ROME: REALISTIC ONE-SHOT MESH-BASED HEAD AVATARS

Presenter: Jaeseong Lee

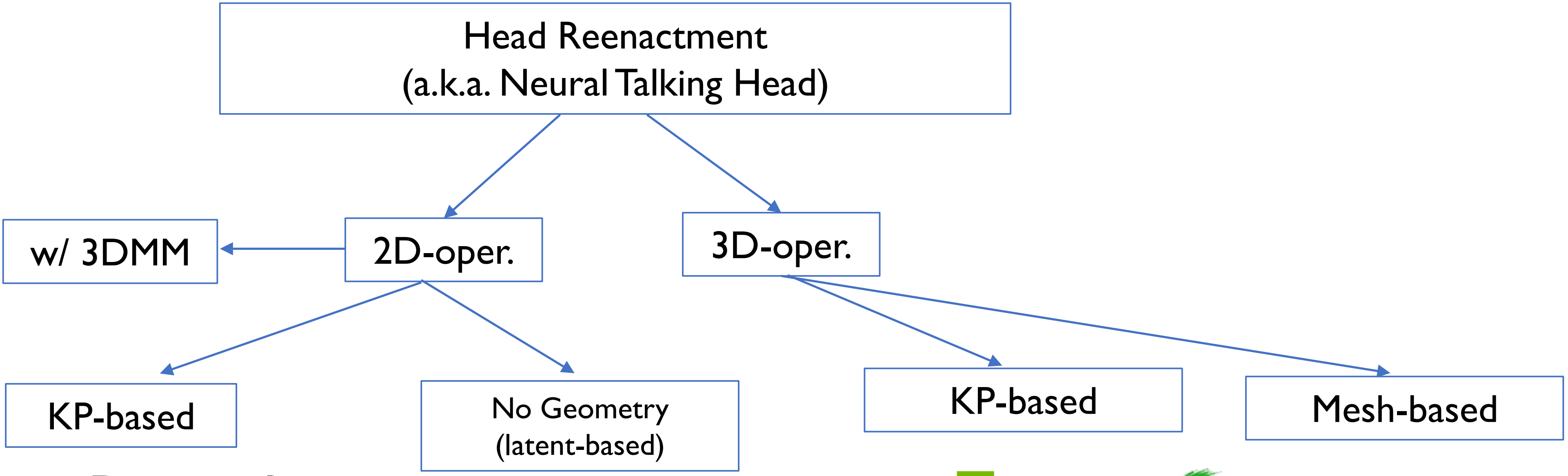
ECCV Under-Review
(Released to Arxiv 17/June)
Samsung AI Moscow

DAVIAN
Data and Visual Analytics Lab

Contents

- Preliminaries
- Overview
- Results
- Limitations

Preliminaries – Taxonomy



Head Reenactment
(a.k.a. Neural Talking Head)

2D-oper.

3D-oper.

w/ 3DMM

KP-based

No Geometry
(latent-based)

KP-based

Mesh-based

Tencent AI Lab

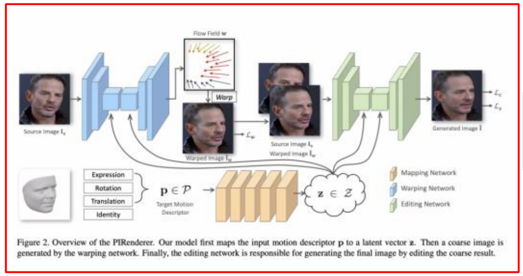
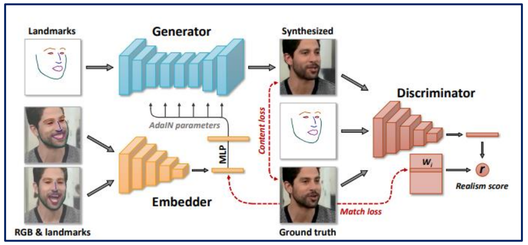


Figure 2. Overview of the PIRenderer. Our model first maps the input motion descriptor p to a latent vector z. Then a coarse image is generated by the warping network. Finally, the editing network is responsible for generating the final image by editing the coarse result.

