Tsinghua University
Tsinghua Shenzhen International Graduate School

DT-NeRF: Decomposed Triplane-Hash Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis

ICASSP 2024 KOREA

Yaoyu Su, Shaohui Wang, Haoqian Wang

Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

Introduction

Contributions

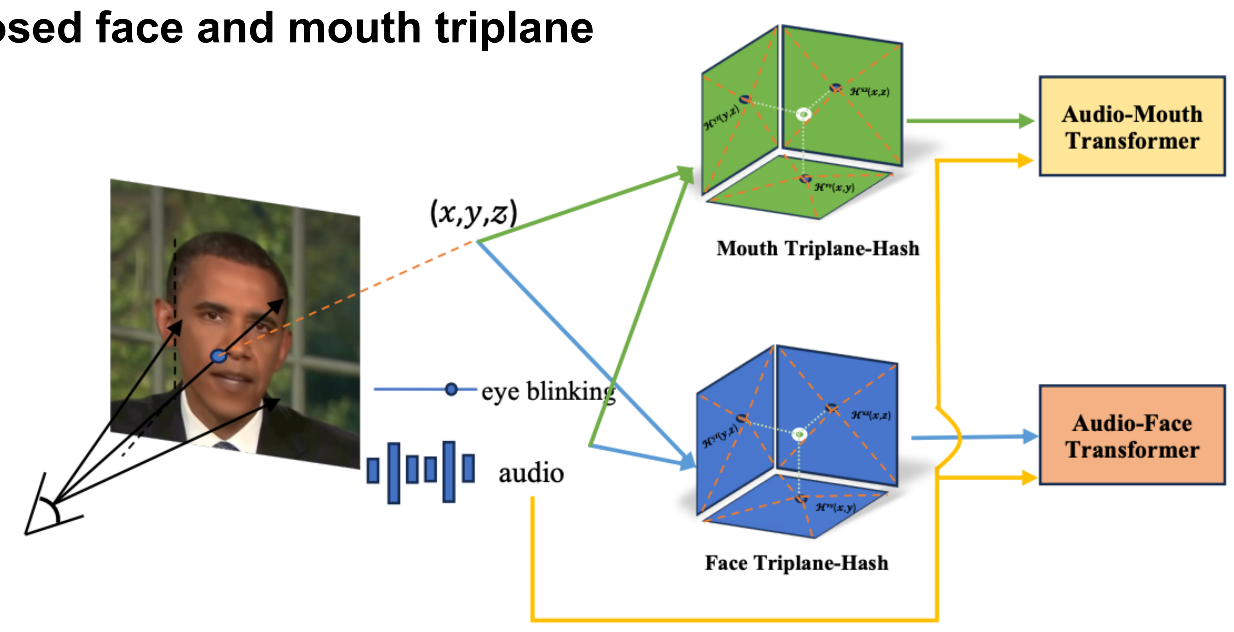
1. **Decomposed Triplane-Hash Representation**: Specifically designed for the mouth and facial areas, it captures the details of facial expressions driven by audio.
2. **Audio-mouth-face-align transformer:** Utilized audio feature as query vector within a transformer model to accurately align the audio cues with coordinate space of the talking portrait.
3. **Spatial Fusion in Volumetric Rendering**: Enhances facial information, ensuring the animation reflects true lip synchronization and expressions.

