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

Haoyu Ma, Tong Zhang, Shanlin Sun, Xiangyi Yan, Kun Han, Xiaohui Xie
University of California, Irvine
Paper ID: 216

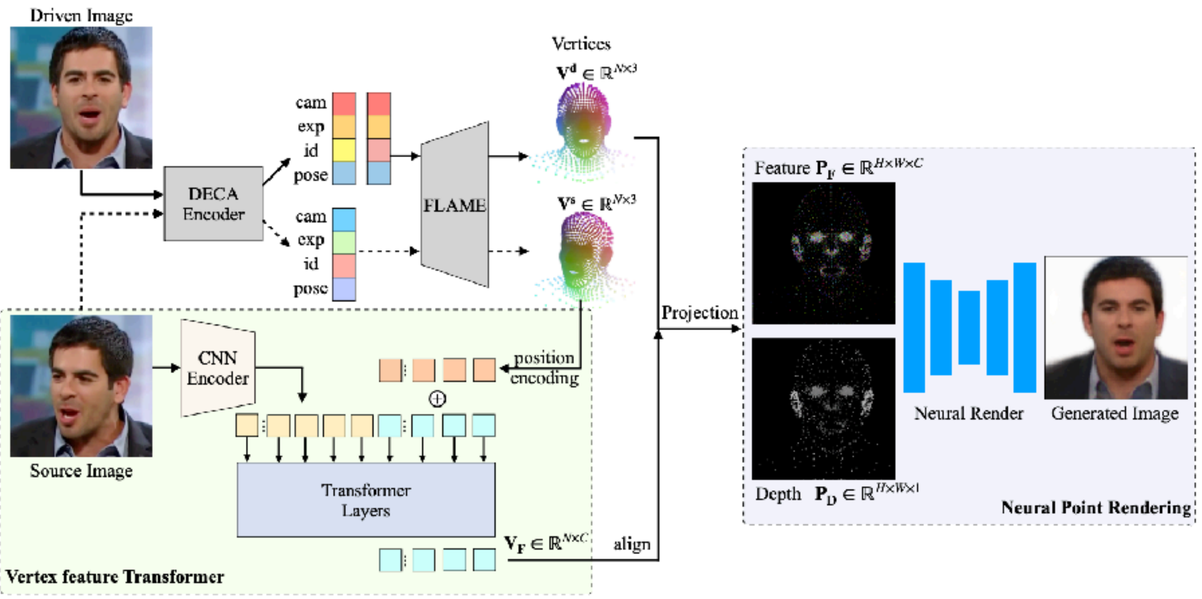
❖ Introduction

CVTHead: efficient and controllable head avatar generation from a single image with point-based neural rendering

❖ Drawbacks of current methods

- Neural head avatar
  - require video inputs or multi-view images
  - subject-specific
- Mesh-guided one-shot face reenactment
  - warp-based: only work for a limited range of head pose
  - graphics-based: tedious differentiable rendering

❖ Methodology



- Head mesh reconstruction:
  - pre-trained DECA [2] & deformation model [1] (optional)
  - source identity & driven expressions & pose
- Vertex-feature transformer:
  - pixel-aligned features may lead to misleading feature for invisible/occluded 3D points
  - vertex token + image token
- Neural vertex rendering:
  - project vertices and corresponding feature descriptors onto the vertex feature image $\mathbf{P_F}$ and depth image $\mathbf{P_D}$
  - Neural render: $\hat{\mathbf{I}} = \mathcal{G}(\mathbf{P_F}, \mathbf{P_D})$

❖ Results

- Quantitative results on talking face synthesis

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

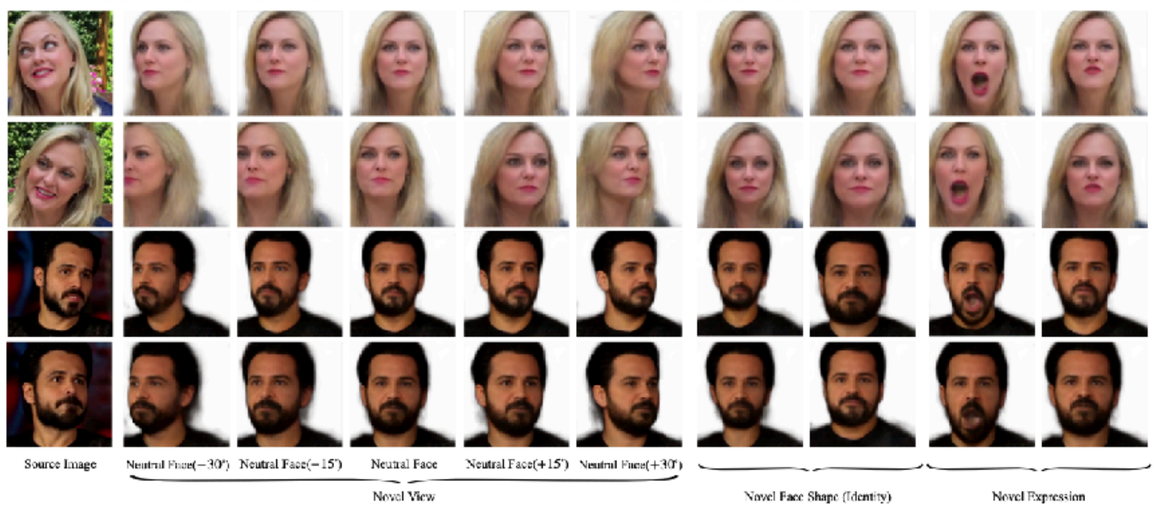
Table 1. Results of self-reenactment on the VoxCeleb1 and VoxCeleb2 (↑ means larger is better, ↓ means smaller is better.)

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.

- Qualitative results

self-reenactment cross-reenactment



3DMM-based face animation with novel views, identity, and expressions

❖ Reference

[1] Khakhulin, Taras, et al. "Realistic one-shot mesh-based head avatars." *ECCV*, 2022.
[2] Feng, Yao, et al. "Learning an animatable detailed 3D face model from in-the-wild images." TO*G, 2021*

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

Contact: *haoyum3@uci.edu*