FNeVR: Neural Volume Rendering for Face Animation

Bohan Zeng[1,†], Boyu Liu[1,†], Hong Li[1], Xuhui Liu[1,*], Jianzhuang Liu[2], Dapeng Chen[3], Wei Peng[3], Baochang Zhang[1,4,*]

[1]Beihang University, [2]Huawei Noah's Ark Lab, [3]Huawei, [4]Zhongguancun Laboratory

*Corresponding author, email: xhliu@buaa.edu.cn, bczhang@buaa.edu.cn

†Equal contributions

Motivation

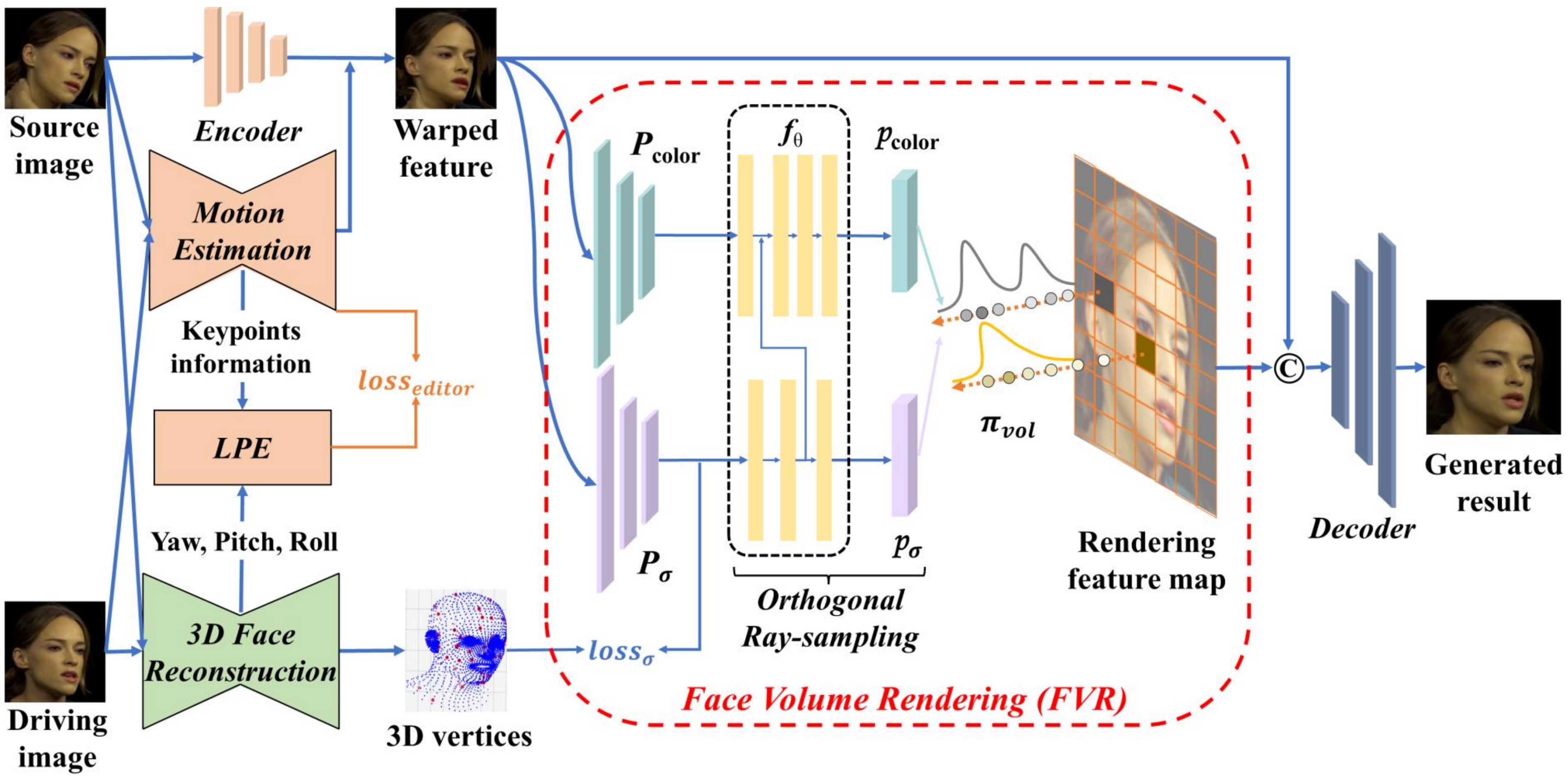
Face animation methods can be mainly divided into three categories.

