Efficient Region-Aware Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis

Supplementary Material

A. Overview

In the supplemental document, we introduce the details of our torso-nerf with Adaptive Head Encoding, model architecture details, user study details, additional experiments and analysis, ethical considerations, and the discussion of this work.

B. Torso-NeRF Details

We combine the proposed Adaptive Pose Encoding and the 2D deformable neural field from RAD-NeRF [14] to render the torso part. As described in Section 3.4 of the main paper, we init three points in the 3D canonical space with trainable homogeneous coordinates:

$$\mathbf{X}_{keys} = (\mathbf{x}_{keys}, \mathbf{y}_{keys}, \mathbf{z}_{keys}, \mathbf{1})^T \in \mathbb{R}^{4\times 3}. \quad (1)$$

For each frame, we form the pose of head $\mathbf{P}$ as:

$$\mathbf{P} = \begin{pmatrix} \mathbf{R} & \boldsymbol{t} \\ 0 & 1 \end{pmatrix} \quad (2)$$

and apply it to transform the key points:

$$\hat{\mathbf{X}}_{keys} = \mathbf{P}^{-1}\mathbf{X}_{keys}. \quad (3)$$

where $\hat{\mathbf{X}}_{keys}$ is the transformed coordinates. Then we convert $\hat{\mathbf{X}}_{keys}$ to the ordinary coordinates and project them onto the plane $\mathbf{Z} = 1$ to calculate their 2D coordinates $\bar{\mathbf{X}}_{keys} \in \mathbb{R}^{2\times 3}$ on the imaging plane, where

$$\bar{\mathbf{X}}_{keys}(i,j) = \gamma \cdot \hat{\mathbf{X}}_{keys}(i,j)/\hat{\mathbf{z}}_{keys}(j), \quad (4)$$

and γ is the coefficient learned by the network.

The overview of the torso-NeRF is shown in Figure 1. We use $\bar{\mathbf{X}}_{keys}$ to condition the 2D deformable neural field [14] for rendering the pixel-wise color and alpha of the torso at the image pixel coordinate $\mathbf{x}_{pixel}$. Specifically, to render the pixel at $\mathbf{x}_{pixel} \in \mathbb{R}^2$ on the image , we firstly feed $\bar{\mathbf{X}}_{keys}$ and the pixel coordinate $\mathbf{x}_{pixel}$ into an MLP, and add the output $\Delta\mathbf{x}$ to $\mathbf{x}_{pixel}$ for a 2D deformation. The deformed coordinate is then encoded by the 2D multiresolution hash encoder $\mathcal{H}^t$. Finally, another MLP is used to calculate the pixel-wise transparency α and color $\mathbf{c}_t$.

The implicit function of the torso-NeRF can be formulated as:

$$\mathcal{F}^{\mathcal{T}} : (\mathbf{x}_{pixel}, \bar{\mathbf{X}}_{keys}; \mathcal{H}^t) \rightarrow (\mathbf{c}_t, \alpha) \quad (5)$$

During training, the coordinates $\mathbf{X}_{key}$ can be optimized to gain the ability in representing the implicit relationship between the poses of the head and torso. And due to only linear transformations involved during forwarding, the torso quality is improved without a significant increase in the amount of calculation.

Methods	AD-NeRF	RAD-NeRF	**ER-NeRF**
Stability	1.33	2.89	**3.89**
Image Quality	2.67	3.33	**4.00**

Table 1. **User Study of Torso Quality.** The rating is of scale 1-5, the higher the better.

User Study. We also conduct a user study to evaluate the synthesized torso part. We invite the attendees to rate the stability and image quality of generated torsos in the *head reconstruction setting*. To compare our method, we selected AD-NeRF [6] and RAD-NeRF [14] as the baselines since they are the only two NeRF-based methods that can synthesize the torso part and have released their codes. The results are reported in Table 1. We can observe that our ER-NeRF achieves the best both on Stability and Image Quality by just adding a straightforward encoding step *without any deep neural network*, which demonstrates the high efficiency of our Adaptive Pose Encoding.