Motivations

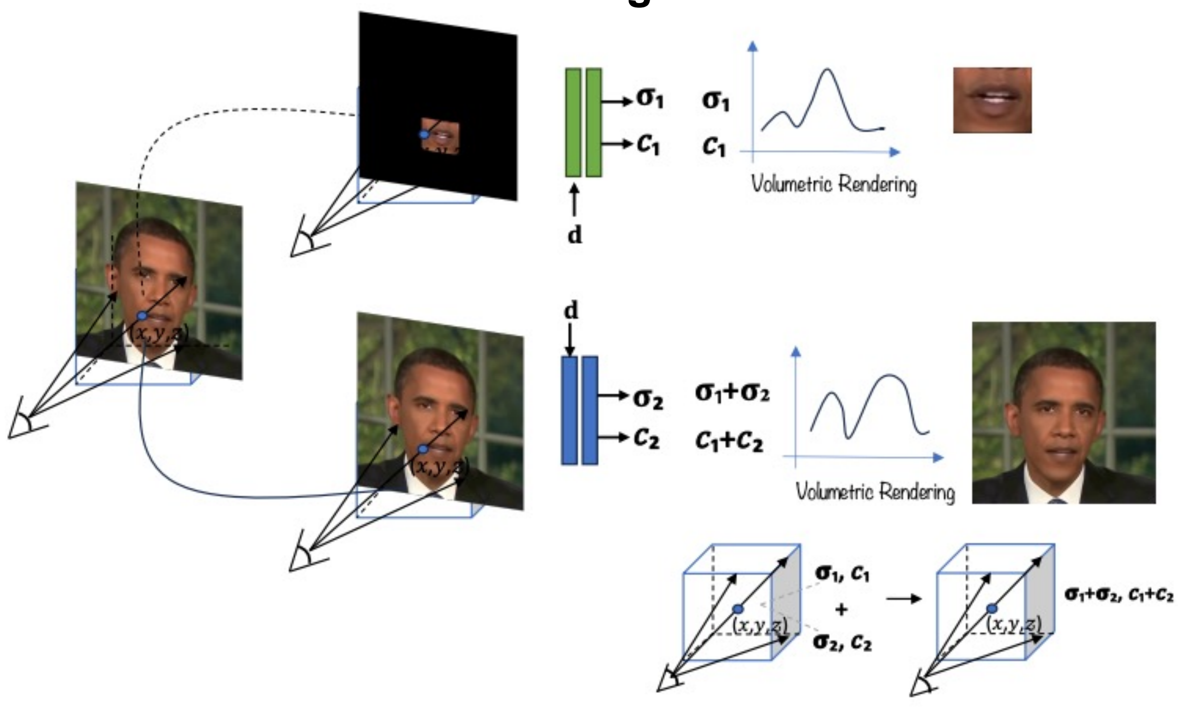
Challenges:
1.Effectively synchronizing audio signals with facial and mouth dynamics.
2.Improving the representation of the mouth and face to achieve high-quality, real-time audio-driven facial synthesis.
Approaches:
1.Utilizing audio features as query vectors, spatial coordinates as key vectors, spatial points features as value vectors for a transformer to align the audio with features of the spatial points. This aims to optimize the density and color networks in NeRF, facilitating the transition from a canonical space to a dynamic space.
2.Utilizing the additive properties of color and volumetric density within the same NeRF space to achieve a seamless integration of mouth and face triplane-hash representation.