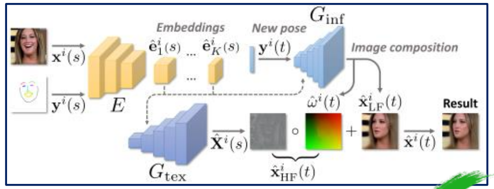
PIRenderer(ICCV 2021)

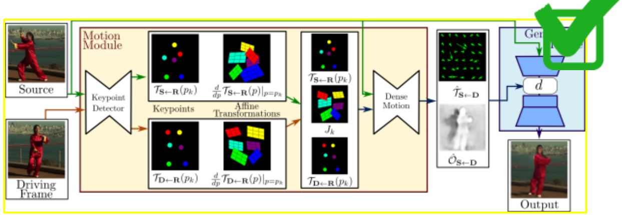
Few-shot TH(ICCV 2019)

Samsung Research



Fast-bi(ECCV 2020)

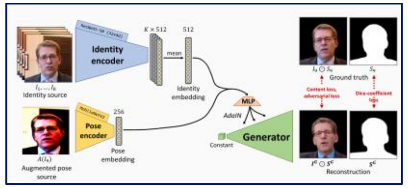




Snap Inc.

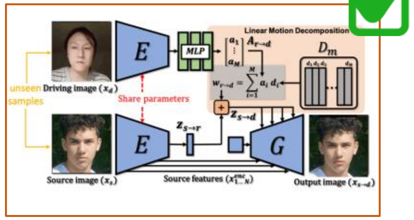
FOMM(NIPS 2019) ✓

Samsung Research



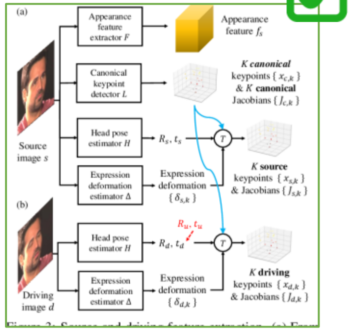
LPD(CVPR 2020)

Inria



LIA(ICLR 2022) ✓

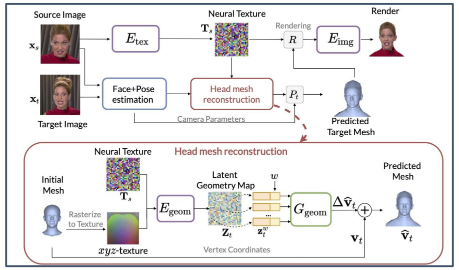
NVIDIA



(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K canonical keypoints & K canonical Jacobians
Head pose estimator H
Expression deformation estimator Δ
K source keypoints & Jacobians
Source image s
(b)
Driving image d
K driving keypoints & Jacobians

OSFV(CVPR 2021) ✓

Samsung Research



ROME(ECCV 2022(?))

OVERVIEW(CONT'D) - STRENGTHS

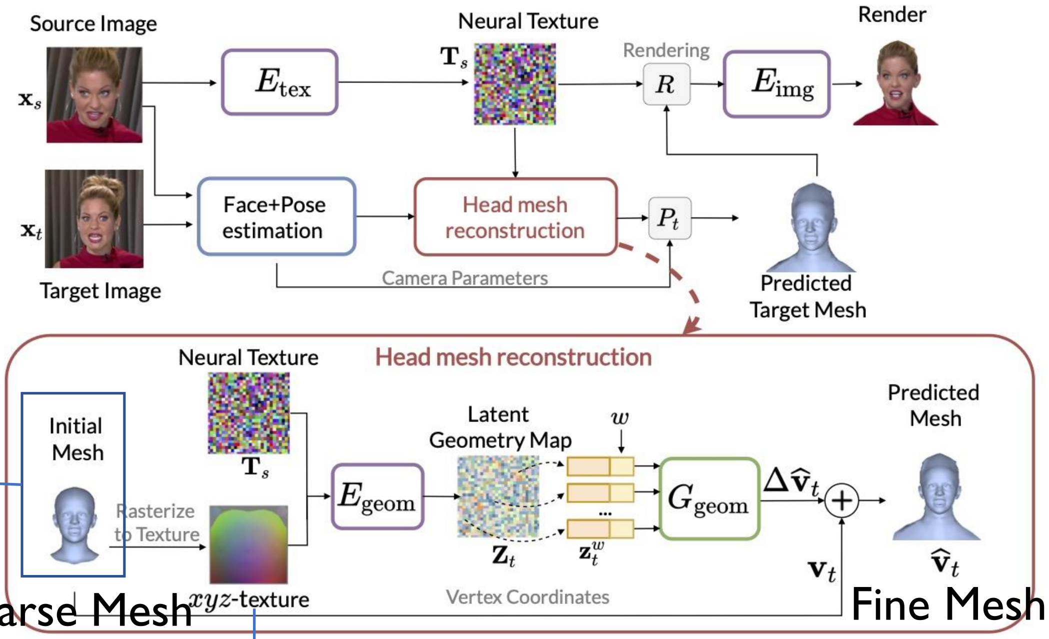
- Strengths from two Perspectives
  - Head recon. : Viable to handle non-facial parts(e.g., hear and torso)
  - Talking head : Viable to handle unseen facial parts or large drv/src-discrepancy (e.g., one-side facing smthng)