- **Model-free methods:** limited ability to produce sufficiently realistic images (e.g. FOMM [1])
- **Landmark-based methods:** disadvantages in identity preservation
- **3D structure-based methods:** deficiency in motion transfer (e.g. Face vid2vid [2])

FNeVR takes the merits of **2D motion warping** on facial expression transformation and **3D volume rendering** on high-quality image synthesis in a unified framework.

Source **Driving** **FOMM (2019)** **Face vid2vid (2021)** **FNeVR (ours)**

Framework and Proposed Method



Framework and Proposed Method

1. 2D Motion Estimation

Motion estimation by 2D keypoints $\{p_{S,k}, p_{D,k} \in \mathbb{R}^2\}$ and Jacobians $\{J_{S,k}, J_{D,k} \in \mathbb{R}^{2\times 2}\}$ [1]:

$$\mathcal{T}_{S\leftarrow D,k}(z) \approx p_{S,k} + J_{S,k} J_{D,k}^{-1}(z - p_{D,k})$$

By 2D warping, we obtain the warped feature F_w.

2. 3D Face Reconstruction

Reconstruct a 3D face mesh v from 2D image by encoding parameters $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}$ [3]:

$$v = W\left(\boldsymbol{T_p}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}), \boldsymbol{J}(\boldsymbol{\beta}), \boldsymbol{\theta}, \boldsymbol{\mathcal{W}}\right)$$

By using Gaussian function to process v, we obtain the 3D mesh feature of the reconstruction result F_m.

Framework and Proposed Method

3. Face Volume Rendering (FVR)

3.1. 3D Feature Extraction

$$F_\sigma = P_\sigma(F_w) \in \mathbb{R}^{H\times W\times D\times N_\sigma}$$

$$F_{color} = P_{color}(F_w) \in \mathbb{R}^{H\times W\times D\times N_{color}}$$

$$\sigma \text{ matching loss: } \mathcal{L}_\sigma = \exp(-\alpha_1 < F_\sigma \cdot F_m >) - \alpha_2$$

