CVTHead: One-shot Controllable Head Avatar with Vertex-feature Transformer

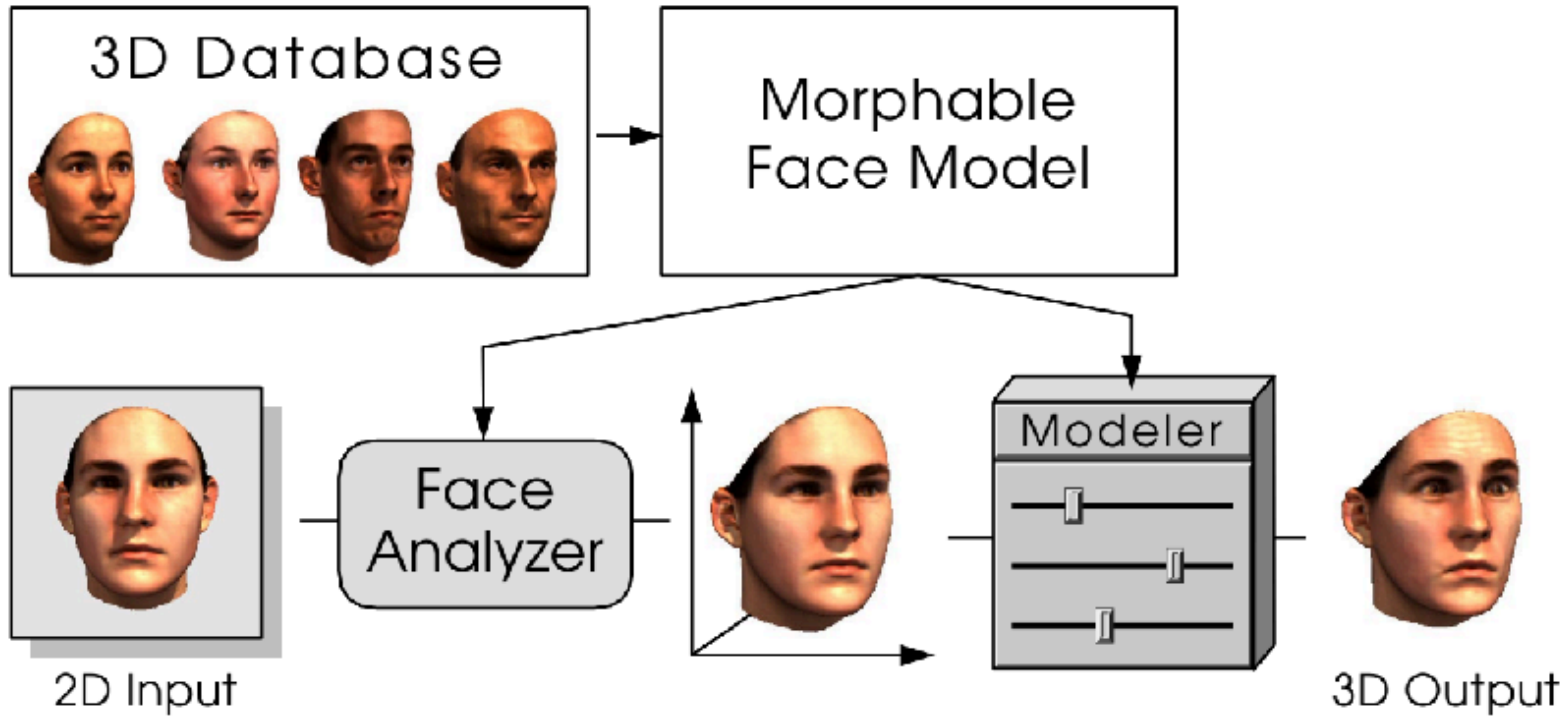
WACV 2024

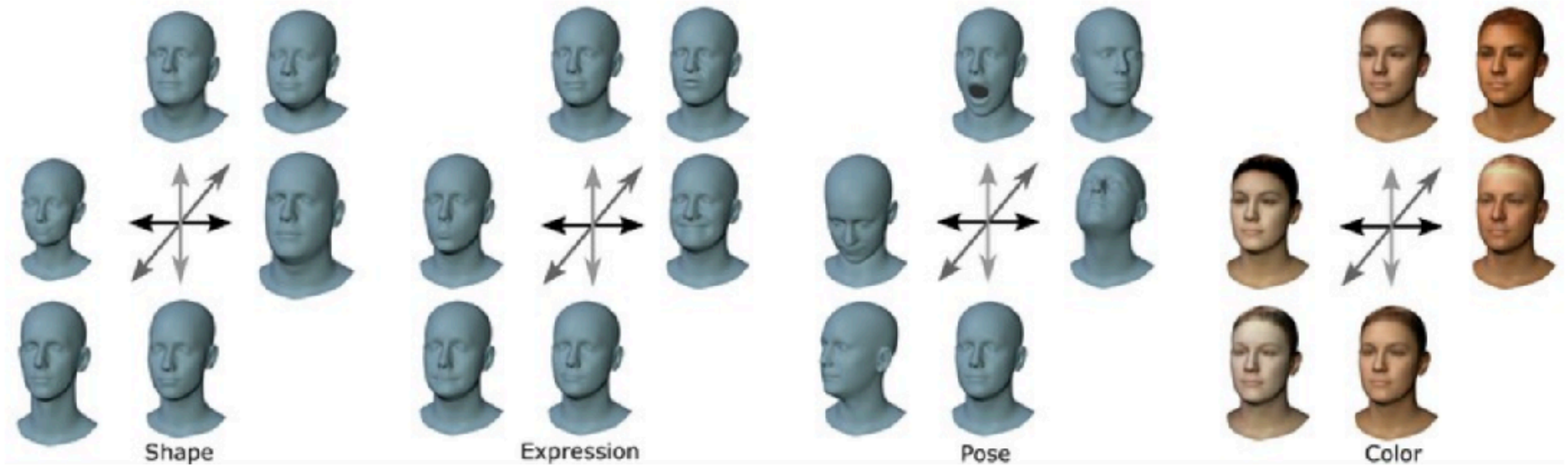
Haoyu Ma, Tong Zhang, Shanlin Sun, Xiangyi Yan, Kun Han, Xiaohui Xie