* From DECA(SIGGRAPH 2021)

** From DAGAN(CVPR 2022)

Overview – Pipeline(cont'd)

1. Rule-based data process



Coarse Mesh

Fine Mesh

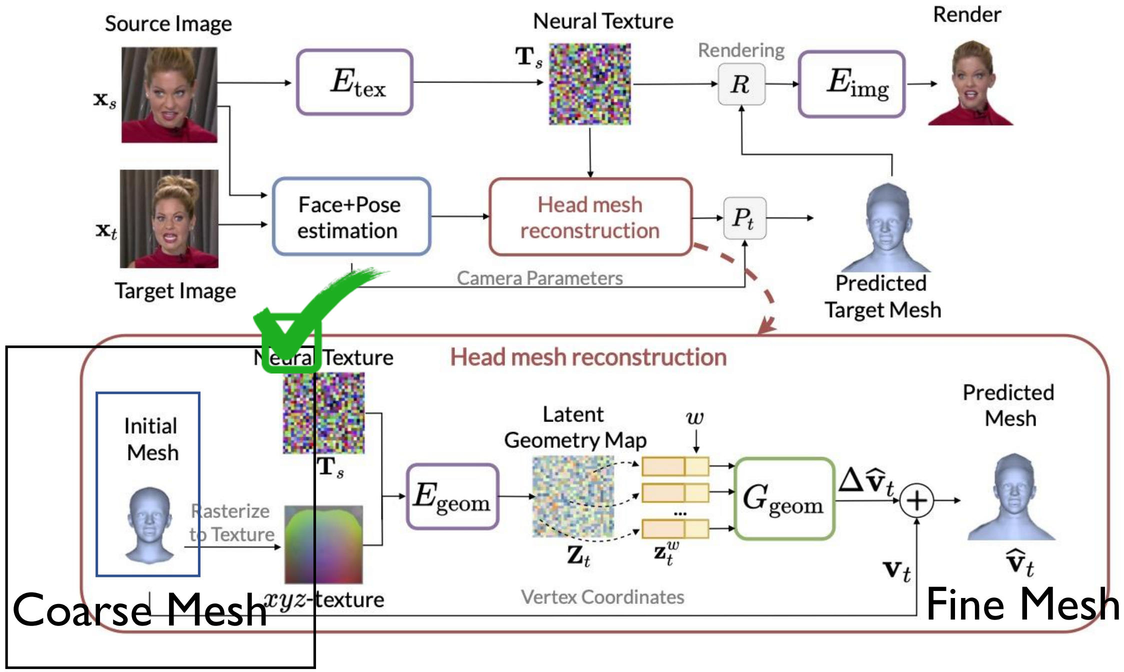
From off-the-shelf 3DMM extractor(DECA)

* 3DMM: representing human head from 3D scanned mesh PCA-based parameterized(Shape and Expression)

Implies which 2d coordinate corresponds to vertex

Overview – Pipeline(cont’d)