Methods

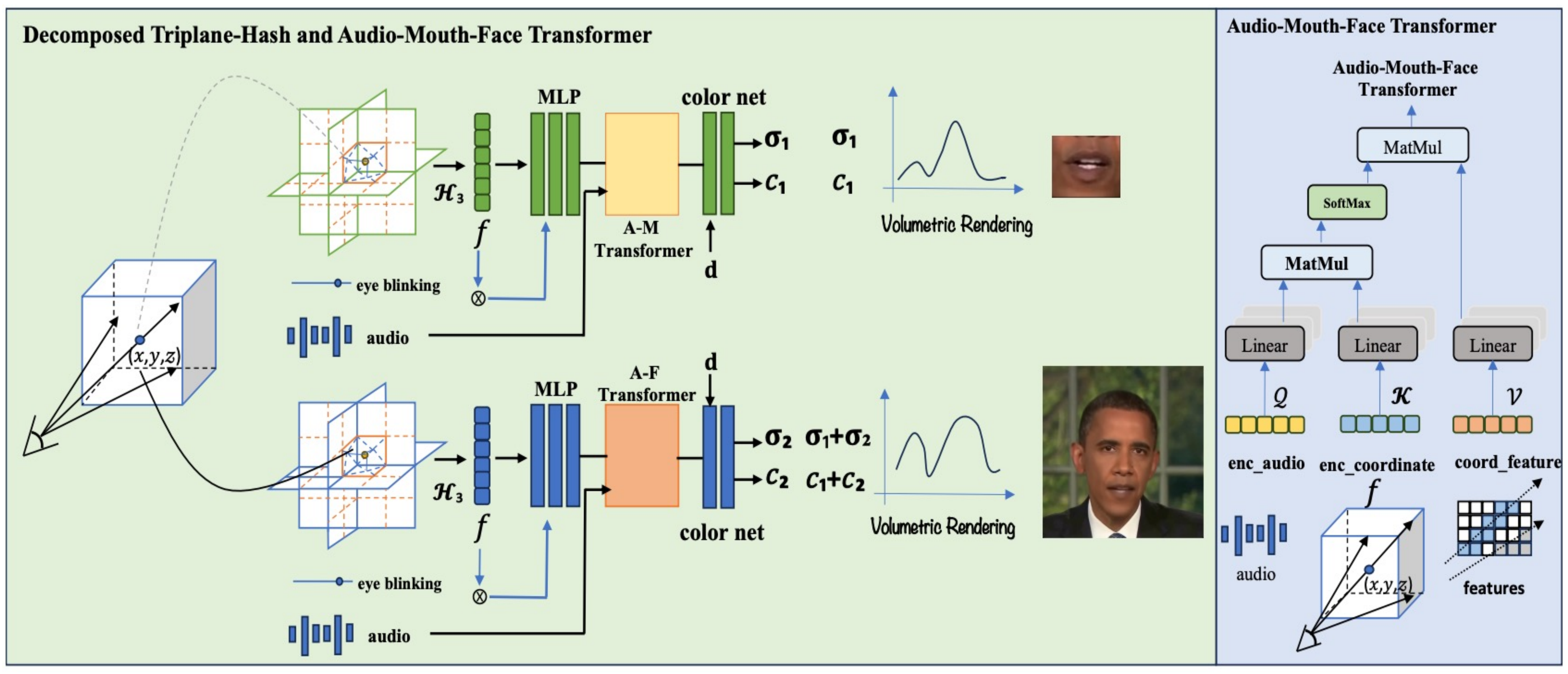
Decomposed face and mouth triplane



Spatial fusion in volumetric rendering



Pipeline



Audio-mouth-face align transformer

$$\mathcal{F}^{C}:(\mathbf{a},\mathbf{x},\mathbf{x\_feat};\mathcal{Q},\mathcal{K},\mathcal{V})\rightarrow(\mathbf{c},\sigma),$$
$$\mathbf{q}=\mathcal{Q}(\mathbf{a}),$$
$$\mathbf{k}=\mathcal{K}(\mathbf{x}),$$
$$\mathbf{v}=\mathcal{V}(\mathbf{x\_feat}),$$
$$\mathrm{attn}=\mathrm{Softmax}(q\cdot k)\cdot\mathbf{v},$$