University of California, Irvine

Background: 3D Morphable Face Models (3DMM)

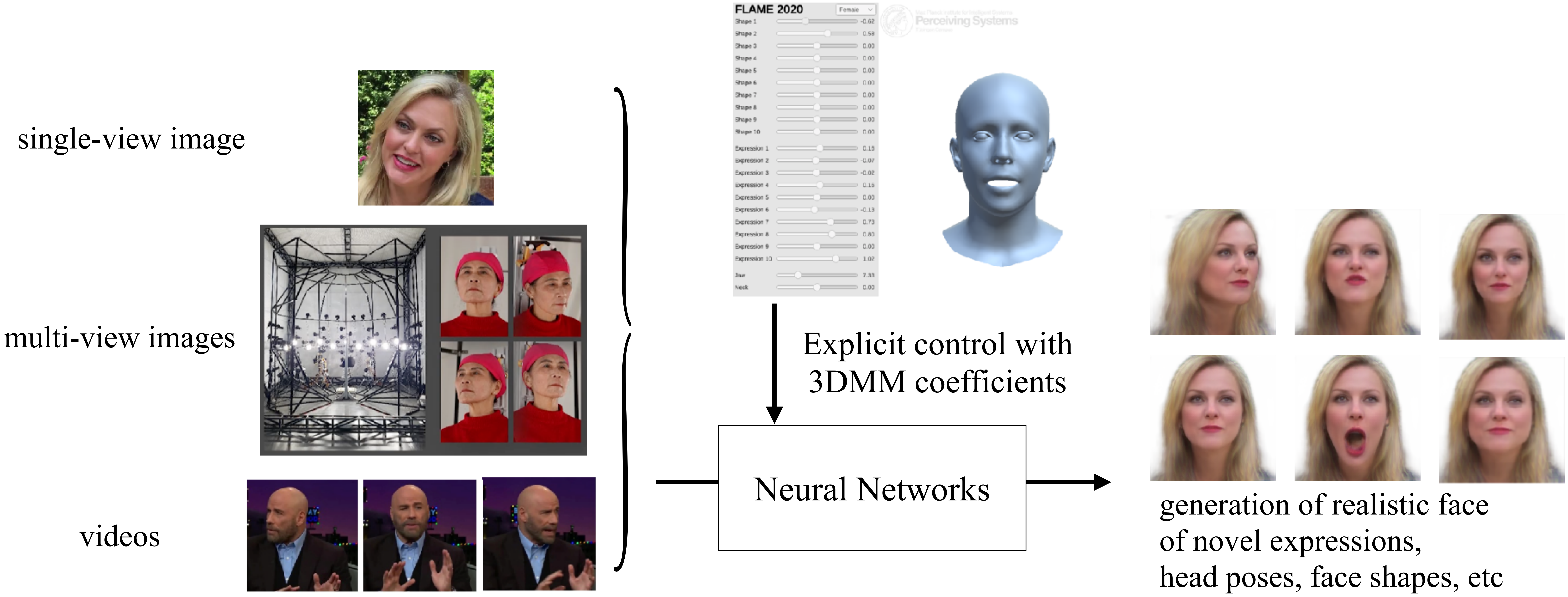
- Parametric model:
  - explicit control of shape, expression, head pose, texture, etc by coefficients
  - no information on detailed regions such as hair





[1] Volker Blanz, et al. "A Morphable Model For The Synthesis Of 3D Faces." *TOG, 1999*
[2] Li, Tianye, et al. "Learning a model of facial shape and expression from 4D scans." *TOG, 2017*