2. How 3DMM works? How to reconstruct the Initial Mesh?

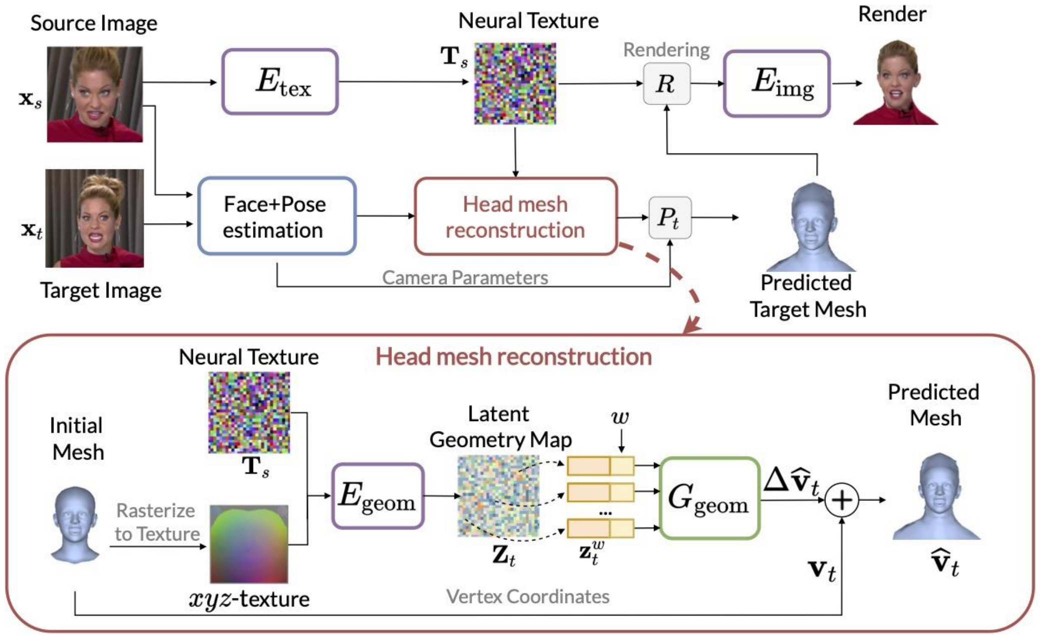


*Src: Shape
*Tgt: Expression and Cam/Head pose

$$\boldsymbol{v}(\phi, \psi, \theta) = \mathrm{W}(\boldsymbol{v}_{base} + \beta\phi + D\psi, \theta)$$

Where,
ϕ: Shape parameter
ψ: Expression parameter
θ: Head pose parameter

Overview – Loss functions



Novel losses

$$\mathcal{L}_{\text{occ}} = \lambda_{\text{hair}} \left\| \hat{\mathbf{o}}_t^{\text{hair}} - \mathbf{s}_t^{\text{hair}} \right\|_2^2 + \lambda_{\text{o}} \left\| \hat{\mathbf{o}}_t - \mathbf{s}_t \right\|_2^2.$$

$$\mathcal{L}_{\text{chm}} = \frac{1}{2N_t} \sum_{\hat{p}_t \in \hat{\mathbf{P}}_t} \left\| \hat{p}_t - \arg\min_{p \in \mathbf{P}_t} \| p - \hat{p}_t \| \right\| + \frac{1}{2N_t} \sum_{p_t \in \mathbf{P}_t} \left\| p_t - \arg\min_{\hat{p} \in \hat{\mathbf{P}}_t} \| \hat{p} - p_t \| \right\|.$$

$$\mathcal{L}_{\text{lap}} = \frac{1}{V} \sum_{i=1}^{V} \left\| \Delta\hat{\mathbf{v}}_i - \frac{1}{\mathcal{N}(i)} \sum_{j \in \mathcal{N}(i)} \Delta\hat{\mathbf{v}}_j \right\|_1,$$

