PHƯƠNG PHÁP ER-NERF TRONG BÀI TOÁN TALKING FACE

Presented by:

Nhóm

THÀNH VIÊN

Nguyễn Thị Yến - B22DCCN928

Lê Thị Hải Yến - B22DCCN927

NỘI DUNG

- GIỚI THIỆU CHUNG
- NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH
- DEMO VÀ PHÂN TÍCH THỰC NGHIỆM
- ĐÁNH GIÁ, PHÂN TÍCH VÀ NHẬN XÉT

1. GIỚI THIỆU CHUNG

- **Tiêu đề của bài báo:** Efficient Region-Aware Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis
- **Tác giả:** Jiahe Li, Xiao Bai (Beihang) + Lin Gu (RIKEN/The University of Tokyo)
- **Thời gian công bố:** Paper ER-NeRF được công bố trên arXiv vào tháng 8/2023, phiên bản cập nhật v2 vào tháng 8 cùng năm.
- **Nơi công bố:** ICCV 2023 – hội nghị top 1 lĩnh vực Computer Vision
- **Giá trị nghiên cứu và lý do chọn:**
  - Là paper SOTA mới nhất (2023) về real-time talking portrait NeRF
  - Đạt đồng thời 4 thứ cùng lúc mà các paper trước không làm được:
    - Real-time rendering ≥ 55–60 fps
    - Chất lượng hình ảnh cao nhất hiện tại (PSNR 35.29, LPIPS 0.041)
    - Model nhỏ nhất chỉ ~28 MB
    - Thời gian train nhanh nhất (hội tụ chỉ 3–4 giờ trên 1 GPU 3090)

1. GIỚI THIỆU CHUNG

- **Lịch sử phát triển từ NeRF gốc đến Talking Portrait real-time**

Năm	Paper	Nội dung chính	Hạn chế còn lại
2020	NeRF (Mildenhall et al.)	Phát minh Neural Radiance Fields: MLP dự đoán màu + density từ (x,y,z,θ,ϕ) → volume rendering	Train & inference cực chậm, chỉ làm được cảnh tĩnh
2021	AD-NeRF (Guo et al.)	Paper đầu tiên đưa audio feature vào NeRF → sinh video talking head, tách head/torso riêng	<1 fps, model ~300 MB, train vài ngày
2023	ER-NeRF (Li et al.)	Real-time 60 fps + chất lượng cao nhất + model 28 MB + train 3–4 giờ → SOTA hiện tại	← Đây là paper chính sẽ trình bày

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

1. Giới thiệu bài toán

a. Mục tiêu

- Từ một video ngắn 3–5 phút của một người + một đoạn audio bất kỳ

→ Tạo video người đó nói theo đúng audio mới, giữ nguyên gương mặt, biểu cảm tự nhiên, môi đồng bộ 100%, có thể thay pose, thay background, chạy real-time ≥ 50 fps trên laptop.

b. Bài toán

- **Input:**
  - Video ngắn 512×512 của 1 người
  - Arbitrary driving audio (có thể là tiếng Việt, tiếng Anh, tiếng Nhật…)
- **Output:** Video 512×512 ≥ 30 fps, môi khớp chuẩn, vùng cổ-torso tự nhiên, có chuyển động nhẹ, background có thể thay tuỳ ý.

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

1. Giới thiệu bài toán

c. Động cơ nghiên cứu

- **AD-NeRF** (2021): chất lượng tốt nhưng rất chậm (<1 fps)
- **RAD-NeRF** (2022): lần đầu real-time (~35 fps) nhưng chất lượng chưa cao hơn AD-NeRF không nhiều, model to, train lâu
- **Các method diffusion** (2023–2024): chất lượng cực đẹp nhưng inference 5–10 giây/frame → không real-time

→ Cần một method vừa real-time, vừa high-fidelity, vừa model nhỏ, vừa train nhanh

→ **ER-NeRF** ra đời.

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

a. Pipeline tổng quan của ER-NeRF

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

Bài báo đề xuất ba cải tiến quan trọng để giải quyết các vấn đề cốt lõi của talking-head NeRF:

(1) encoding không gian kém hiệu quả

→ **Tri-Plane Hash Representation**

(2) khớp nối âm thanh với hình học thiếu chính xác

→ **Region Attention Module**

(3) sự tách biệt giữa head NeRF và torso NeRF gây artifacts

→ **Torso – Adaptive Pose Encoding**

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(1) Tri-Plane Hash Representation

- **Vấn đề của phương pháp trước (RAD-NeRF/AD-NeRF):** Các mô hình NeRF cho talking head trước ER-NeRF thường dùng: 3D multi-resolution hash grid encoding (Instant-NGP). Nhưng trong setting "head-only NeRF", việc dùng full 3D hash encoding dẫn đến:
  - Hash collision cao:
    - Vì phần lớn không gian 3D của head volume không chứa thông tin.
    - NeRF chỉ cần mô hình hóa một vùng nhỏ (khuôn mặt), nhưng grid lại bao trùm toàn bộ bounding box → nhiều ô trống.
  - Dữ liệu hiệu quả thấp:
    - 3D hash grid chia nhỏ không gian thành voxel → số lượng cell tăng theo $O(n^3)$.
    - Phần lớn voxel không bao giờ được ray sample.
  - MLP phải học quan hệ vị trí 3D phức tạp: gây overfitting, dễ xuất hiện artifacts khi head xoay.

