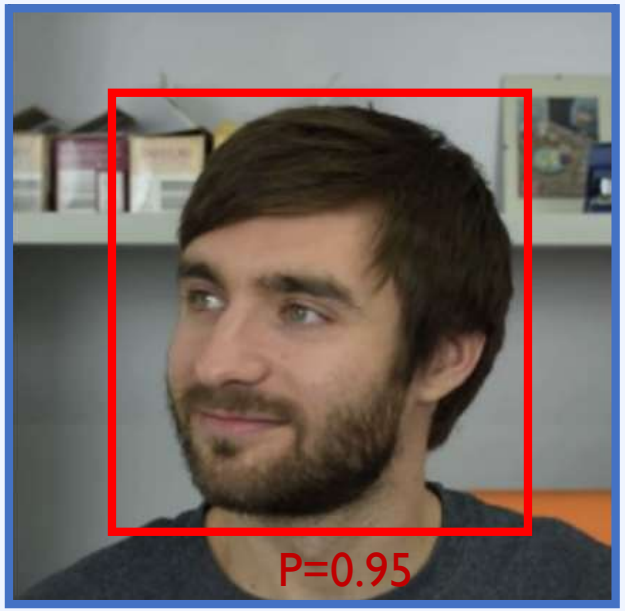
$$\mathcal{D}_\theta(\mathbf{p}, \vec{v}, \delta, \gamma) = (RGB, \sigma)$$



- Pose
  - Define canonical space of head
  - Same method as NeRF
  - Cannot handle torso direction

- Per-frame Latent codes
  - Details that Expression parameters cannot handle
    - Eyeballs
    - Torso
    - Wrinkles ..

Method – Bbox Ray sampling



- Sample rays inside of face bounding box
  - Focus on important areas
  - Similar to center crop of NeRF

Method – Fixed Background

- Want to decouple static background with dynamic foreground
  - Want the background to be fixed with different inputs
- How?
  - Replace the last sample of the ray with background rgb value
  - Easiest way for the model to learn the concept
  - For background rays, densities converge to zero except for the last sample

Background

Ground Truth

Method

- Volume rendering of NeRF

$$\mathcal{C}(\mathbf{r};\theta,P,\delta,\gamma)=\int_{z_{\text{near}}}^{z_{\text{far}}}\sigma_\theta\left(\mathbf{r}\left(t\right)\right)\cdot\mathbf{RGB}_\theta\left(\mathbf{r}\left(t\right),\vec{d}\right)\cdot T(t)dt, \tag{2}$$

- Hierarchical sampling and loss function + l2 regularization of latent codes

$$L_{total}=\sum_{i=1}^{M}L_i(\theta_{coarse})+L_i(\theta_{fine}) \tag{4}$$

with

$$L_i(\theta)=\sum_{j\in\text{pixels}}\left\|\mathcal{C}\left(\mathbf{r}_j;\theta,P_i,\delta_i,\gamma_i\right)-I_i[j]\right\|^2. \tag{5}$$

Method

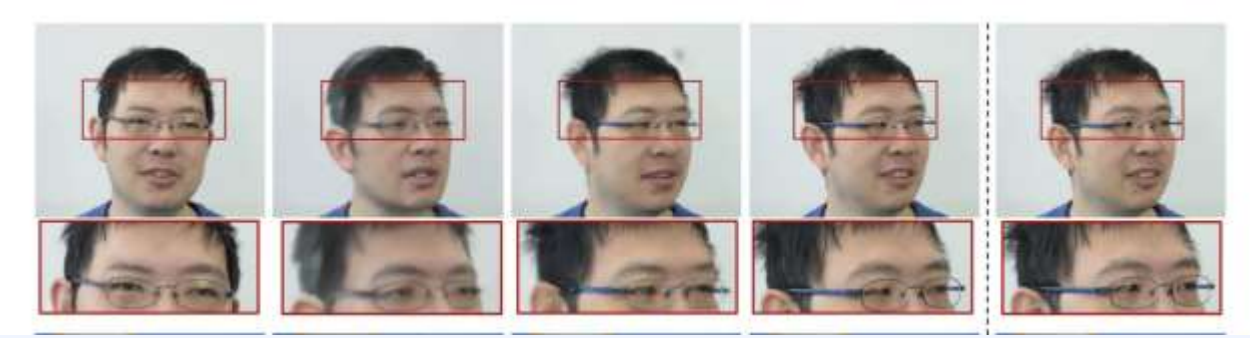
- Training details
  - Batch of 2048 rays
  - 64 points per ray
  - 400K iterations(but generally converge after 600K)