Background: 3DMM-based face generation

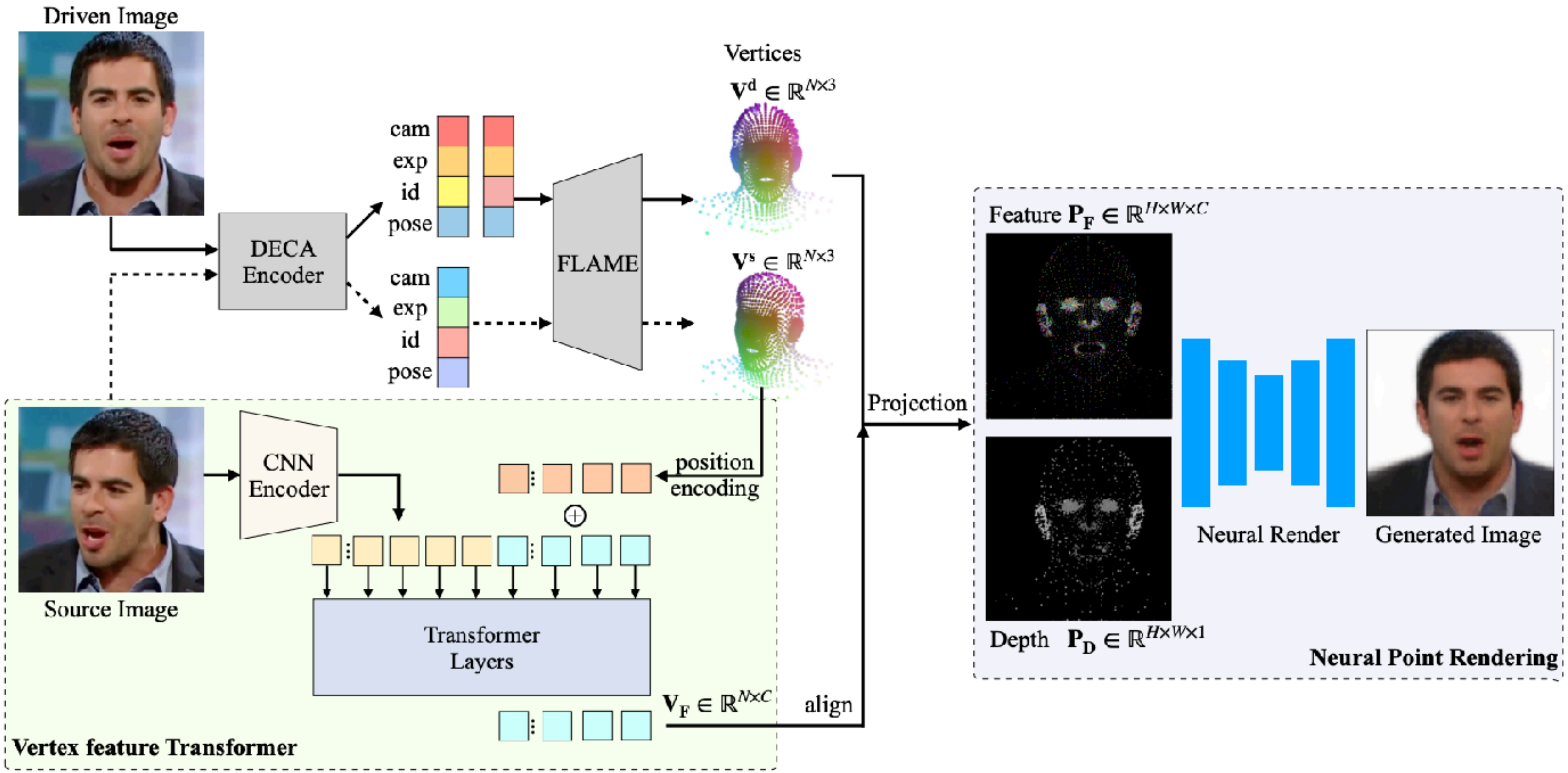


[1] Li, Tianye, et al. "Learning a model of facial shape and expression from 4D scans." *TOG, 2017*