Where,
p_t : sampled set of 2D points in the predicted segmentation mask s_t
$\hat{p_t}$: projected 2D vertices coordinate at the target image

align

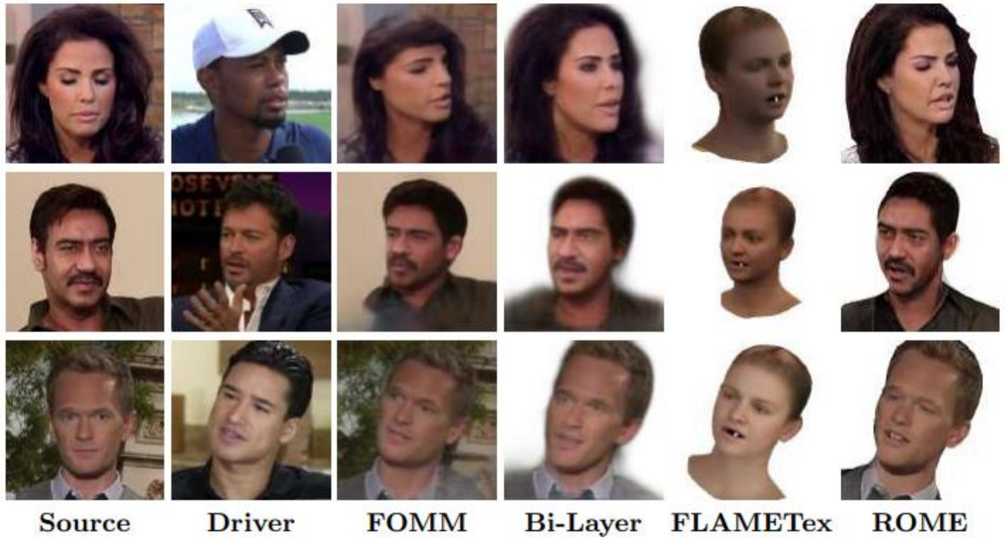
regularization

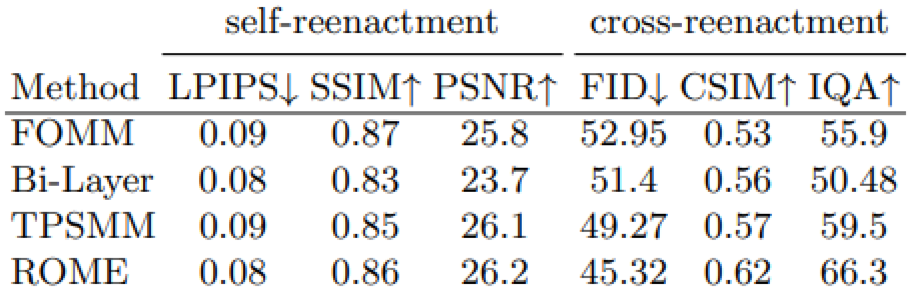
Widely-used losses

$$L_{adv} + L_{perc} + L_{arcface} + L_{dice}$$

Results(cont'd)