- Dataset
- 2 min video of a single person, static camera (~6000 frames)
  - Conversation, rotation of head
  - 512 x 512 size frames
  - Last 1000 frames for test sequence

Results

- Good at modeling view dependent changes

- Good at modeling head rotations



Results

- Quantitative results

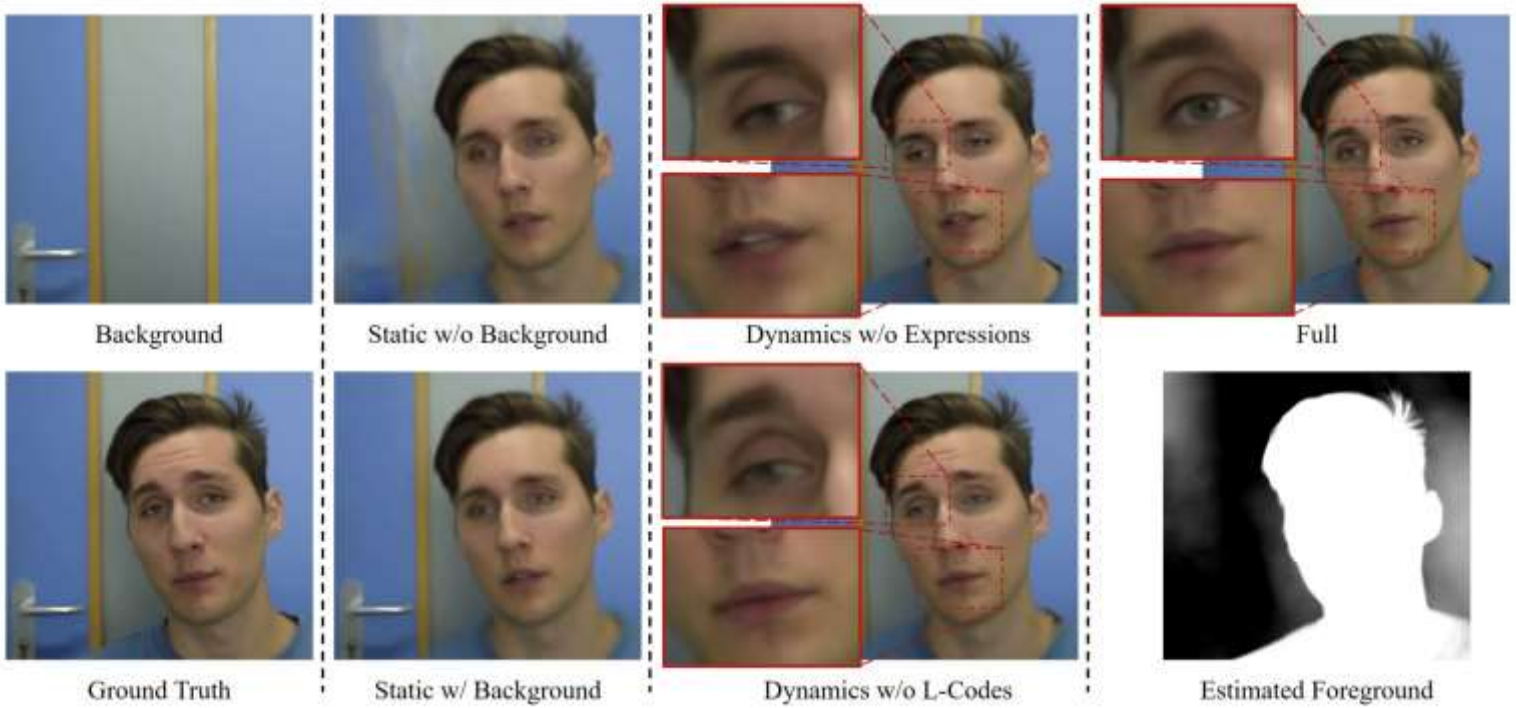
Method	$L_1 \downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓
FOMM [24]	0.036	23.77	0.91	0.16
DVP [16]	0.021	25.67	0.93	0.10
Ours (no BG)	0.035	23.52	0.90	0.18
Ours (no dyn.)	0.024	26.65	0.93	0.11
Ours (full)	**0.019**	**26.85**	**0.95**	**0.06**

Table 1: Quantitative evaluation of our method in comparison to state-of-the-art facial reenactment methods based on self-reenactment (see. Fig. 5). Ours (no dyn.) refers to our method without conditioning on dynamics. Ours (no BG) is our method without background image input.