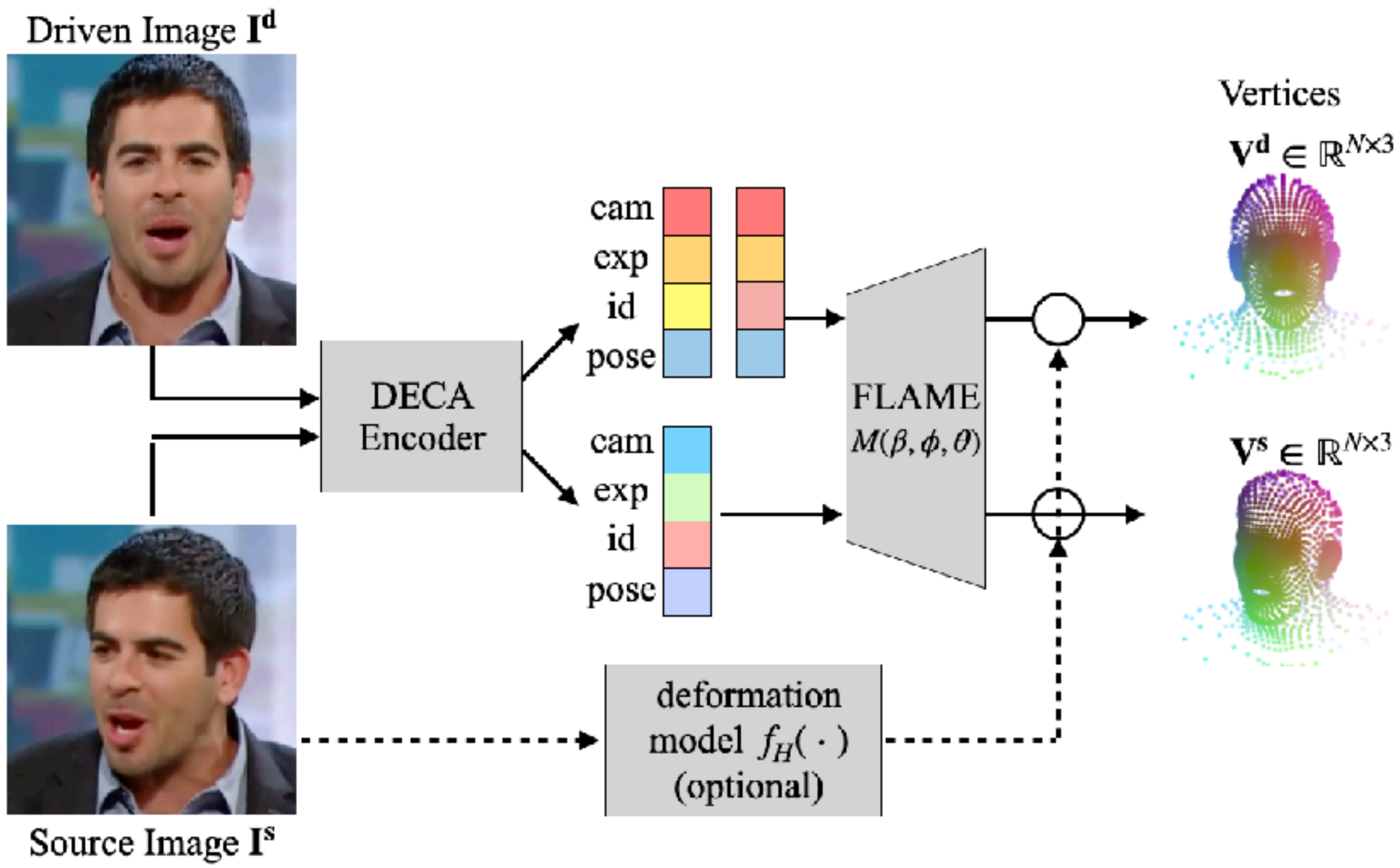
CVTHead: Framework

Efficient and controllable head avatar generation from a single image with point-based neural rendering

(1) head mesh reconstruction; (2) vertex feature transformer; (3) neural point rendering



CVTHead: Head mesh reconstruction



- FLAME [1] Parametric head model:
  - $M(\beta, \phi, \theta)$
  - face shape β, expression ϕ, head pose θ

- pre-trained DECA [2] and hair deformation model [3] (optional) to obtain mesh vertices:

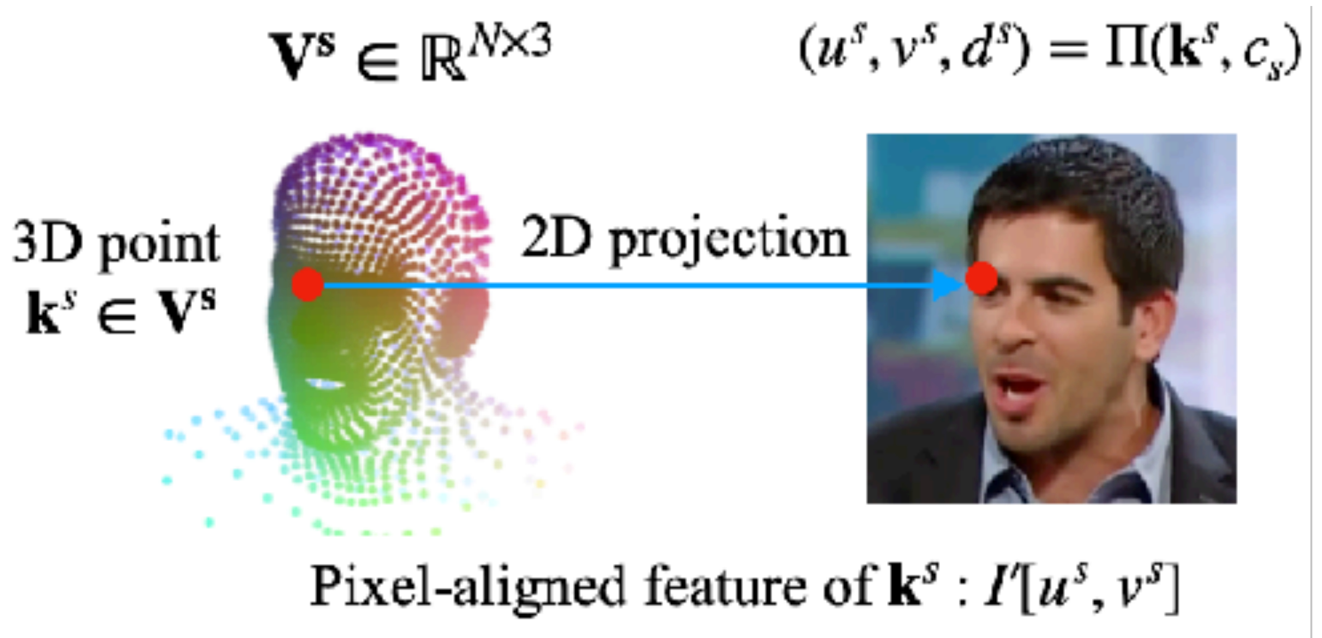
$$\mathbf{V^s} = M(\beta^s, \phi^s, \theta^s) + f_H(\mathbf{I^s}) \in \mathbb{R}^{N\times3}$$