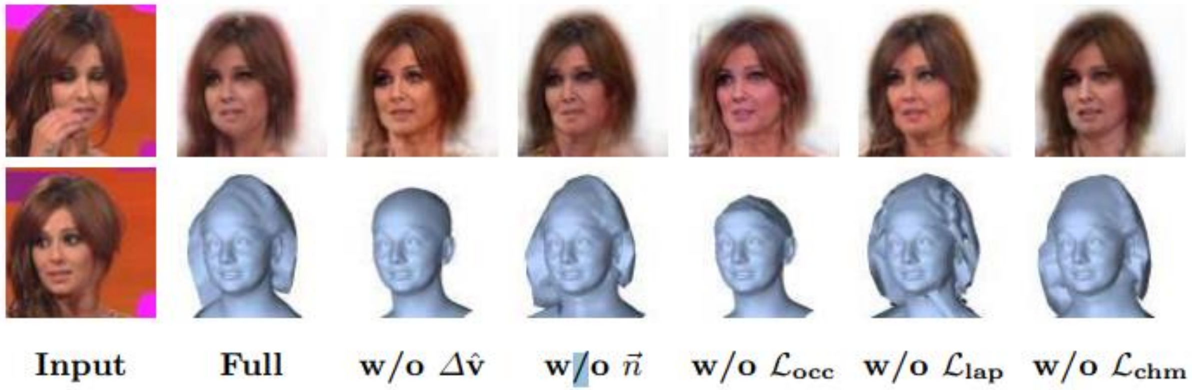
Comparison with Head Reenactment models



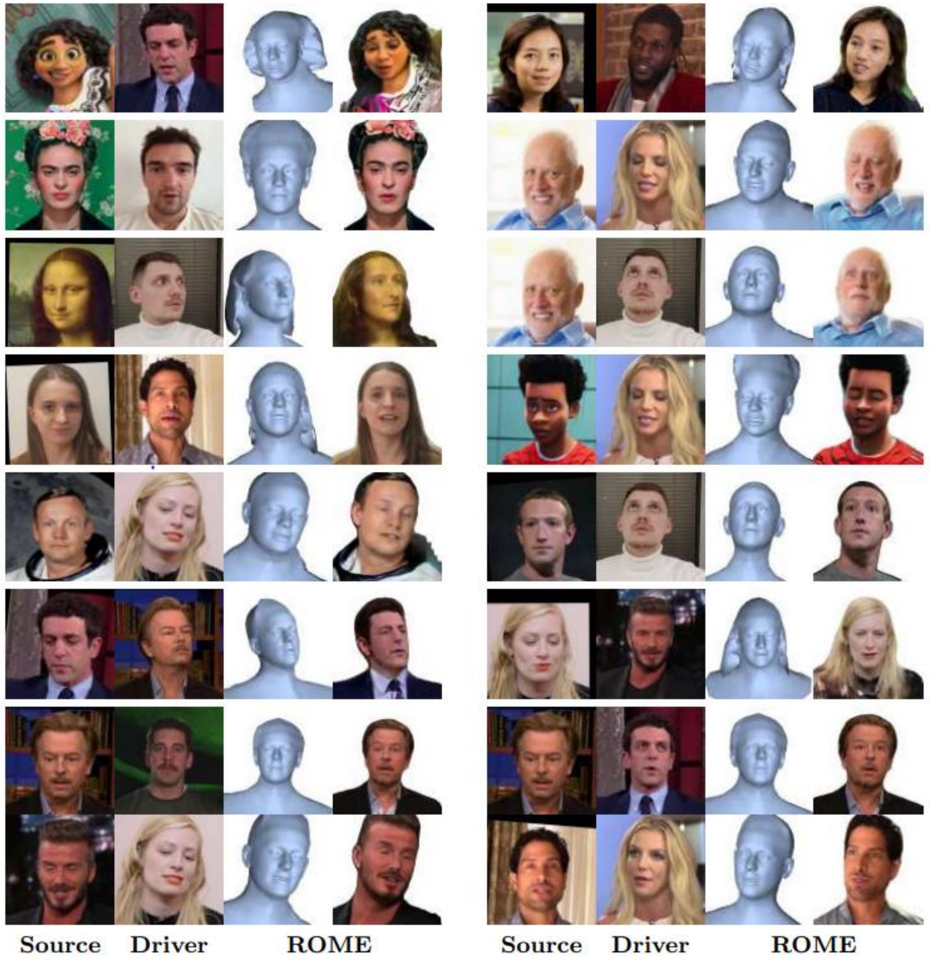


	self-reenactment			cross-reenactment		
Method	LPIPS↓	SSIM↑	PSNR↑	FID↓	CSIM↑	IQA↑
FOMM	0.09	0.87	25.8	52.95	0.53	55.9
Bi-Layer	0.08	0.83	23.7	51.4	0.56	50.48
TPSMM	0.09	0.85	26.1	49.27	0.57	59.5
ROME	0.08	0.86	26.2	45.32	0.62	66.3

Ablation Studies

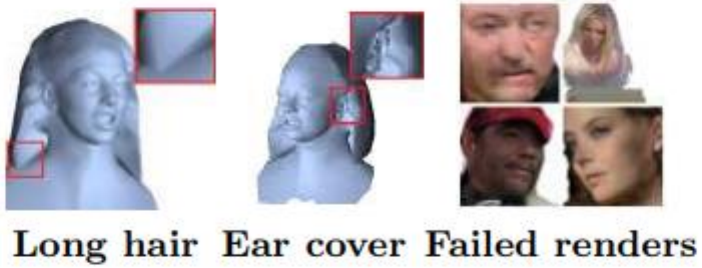


Result



Limitations

- Mesh resolution



Long hair Ear cover Failed renders

(b) Limitations