C. Architecture Details

Audio Feature Extractor. In the experiments, we use the pretrained *DeepSpeech* [7] model to extract raw audio features. We then process these features with the same audio attention module as previous NeRF-based works [6, 11, 14], except for changing the output dimension from 64 to 32.

Region Attention Module. The speech audio branch utilizes an attention vector MLP with 2 layers and 64 hidden dimensions. Conversely, the eye-blinking branch employs a 2-layer MLP with only 16 hidden dimensions.

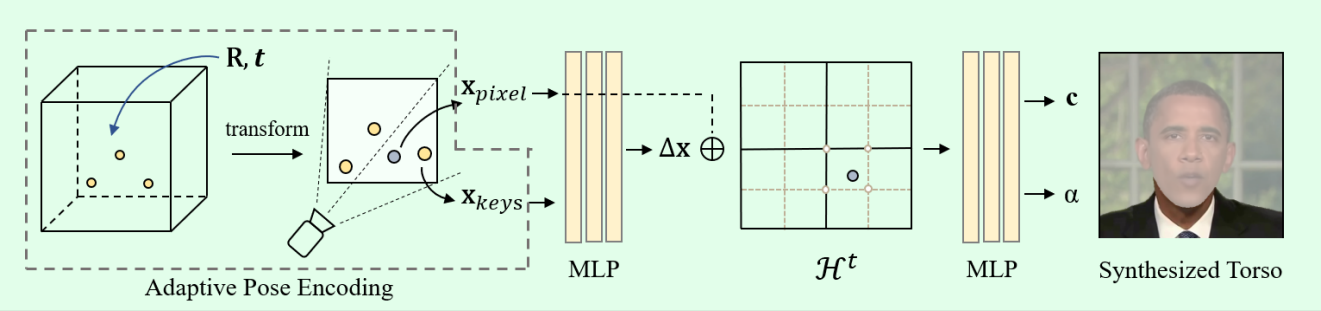


Figure 1. **Overview of the Torso-NeRF**.

Grid	Instant-NGP	Tri-Hash			
		Frontal	Side 1	Side 2	Total
Collision	835186	138345	31041	26048	195434

Table 2. The number of hash collisions occurring in one feature lookup step on a single grid resolution.

Tri-plane Hash Representation. The 2D hash encoders are configured to have 14 resolution levels and a single entry assigned to each level, with a range of multiple resolutions from 64 to 512. The density MLP decoder contains 3 layers, and the color MLP decoder contains 2 layers, both of which have 64 hidden dimensions.

D. User Study Details

The study involves 18 participants with an age range of 20-30 years old. To facilitate more accurate judgments, we combine all generated videos and the ground truth into a single high-resolution video. This allows participants to observe all motions simultaneously. To ensure fairness in the comparison process, we assign a number to each generated result instead of identifying them by their method. Participants are asked to evaluate the three perspectives of the generated portraits: (1) Lip-sync Accuracy; (2) Video Realness; (3) Image Quality. To evaluate the torso-NeRF, we additionally invite the attendees to judge two aspects of the synthesized torso: (1) Stability; (2) Image Quality.

E. Tri-Plane Hash Representation

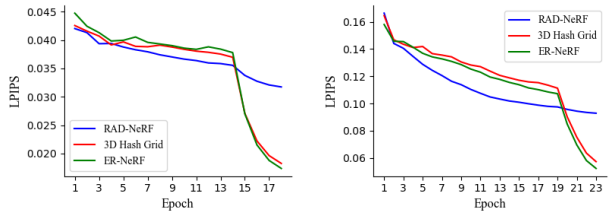
Complexity of Hash Collision Here we give the proof of the complexity $O(R^2 + 2RN)$ in Section 3.2 for our Tri-Hash Representation: 1) For the frontal plane, the projected area is linearly correlated to R^2, thus the collision is $O(R^2)$; 2) The ideal projected area for the other two side planes is $(\lambda R)R$, where λ is an adjustment. But notice only the nearest N points can be sampled at some side areas due to occlusion, so λR is partly correlated to N, and the collision is $O(\lambda R^2 + RN)$. Overall, $O(R^2 + 2RN)$ is given.