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(1) Tri-Plane Hash Representation

- **Ý tưởng:**
  - Nếu xem khuôn mặt là khối 3D, thì mô hình cũ lưu nó bằng cách lưu dữ liệu trong không gian 3 chiều → rất nặng, chậm, dễ lỗi.
  - ER-NeRF chia khối 3D thành 3 mặt phẳng 2D, tức là thay vì lưu trữ thông tin trong không gian 3D, ER-NeRF dùng 3 mặt phẳng 2D trực giao: XY, YZ, XZ

→ Mỗi điểm 3D p = (x, y, z) được ánh xạ sang ba điểm 2D trên ba mặt phẳng.

Hash encoder trên từng mặt phẳng:

$$H_{AB} : (a, b) \rightarrow f_{ab}^{AB} \tag{4}$$

Sau đó ghép 3 plane:

$$f_x = H_{XY}(x, y) \oplus H_{YZ}(y, z) \oplus H_{XZ}(x, z) \tag{5}$$

(*⊕ là phép nối vector.*)

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(1) Tri-Plane Hash Representation

- **Lợi ích:**
  - **Giảm hash collision**: collision ít hơn, embedding ổn định hơn
    - Hash 2D → $O(n^2)$ cell
    - Hash 3D → $O(n^3)$ cell
  - **Khối đầu người gần như là một manifold 2.5D**: Về bản chất, bề mặt đầu người có thể biểu diễn tốt bằng các lát cắt 2D theo ba hướng.
  - **Ray sampling hiệu quả hơn**: Mỗi truy vấn ray cần trích xuất đặc trưng từ 3 vector embedding nhỏ → nhanh hơn 3D grid nhiều.
  - **Tái tạo geometry tốt hơn khi xoay đầu:** 2D tri-plane lưu giữ thông tin hướng tốt hơn.

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(1) Tri-Plane Hash Representation

- Input: Tọa độ điểm 3D (x, y, z) trong không gian. (được lấy dọc theo các tia)
- Quy trình (Pipeline):
  - Bước 1 (Chiếu): Chiếu điểm 3D lên 3 mặt phẳng trực giao (XY, XZ, YZ).
  - Bước 2 (Truy vấn): Trích xuất đặc trưng từ bảng băm (Hash Table) tại các tọa độ 2D tương ứng.
  - Bước 3 (Tổng hợp): Gộp kết quả từ 3 mặt phẳng thành một Vector đặc trưng không gian.
- Output: Vector đặc trưng hình học tĩnh.
  - Vector này sẽ được chuyển đến Region Attention Module (RAM) để kết hợp với âm thanh.

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(1) Tri-Plane Hash Representation

(a) Static (b) 3D hash grid (c) Tri-hash (ours)

Hiệu quả của Tri-Hash Representation
trong giảm nhiễu và tái dựng bề mặt đầu

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(2) Region Attention Module

- **Vấn đề của phương pháp trước:** Các mô hình như RAD-NeRF dùng audio feature global (toàn cục), được trộn thẳng vào NeRF MLP. Hậu quả:
  - Lip-motion không chính xác: Vì môi chịu ảnh hưởng mạnh nhất từ audio nhưng các vùng khác bị noise từ audio “tràn sang”.
  - Artifact khi phát âm mạnh ("a", "o", "p")
  - Không đồng bộ giữa răng – môi – lưỡi

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(2) Region Attention Module

- **Ý tưởng RAM (Region-aware Attention Field):**
  - Khi nghe người khác nói, bạn chú ý nhất vào môi, sau đó tới má, còn trán thì chẳng liên quan gì.
  - Nhưng mô hình NeRF cũ lại dùng audio như nhau cho toàn bộ mặt → dẫn đến môi không khớp, méo mặt.
  - Thay vì trộn audio vào NeRF một cách "đồng nhất", ER-NeRF học một attention map trong không gian 3D:
    - Chia mặt thành nhiều vùng nhỏ
    - Học xem audio ảnh hưởng mạnh nhất lên vùng nào:
      - môi → ảnh hưởng mạnh
      - má → ảnh hưởng ít
      - mắt → gần như không ảnh hưởng