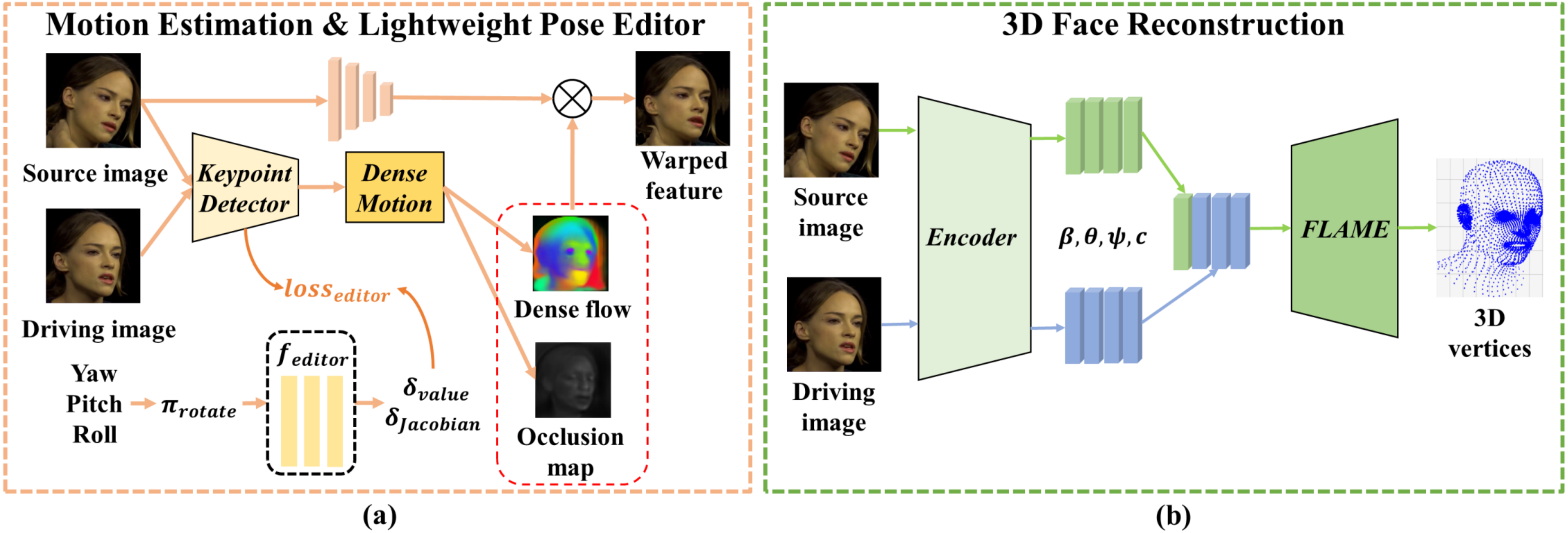
3.2. Orthogonal Adaptive Ray-Sampling

$$p_\sigma, p_{color} = f_\theta(F_\sigma, F_{color}) \in R^{H\times W\times D\times 1} \times R^{H\times W\times D\times M_{color}}$$

3.3. Image Rendering

$$F_{r,i} = \sum_{j=1}^{D} \tau_j \left(1 - \exp(-p_{\sigma,i,j})\right) p_{color,i,j}$$

Illustration of Motion Estimation and LPE modules, and 3D Face Reconstruction module



Framework and Proposed Method

4. Lightweight Pose Editing (LPE)

Estimation of new keypoints δ_{value} and Jacobians $\delta_{Jacobian}$ under a specific rotation angle θ_{rotate}:

$$\delta_{value}, \delta_{Jacobian} = f_{editor}(\theta_{rotate}, p_S, J_S) \in \mathbb{R}^{K\times 2} \times \mathbb{R}^{K\times 2\times 2}$$

Pose editor loss:

$$\mathcal{L}_{editor} = \lambda_1 L_1(p_D, \delta_{value}) + \lambda_2 L_1(J_D, \delta_{Jacobian})$$

Experiments and Results

Same-Identity Reconstruction: **state-of-the-art** on VoxCeleb [4]

Method	$\mathcal{L}_1 \downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	AKD $\downarrow$	AED $\downarrow$	FID $\downarrow$
Bilayer	0.1197	0.4247	15.219	0.3968	12.60	0.0546	219.8
FOMM	0.0450	0.1099	23.210	0.7475	1.383	0.0244	11.56
Face vid2vid	0.0485	0.1051	22.642	0.7268	1.616	0.0395	9.142
Face vid2vid-S	0.0445	0.0901	23.357	0.7473	1.421	0.0243	9.151
DaGAN	0.0462	0.0981	23.263	0.7536	1.441	0.0247	9.660
PIRender	0.0566	0.0850	21.040	0.6550	2.186	0.2245	11.88
FNeVR (ours)	**0.0404**	**0.0804**	**24.292**	**0.7773**	**1.254**	**0.0231**	**8.443**

Experiments and Results

Cross-Identity Reenactment: **best overall performance** with **less computation and memory cost (FLOPs and Parameters)** on VoxCeleb [4] and VoxCeleb2 [5]

Method	VoxCeleb		VoxCeleb2	
	FID↓	CSIM ↑	FID↓	CSIM ↑
FOMM	106.9	0.5491	138.1	0.5228
Face vid2vid-S	106.6	**0.6447**	148.6	**0.6290**
DaGAN	110.3	0.5305	139.6	0.4932
FNeVR (ours)	**98.23**	0.5505	**133.9**	0.5282

Method	Flops(G)	Params(M)	FPS
Face vid2vid	231.038	125.216	17.790
Face vid2vid-S	636.941	173.109	13.219
DaGAN	**75.642**	74.660	26.753
FNeVR (ours)	130.109	**61.378**	**36.568**