The Number of Hash Collisions. Here we give the evaluation during one lookup step to directly verify our effect on hash collision reduction. The hashtable size is set to 2^{14} and divided by 3 for each planar grid in our Tri-Hash, with the grid resolution of 512, the max in the experiment. Adjustments of $1/8$ and $1/4$ are applied due to bilinear interpolation. The point coordinates are scaled up to encourage uniform hashing. In practice, the benefit of our method would be more obvious, since indeed the coordinates cannot be uniformly separated among the hash table and so the overlapping of grids becomes more serious.

F. Additional Experiments

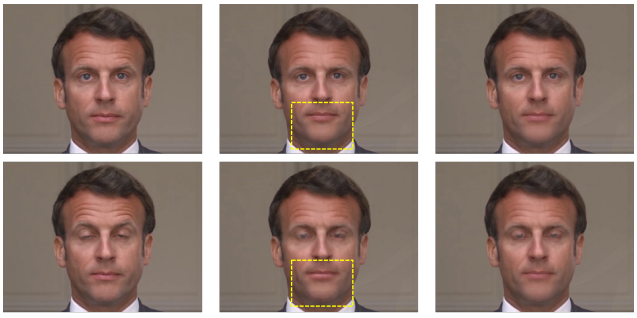
LPIPS Finetune. It may seem counter-intuitive that the overall LPIPS [16] finetuning is less effective for RAD-NeRF [14] but has a significant impact on the high-frequency details of our ER-NeRF despite having a smaller model size. This phenomenon is likely due to differences in training difficulty. Our ablation study shows that even a simplified architecture with only a 3D hash grid backbone and an audio feature dimension of 32 can reproduce fine details. On the other hand, RAD-NeRF uses a more complex architecture with an additional hash grid and higher-dimensional audio features to improve lip-sync performance, which increases the training difficulty and makes the network harder to optimize. As a result, the LPIPS finetuning has a weaker impact on its rendering quality. The variations in LPIPS loss during training are illustrated in Figure 2.

Region Attention for Eye Blinking. We perform an ablation study on the eye-blinking branch of the Region Attention Module in isolation. When we skip the region attention mechanism and directly concatenate the AU45 with the input of the MLP decoder, some unnatural facial movements appear, like jittering and unreasonable lip movements with eye blinking (Figure 3). This might be due to the module's inability to accurately identify the regional impact of eye blinking and thus learns an incorrect motion mapping with other facial regions. The results indicate that our Region



(a) Obama (b) May

Figure 2. The validation LPIPS loss on our Obama dataset and May dataset with different architectures. A complex network is much harder to be optimized by the LPIPS finetune and reproduce fine details.



Ground Truth w/o Att. with Att.

Figure 3. **Ablation on Region Attention for Eye Blinking**. Some unnatural facial movements appear when directly concatenating the AU45 with the input to control eye blinking. After applying the proposed region attention mechanism, the robustness has been improved.

Attention Module can help decouple different semantic motions and improve robustness.

G. Comparison with GeneFace and DFRF

In table 3 and 4, we have also compared our ER-NeRF with two current SOTA methods GeneFace [15] and DFRF [11], both of which are designed for different settings, notably. Meanwhile, since the code of GeneFace is released too close to the submission deadline, it was not taken into the baselines in the main paper. We consider the comparisons not entirely fair for them, and the results are just for reference.

H. Additional Qualitative Comparison

We show some additional generated key frames on the Testset A under the *lip synchronization setting* with high resolution in Fig. 4. In this setting, we only synthesize the head part. The results show that our ER-NeRF can outperform most baselines in image quality while retaining a high lip-sync accuracy. We strongly recommend watching our supplemental video for better visualization and more results.

Methods	PSNR ↑	LPIPS ↓	FID ↓	LMD ↓	AUE ↓	Sync ↑
DFRF	30.74	0.0881	13.32	3.553	2.538	4.385
GeneFace	30.24	0.0817	11.16	3.496	2.854	5.403
ER-NeRF (Ours)	**33.10**	**0.0291**	**10.42**	**2.740**	**1.629**	**5.708**

Table 3. DFRF and GeneFace at the *head reconstruction setting*.