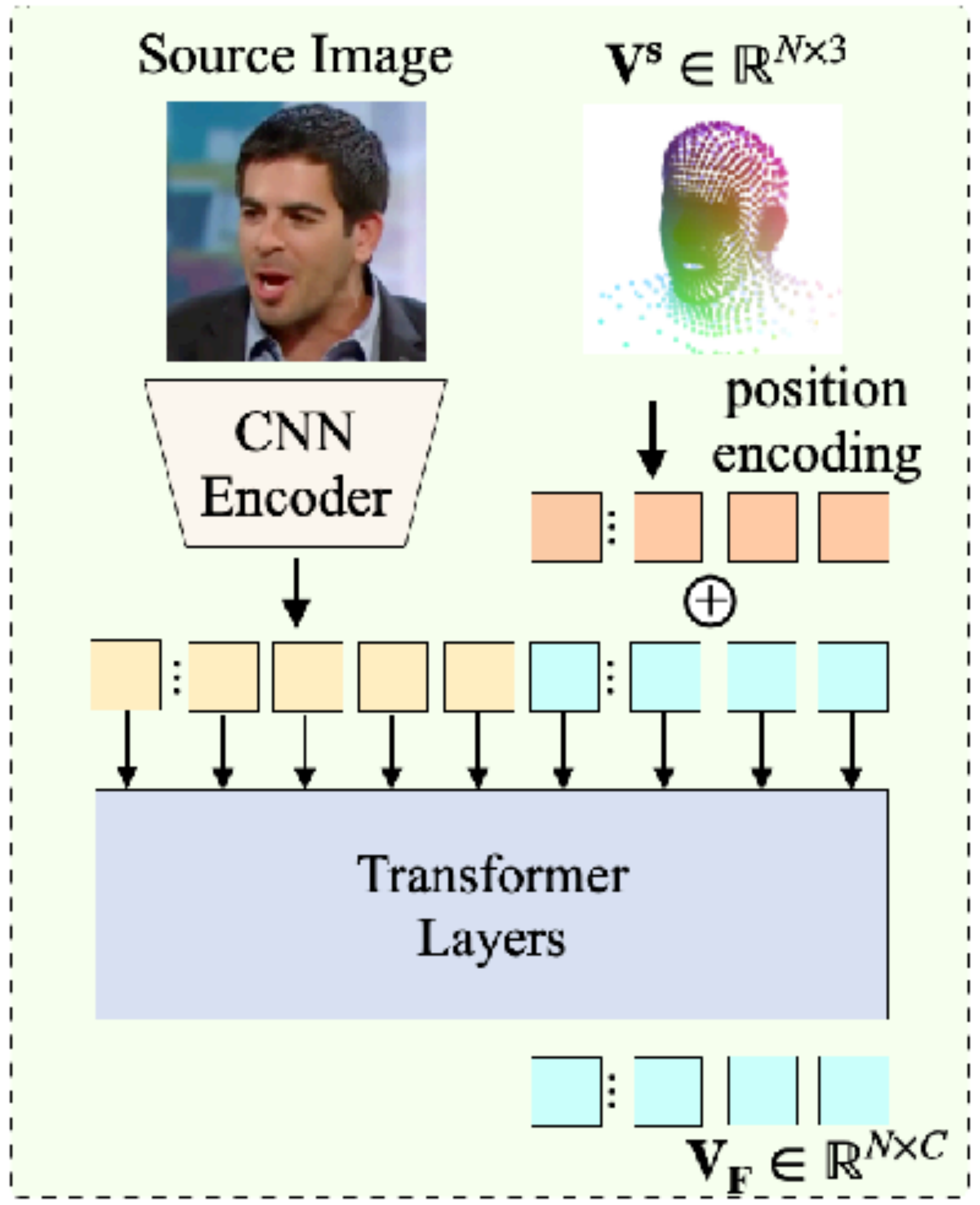
$$\mathbf{V^d} = M(\beta^s, \phi^d, \theta^d) + f_H(\mathbf{I^s}) \in \mathbb{R}^{N\times3}$$

[1] Li, Tianye, et al. "Learning a model of facial shape and expression from 4D scans." *TOG, 2017*
[2] Feng, Yao, et al. "Learning an animatable detailed 3D face model from in-the-wild images." TO*G, 2021*
[3] Khakhulin, Taras, et al. "Realistic one-shot mesh-based head avatars." *ECCV*, 2022

CVTHead: Vertex feature transformer

---- Obtain feature vector of each vertex in the canonical space from source image





Vertex feature Transformer

3D point $\mathbf{k}^s \in \mathbf{V^s}$
2D projection $(u^s, v^s, d^s) = \Pi(\mathbf{k}^s, c_s)$

Limitations of pixel-aligned features [1]:

- require accurate 3D mesh to locate 2D pixels
- misleading feature for occluded 2D projections

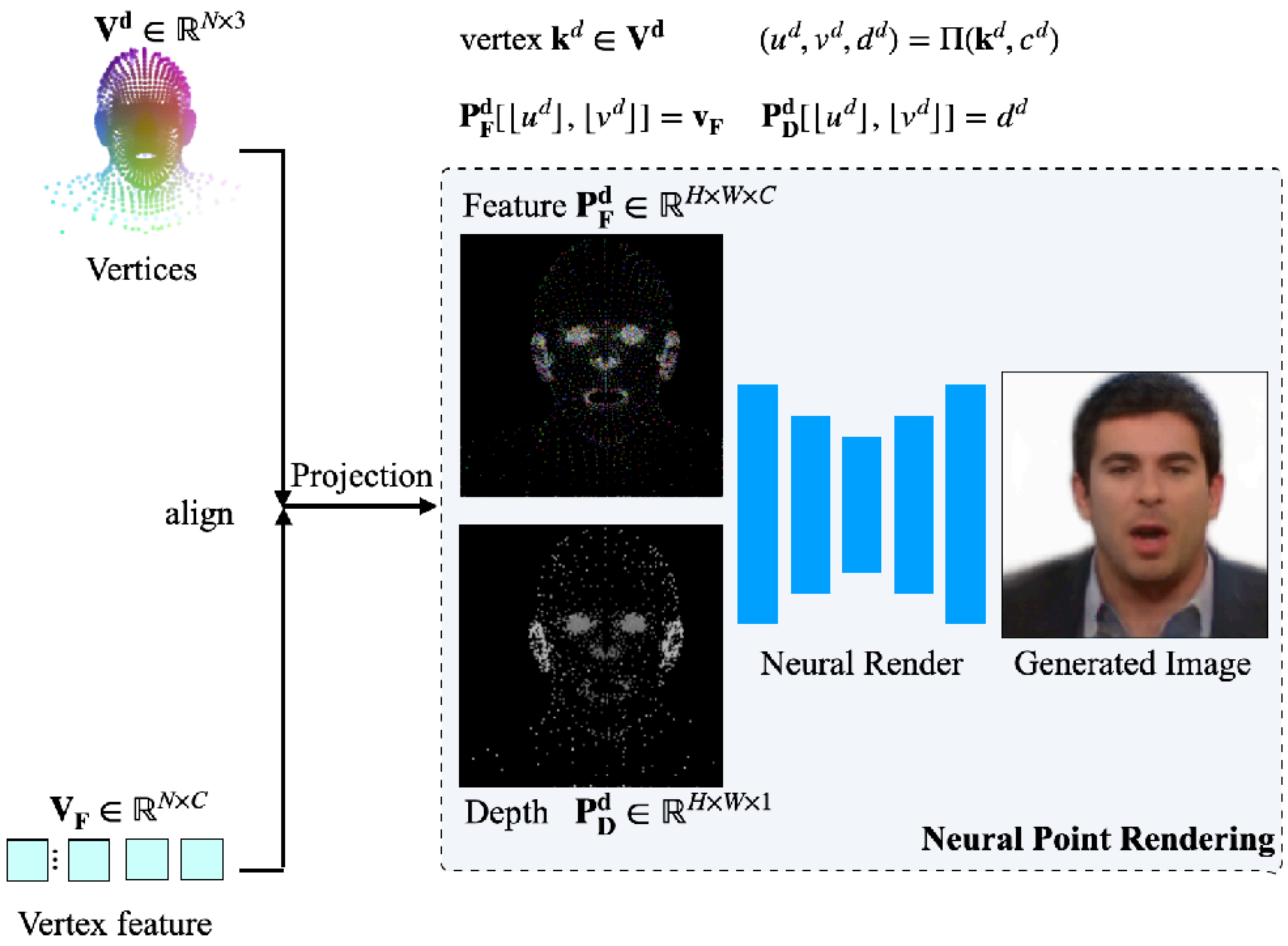
Vertex feature as learnable token $\mathbf{X_v} \in \mathbb{R}^{N\times C'}$
2D projection as positional encoding $(u^s, v^s, d^s) \rightarrow \mathbf{E^s_{uv}}, \mathbf{E^s_{dep}}$
transformer inputs: vertex token & image token