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(2) Region Attention Module

- **Công thức của Region Attention**
  - External attention:

$$A = \mathrm{ReLU}(FM_k^T), \qquad V_{\mathrm{out}} = AM_v \tag{7}$$

  - Channel-wise attention để reweight audio:

$$q_{\mathrm{out}} = v \odot q \tag{8}$$

    - Speech audio:

$$v_{a,x} = \mathrm{MLP}_a(H_3(x))$$

$$a_{r,x} = v_{a,x} \odot a \tag{9}$$

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(2) Region Attention Module

- **Lợi ích:**
  - Môi học chính xác vì attention cao tại vùng môi
  - Mắt/mũi/tiền cảnh ít bị nhiễu
  - Chuyển động khớp audio frame-by-frame
  - Xử lý tốt pose khó vì attention được conditioned theo vị trí

Rendering Result | 2D Audio Attention Map | 3D Audio Attention Map | 2D Blink Attention Map | 3D Blink Attention Map

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(3) Torso – Adaptive Pose Encoding

- **Vấn đề của phương pháp trước:** Vấn đề của các mô hình head-NeRF:
  - Head NeRF và Torso NeRF được train riêng
  - Thiếu ràng buộc về pose → dẫn đến lỗi kinh điển:
  - cổ bị lệch
  - đầu trôi khỏi vai
  - khi xoay đầu, thân không xoay tương ứng

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(3) Torso – Adaptive Pose Encoding

- Quy trình (Pipeline):
  - Bước 1 (Head Pose): Lấy thông tin tư thế đầu (Xoay R, Tịnh tiến t) từ quá trình tracking.
  - Bước 2 (Adaptive Encoding): Ánh xạ tư thế đầu vào các điểm chốt (Trainable Keypoints) để tính toán độ biến dạng của thân.
  - Bước 3 (Synthesis): Dùng mạng 2D tạo ra hình ảnh phần thân khớp với tư thế đầu.
  - Kết quả: Ghép đầu (3D) và thân (2D) lại để ra video hoàn chỉnh.

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(3) Torso – Adaptive Pose Encoding

- **Mô hình Torso-NeRF (kỹ thuật):** Torso-NeRF có cấu trúc gần giống head-NeRF nhưng đơn giản hơn:
  - Giảm số layer trong MLP
  - Dùng canonical space để giữ ổn định
  - Ánh xạ điểm trong torso volume qua pose transform rồi mới đánh giá σ, c

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(3) Torso – Adaptive Pose Encoding

- **Công thức Adaptive Pose Encoding:**
  - Khởi tạo N=3 điểm trong không gian 3D:

$$X_{\text{keys}} = (x_{\text{keys}}, y_{\text{keys}}, z_{\text{keys}}, 1)^T \in \mathbb{R}^{4\times 3} \qquad (13)$$

  - Ứng dụng head pose P = (R, t):

$$\hat{X}_{\text{keys}} = P^{-1} X_{\text{keys}} \qquad (15)$$

  - Chiếu xuống mặt phẳng Z = 1 để lấy tọa độ 2D:

$$\bar{X}_{\text{keys}}(i,j) = \gamma \cdot \frac{\hat{X}_{\text{keys}}(i,j)}{\hat{z}_{\text{keys}}(j)} \qquad (16)$$

  - Sử dụng $\bar{X}_{\text{keys}}$ để điều khiển torso-NeRF:

$$F_T : (x_{\text{pixel}}, \bar{X}_{\text{keys}}; H_t) \rightarrow (c_t, \alpha) \qquad (17)$$

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(3) Torso – Adaptive Pose Encoding

- **Công thức Adaptive Pose Encoding:**
  - Ý nghĩa:
    - Thay vì đưa cả ma trận pose vào mạng (rất nhiễu → dễ sai)
    - ER-NeRF chỉ dùng 3 keypoints, transform theo pose → tạo 3 điểm neo cho phần thân.
    - Khi đầu nghiêng sang trái, cả ba keypoints sẽ nghiêng tương ứng → torso-NeRF biết phải xoay vai và cổ theo.