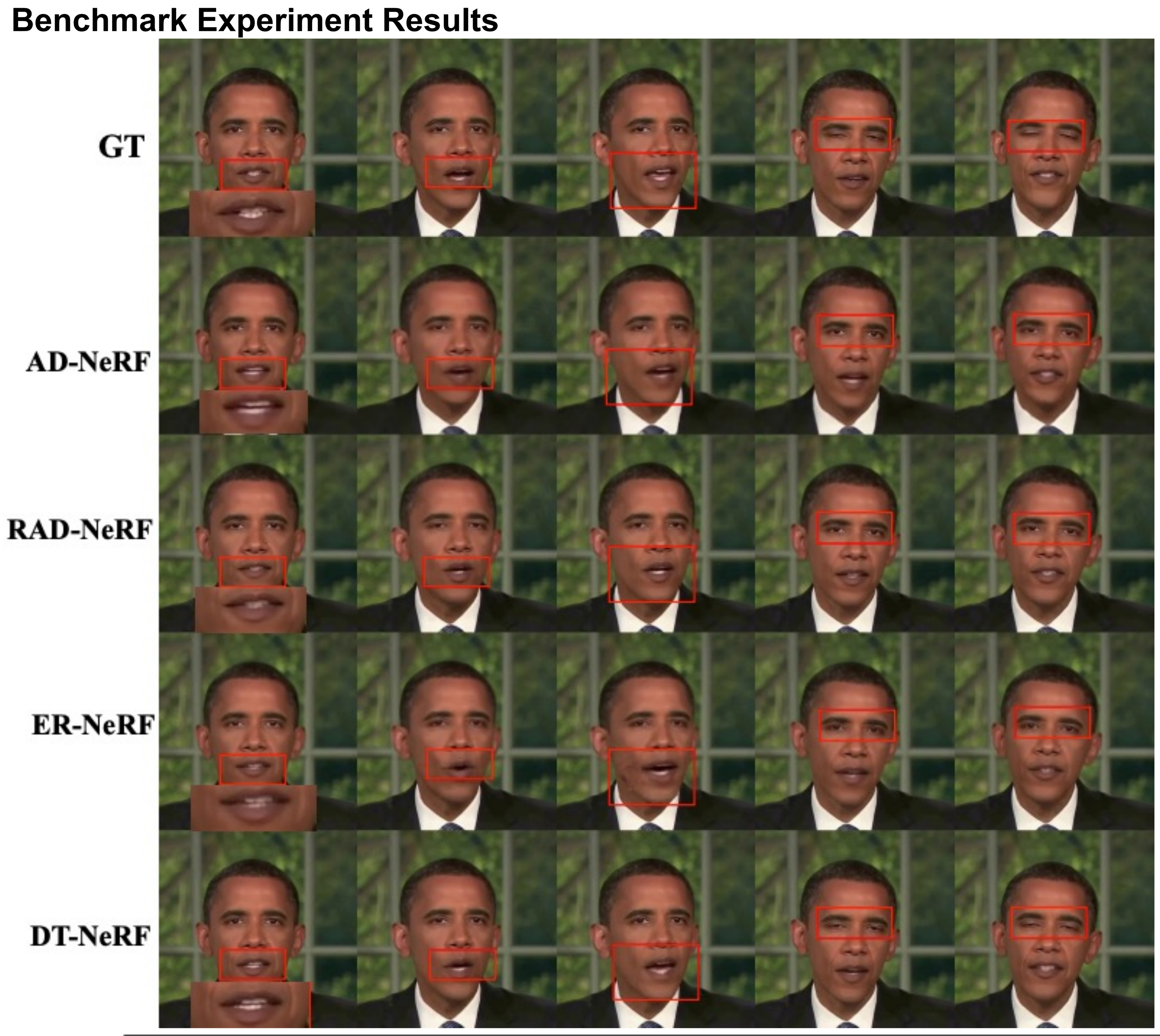
Loss Function: Two stage fine-tune

$$\mathcal{L}_{coarse}=\sum_{i\in I_{face}}\left\|C(\mathrm{i})-\hat{C}(\mathrm{i})\right\|_2^2+\lambda\sum_{j\in I_{mouth}}\left\|C(\mathrm{j})-\hat{C}(\mathrm{j})\right\|_2^2$$

$$\mathcal{L}_{fine}=\sum_{i\in\mathcal{P}}\left\|C(\mathrm{i})-\hat{C}(\mathrm{i})\right\|_2^2+\lambda\,\mathrm{LPIPS}(\hat{\mathcal{P}},\mathcal{P})$$

Experiments

Benchmark Experiment Results



Methods	PSNR ↑	LPIPS ↓	FID ↓	LMD ↓	Time	FPS
Ground Truth	N/A	0	0	0	-	-
AD-NeRF [15]	30.75	0.1034	24.514	3.345	18h	0.08
RAD-NeRF [21]	34.00	0.0387	10.835	2.696	5h	32
ER-NeRF [22]	35.37	0.0185	9.675	2.604	**2h**	**34**
DT-NeRF(Ours)	**35.39**	**0.0169**	**9.472**	**2.601**	2.5h	32

Generalization Experiment

Methods	PSNR ↑	LPIPS ↓	FID ↓	LMD ↓	Time	FPS
Ground Truth	N/A	0	0	0	-	-
ER-NeRF [22]	**30.80**	0.054	12.110	5.54	**2h**	**34**
DT-NeRF(Our)	30.45	**0.048**	**11.274**	**5.34**	2.5h	32

Ablation Study

Methods	PSNR ↑	LPIPS ↓	FID ↓	LMD ↓	Time
Ground Truth	N/A	0	0	0	-
w/o T w/o F	35.35	0.0362	10.287	2.687	1.5h
w/o T w F	35.17	0.0173	**9.22**	2.661	2.5h
w/o S w/o F	**35.54**	0.0381	10.949	2.663	1.5h
w/o S w F	35.21	0.0172	9.550	2.662	2.5h
w T w S w F	35.39	**0.0169**	9.472	**2.601**	2.5h

T: transformer, F: finetune, S: space fusion

Acknowledgements: This research was funded through National Key R&D Program of China (No. 2022YFB36066) and the Shenzhen Science and Technology Project (Grants JCYJ20220818101001004, JSGG20210802153150005).