Methods	A: LMD ↓	A: Sync ↑	B: LMD ↓	B: Sync ↑
DFRF	6.551	4.854	8.126	4.127
GeneFace	**5.465**	5.849	**7.237**	6.275
ER-NeRF (Ours)	6.254	**6.242**	8.150	**6.830**

Table 4. DFRF and GeneFace at the *lip synchronization setting*.

I. Ethics Considerations

Our proposed ER-NeRF synthesizes high-fidelity talking portraits with accurate lip-audio synchronization. The generated portrait video is highly realistic and difficult for people to distinguish fake from real. We hope it can facilitate a wide range of applications, such as digital humans, video production, and human-computer interaction assistance. On the other hand, however, such techniques may be misused for malicious purposes and make harm. It's significant to tell the users whether a video is real or fake. Recent studies have already achieved success in deepfake detection for face swapping, reenactment and other generating videos [5, 17, 3, 2, 12, 4], but it remains a challenge to discriminate synthesized high-fidelity portraits from recent NeRF-based methods. Besides sharing our generated results to the deepfake detection communication and to help develop more powerful deepfake detectors, we also provide some possible perspectives to fight against the malicious use of talking portrait synthesis:

- **Protect real portrait speech videos**. Since current NeRF-based techniques rely heavily on specific training videos, protection for these real videos is valid to prevent misuse. For example, we can add digital watermarks to the portrait part which can be easily detected even in the generated fake videos.

- **Limit the use of deepfake techniques**. Nowadays, little cost of deepfakes leads to an unconstrained use of these techniques. The negative impact of the malicious use of deepfakes can be amplified when they are unintentionally created and shared by the public on social media platforms. Even though the creators may have no malicious intent, the spread of these deepfakes can still have harmful consequences. We suggest the laws should state how to properly make use of these face-generation techniques.

Figure 4. **Additional Qualitative Comparisons**. We show the synthesized head results of the *lip synchronization setting* on Testset A. **(a)** Ground truth; **(b)** AD-NeRF [6]; **(c)** SynObama [13]; **(d)** RAD-NeRF [14]; **(e) ER-NeRF (ours)**; **(f)** LSP [9]; **(g)** SSP-NeRF [8]; **(h)** Wav2Lip [10]; **(i)** PC-AVS [18].

On the other hand, the public should also be aware of the potential harm of deepfakes and treat them cautiously.

J. Limitation and Future Work

Compare to the one-shot methods like Wav2Lip [10], our method has some advantages in results quality and resolution, however, needs per-scene training when generating new target portraits. Enabling the generative ability may be the target we work for.

Besides, the proposed method has two main limitations. Firstly, our method still encounters a challenge with the small scale of a single training video, leading to a weak lip-audio synchronization with out-of-domain audio, such as some cross-lingual speech or singing voice. Currently, we rely on a pretrained speech recognition model to extract audio features. We have noticed that some recent works [15, 1] employed a pretrained model to enhance their generalizability. In future work, we will consider incorporating priors from large audiovisual datasets to address this limitation. Secondly, although our method has improved the robustness and image quality of the torso part, there remain some blurry regions. We analyze this may be caused by uncertain movements and the form of representation itself. In future work, we will focus on addressing this issue.

References

[1] Aggelina Chatziagapi, ShahRukh Athar, Abhinav Jain, Rohith Mysore Vijaya Kumar, Vimal Bhat, and Dimitris Samaras. Lipnerf: What is the right feature space to lip-sync a nerf. In *International Conference on Automatic Face and Gesture Recognition 2023*, 2023. 5

[2] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022. 3

[3] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 3

[4] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9468–9478, 2022. 3

[5] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 666–667, 2020. 3

[6] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 1, 4

[7] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 1

[8] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 106–125. Springer, 2022. 4

[9] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: Real-time photorealistic talking-head animation. *ACM Trans. Graph.*, 40(6), dec 2021. 4

[10] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 4, 5

[11] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 666–682. Springer, 2022. 1, 3

[12] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 3

[13] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 4

[14] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022. 1, 2, 4

[15] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In *The Eleventh International Conference on Learning Representations*, 2022. 3, 5

[16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2

[17] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 3

[18] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 4