Benefits:

- solve the limitation of pixel-aligned features
- long-range correspondence among all vertex features

[1] Saito, Shunsuke, et al. "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization." *ICCV*. 2019.

CVTHead: Neural vertex rendering



3D point $\mathbf{k}^d \in \mathbf{V^d}$ and corresponding 2D projection $(u^d, v^d, d^d) = \Pi(\mathbf{k}^d, c^d)$

- vertex projection features $\mathbf{P^d_F} \in \mathbb{R}^{H\times W\times C}$

$$\mathbf{P^d_F}[\lfloor u^d \rfloor, \lfloor v^d \rfloor] = \mathbf{v_F}$$

- generate synthetic image $\hat{\mathbf{I}}^{\mathbf{d}}$ and binary foreground mask $\hat{\mathbf{M}}^{\mathbf{d}}$ with a U-Net $\mathcal{G}(\cdot)$

$$(\hat{\mathbf{I}}^{\mathbf{d}}, \hat{\mathbf{M}}^{\mathbf{d}}) = \mathcal{G}([\mathbf{P^d_F}, \mathbf{P^d_D}])$$

- get rid of tedious differentiable rendering

Benefits of CVTHead

- One-shot
  - a single reference image (v.s. multi-view or video inputs for NeRF-based methods)
  - no fine-tuning or optimization for unseen subjects
- Efficiency
  - a single forward for rendering (v.s. hundreds of forwards per ray for volumetric rendering)
- Generalize well on diverse head poses
  - warping-based methods only work well for a limited range of head pose

Results: Face Reenactment

Comparable performance to state-of-the-art graphics-based methods
Better efficiency