Visualization – Reconstruction

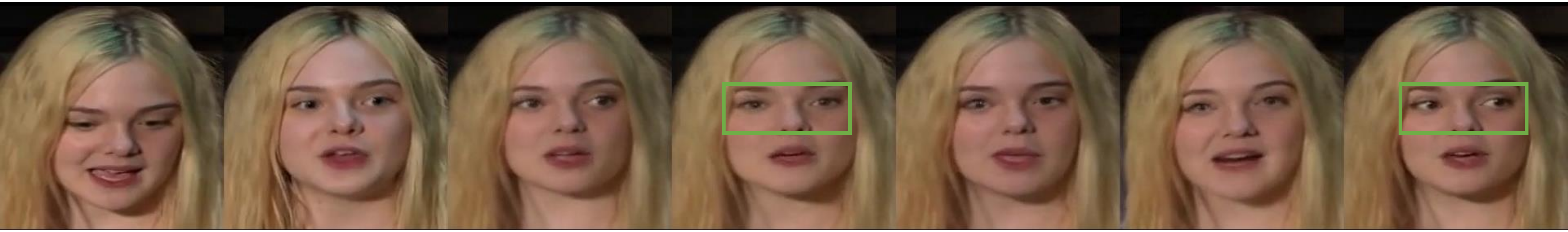
Source | Driving | FOMM (2019) | Face vid2vid-s (2021) | DaGAN (2022) | PIRenderer (2021) | FNeVR

Our generated video is much clearer than FOMM's, with more natural and realistic facial details than FOMM, Face vid2vid-s, DaGAN and PIRenderer(especially eyes and mouth).

Visualization – Reconstruction

Source **Driving** **FOMM (2019)** **Face vid2vid-s (2021)** **DaGAN (2022)** **PIRenderer (2021)** **FNeVR**

Visualization – Reconstruction



Visualization – Reenactment

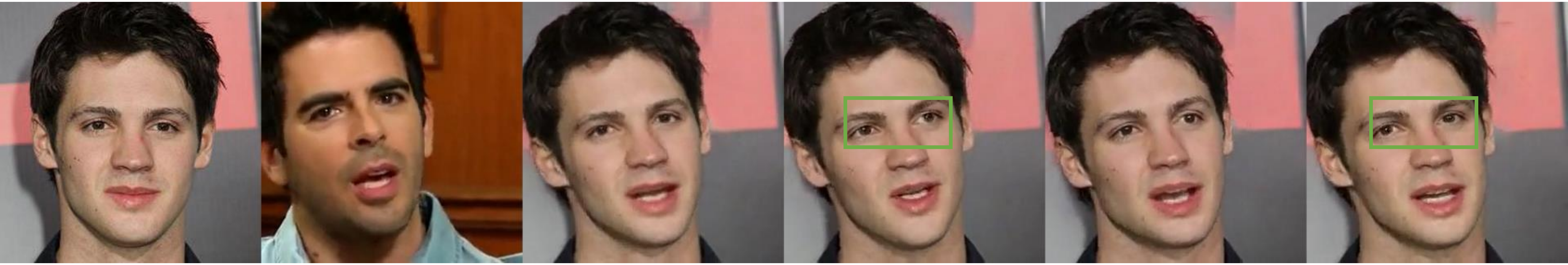
Source | **Driving** | **FOMM (2019)** | **Face vid2vid-s (2021)** | **DaGAN (2022)** | **FNeVR**

Our generated video is much clearer than FOMM's, with more natural and realistic facial details than FOMM, Face vid2vid-s and DaGAN (especially eyes and mouth).

Visualization – Reenactment

Source | Driving | FOMM (2019) | Face vid2vid-s (2021) | DaGAN (2022) | FNeVR

Visualization – Reenactment



Source | Driving | FOMM (2019) | Face vid2vid-s (2021) | DaGAN (2022) | FNeVR

Conclusion

- We propose a Face Neural Volume Rendering (FNeVR) network for face animation, which unifies the 2D motion warping and 3D volume rendering in one framework.
- We innovatively develop a Face Volume Rendering (FVR) module to enhance the facial details of the warped feature and generate high-quality faces. Moreover, we design a Lightweight Pose Editing (LPE) module, which can directly implement pose editing with rotation angles.
- Extensive experiments illustrate that our FNeVR achieves state-of-the-art performance.

References

- [1] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In NeurIPS, 2019.
- [2] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In CVPR, 2021.
- [3] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. TOG, 2017.
- [4] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. INTERSPEECH, 2017.
- [5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. INTERSPEECH, 2018.

Thank you for listening