- But other baseline methods can handle multiple identities..

Results

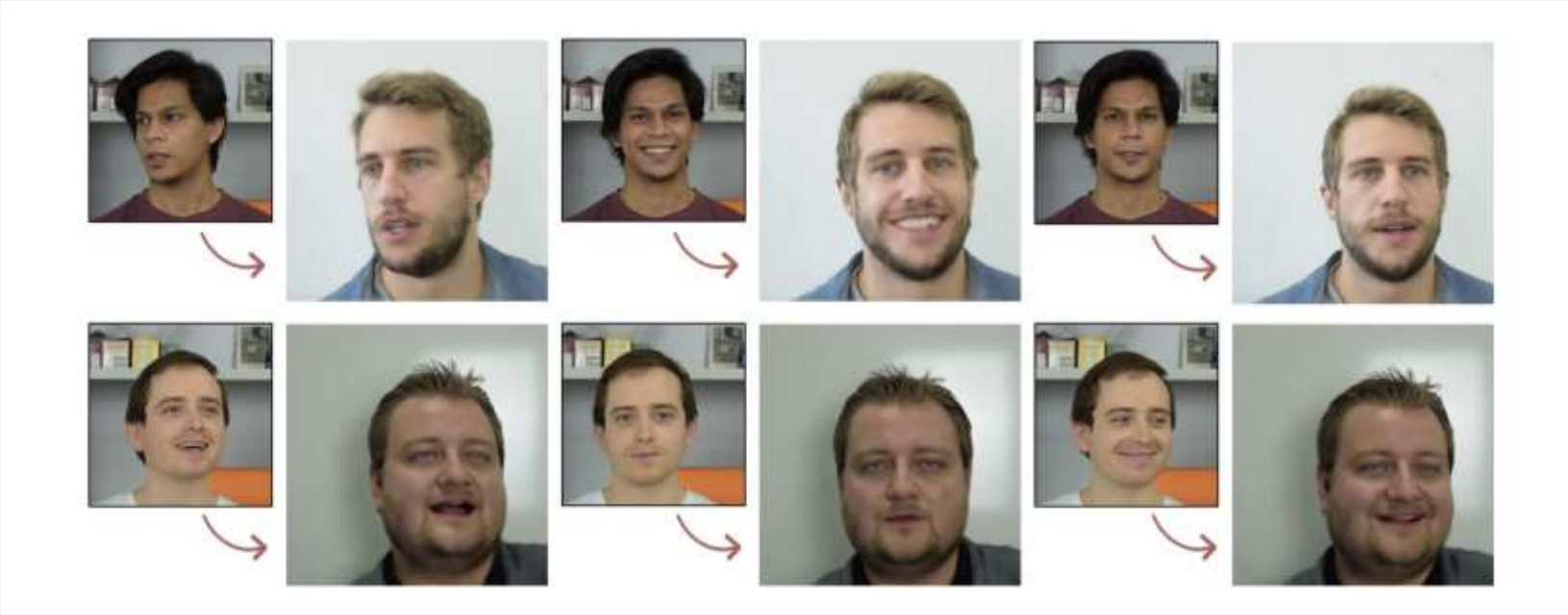
- Ablation study



Method	$L_1 \downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓
Ours (25%)	0.029	24.22	0.93	0.09
Ours (50%)	0.024	25.47	0.94	0.07
Ours (full)	**0.019**	**26.85**	**0.95**	**0.06**

Results

- Cross-reenactment



Results

- Limitations
  - Cannot control details such as eye blinks
  - Cannot control torso
  - Single identity fitting
    - 2D based works generally work on one-shot setting

- Further Works
  - IMAVATAR(I M Avatar: Implicit Morphable Head Avatars from Videos)
    - Use 3DMM directly with deformation field
  - RigNeRF (RigNeRF: Fully Controllable Neural 3D Portraits)
    - Control camera pose & head pose