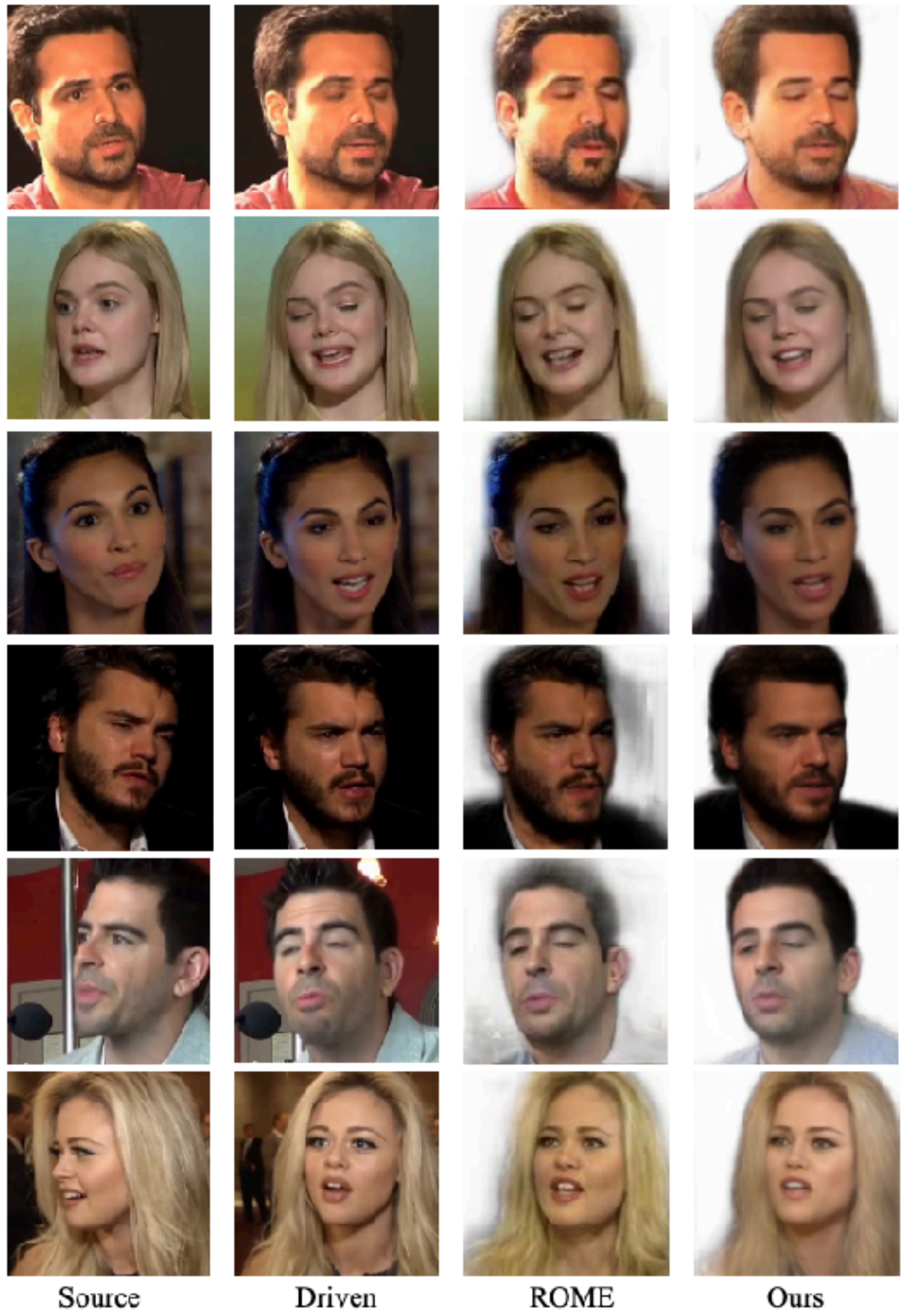
Dataset	VoxCeleb1			
Method	L1 ↓	PSNR ↑	LPIPS ↓	MS-SSIM ↑
FOMM [49]	0.048	22.43	0.139	0.836
Bi-Layer [70]	0.050	21.48	0.108	0.839
ROME [31]	0.048	21.13	0.116	0.838
Ours	0.041	22.09	0.111	0.840
Dataset	**VoxCeleb2**			
Method	L1 ↓	PSNR ↑	LPIPS ↓	MS-SSIM ↑
FOMM [49]	0.059	20.93	0.165	0.793
ROME [31]	0.050	20.75	0.117	0.834
Ours	0.042	21.37	0.119	0.841

Table 1. Results of self-reenactment on the VoxCeleb1 and VoxCeleb2 (↑ means larger is better, ↓ means smaller is better.)

Dataset	VoxCeleb1			
Method	FID ↓	CSIM ↑	IQA ↑	FPS ↑
FOMM [49]	39.69	0.592	37.00	64.3
Bi-Layer [70]	43.8	0.697	41.4	20.1
ROME [31]	29.23	0.717	39.11	12.9
Ours	25.78	0.675	42.26	24.3
Dataset	**VoxCeleb2**			
Method	FID ↓	CSIM ↑	IQA ↑	FPS ↑
FOMM [49]	61.28	0.624	36.20	64.3
ROME [31]	53.52	0.729	37.34	12.9
Ours	48.48	0.712	40.27	24.3

Table 2. Results of cross-identity reenactment.

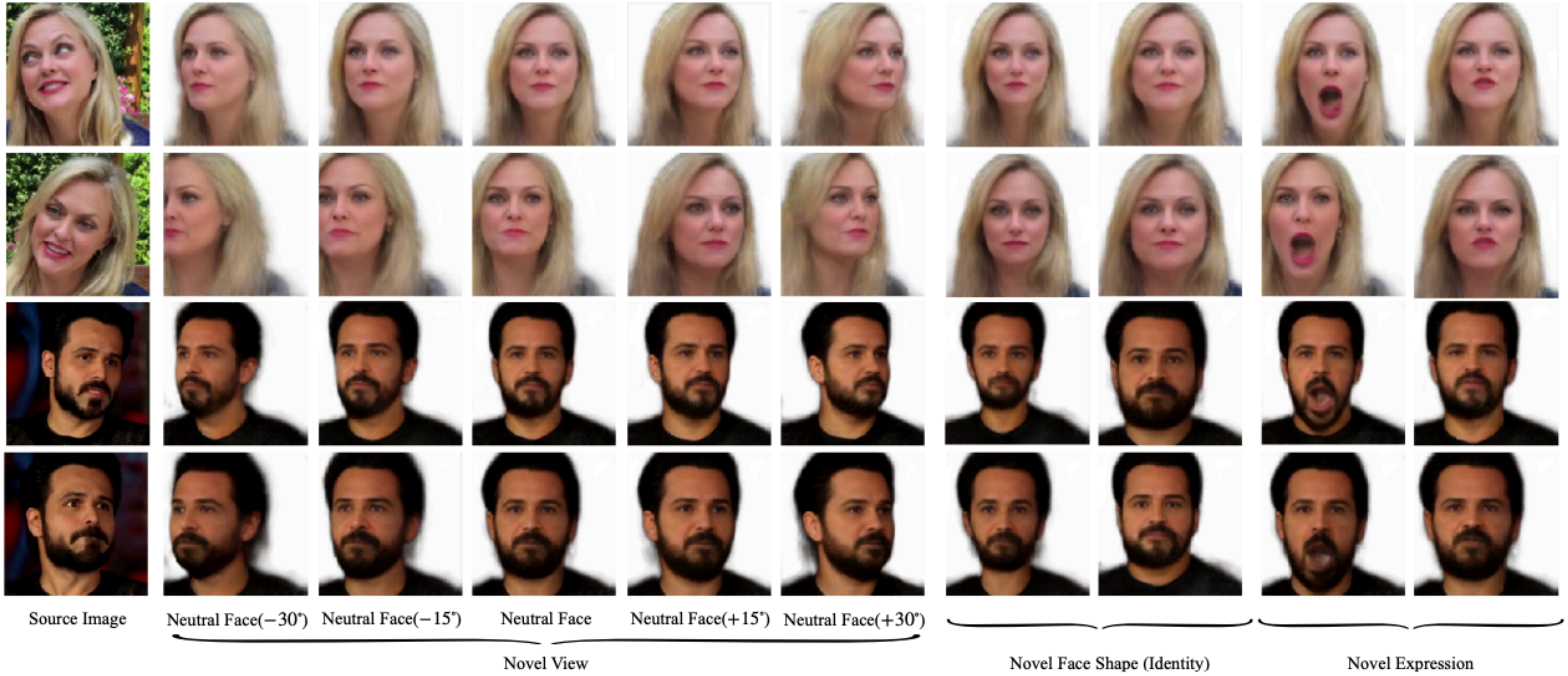
Results: Face Reenactment



self-reenactment

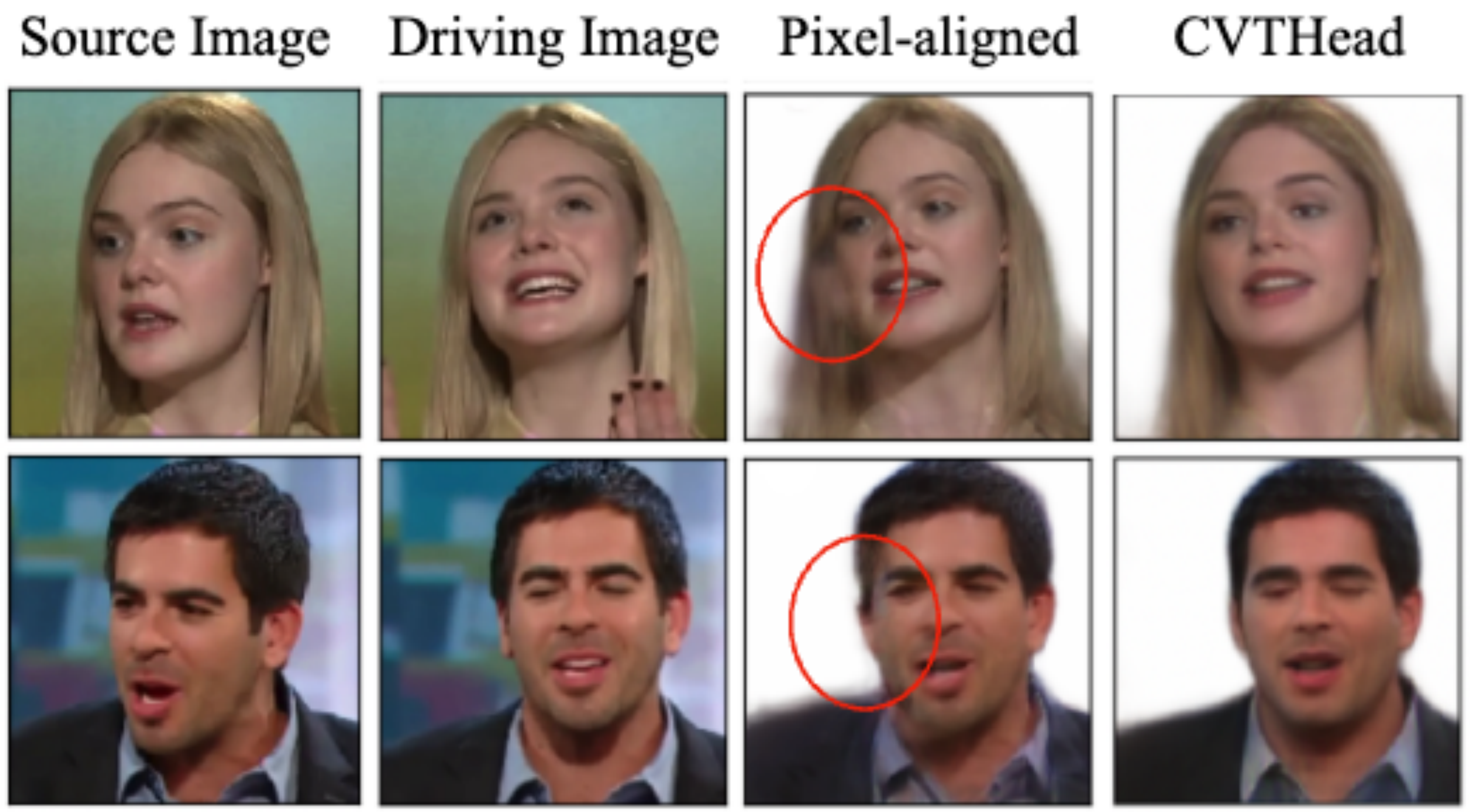
cross-identity reenactment

Results: 3DMM-based Face Animation



face animation with novel views, novel face shapes, and novel expressions

Ablation Study: Comparisons with pixel-aligned features



Method	L1 ↓	PSNR ↑	LPIPS ↓	MS-SSIM ↑
Pixel-aligned features	0.045	21.81	0.107	0.841
CVTHead	0.041	22.09	0.111	0.840

Thanks!

Paper ID: 216