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

(3) Torso – Adaptive Pose Encoding

- **Lợi ích:**
  - Tránh artifact tách đầu – cổ
  - Torso cử động phù hợp hướng đầu quay
  - Tăng tính tự nhiên trong video
  - Chất lượng rendering ổn định khi pose lớn (ngước lên/ nghiêng xuống)

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

b. Thuật toán chính

- **(3) Eyes blinking**
- Mục tiêu:
  - Tách biệt chuyển động mắt khỏi chuyển động môi.
  - Cho phép điều khiển việc chớp mắt một cách chủ động (Explicit control).
- Cơ chế hoạt động:
  - Sử dụng một biến vô hướng (scalar) e để đại diện cho trạng thái mắt (mở/nhắm).
  - Sử dụng mạng MLP riêng để sinh ra vector chú ý vùng mắt ($v_{e,x}$).
  - Công thức (Khác biệt so với Audio):
  - Sử dụng hàm Sigmoid (thay vì ReLU) để tạo mặt nạ vùng mắt:

$$\mathbf{v}_{e,\mathbf{x}} = \mathrm{MLP}_e(\mathcal{H}^3(\mathbf{x})),$$
$$\mathbf{e}_{r,\mathbf{x}} = \mathbf{e} \cdot \mathrm{Sigmoid}(\mathbf{v}_{e,\mathbf{x}}).$$

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

2. Thuật toán

c.Coarse-to-Fine Optimization

- Vấn đề của NeRF truyền thống:
  - Chỉ sử dụng hàm mất mát MSE Loss (Mean Squared Error).
  - Kết quả: Ảnh tái tạo tốt về màu sắc tổng thể nhưng thường bị mờ (blurry), thiếu chi tiết sắc nét (tóc, răng).
- Giải pháp: Huấn luyện 2 giai đoạn (Two-staged Process):
- Giai đoạn 1 (Coarse - Thô): Dùng MSE Loss trên toàn bộ ảnh
  - Mục đích: Học cấu trúc khuôn mặt và màu sắc cơ bản.

$$\mathcal{L}_{coarse} = \sum_{\mathrm{i} \in \mathcal{I}} \left\| C(\mathrm{i}) - \hat{C}(\mathrm{i}) \right\|_2^2 .$$

- Giai đoạn 2 (Fine - Tinh):
  - Kết hợp MSE + LPIPS Loss (Perceptual Loss).
  - Lấy mẫu theo các mảng ảnh (patches) ngẫu nhiên thay vì từng điểm ảnh rời rạc.
  - Mục đích: Tăng cường độ sắc nét và chi tiết tần số cao.

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

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

3. Phân tích kết quả

a. Giải thích kết quả

1. Head Reconstruction – tái tạo khuôn mặt từ video gốc:

Methods	PSNR ↑	LPIPS ↓	FID ↓	LMD ↓	AUE ↓	Sync ↑	Time	FPS	Size (MB)
Ground Truth	N/A	0	0	0	0	7.584	-	-	-
Wav2Lip [32]	-	-	31.08	5.124	3.861	**8.576**	-	19	>400
PC-AVS [57]	18.25	0.2440	101.97	4.816	3.142	8.397	-	32	>500
AD-NeRF [23]	30.75	0.1034	18.60	3.345	2.201	5.205	18h	0.13	5.21
RAD-NeRF [40]	33.13	0.0519	12.05	2.812	2.102	5.052	5h	32	11.8
RAD-NeRF†	**33.26**	0.0486	12.20	2.802	1.750	5.197	-	-	-
ER-NeRF (Ours)	33.10	**0.0291**	**10.42**	**2.740**	**1.629**	5.708	**2h**	**34**	**2.51**

† using AU45 and overall LPIPS finetune.

- PSNR = 33.10 → Mức khôi phục chi tiết ảnh gần với ảnh thật, tương đương hoặc tốt hơn các phương pháp NeRF.
- LPIPS = 0.0291 (thấp nhất) → Mô hình tái tạo được các chi tiết nhỏ như răng, môi, da tốt hơn các baseline.
- FID = 10.42 (tốt nhất) → Ảnh sinh ra "thật" hơn đối với phân phối ảnh thật.
- LMD = 2.740 (tốt nhất) → Mức sai lệch landmark môi rất thấp, khớp khẩu hình tốt hơn các phương pháp còn lại.
- AUE = 1.629 (tốt nhất) → Chuyển động khuôn mặt theo đặc trưng AU (Action Unit) chính xác hơn.
- Sync = 5.708 (cao nhất) → Độ đồng bộ audio–lip vượt trội trong nhóm NeRF-based.
- Inference: 34 FPS + thời gian train 2 giờ + size 2.51 MB→ Đạt real-time, nhanh hơn RAD-NeRF (5 giờ + 11.8 MB).

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

3. Phân tích kết quả

a. Giải thích kết quả

1. Head Reconstruction – tái tạo khuôn mặt từ video gốc:

Nguyên nhân đặt kết quả tốt:

- Tri-Plane Hash Representation:
  - Giảm xung đột hash → MLP không bị nhiễu → học chi tiết tốt hơn.
  - Prune vùng rỗng (empty space) → tăng hiệu quả, giảm thời gian học.
- Region Attention Module:
  - Học rõ "vùng nào" chịu ảnh hưởng của âm thanh → mô hình tạo được cử động môi tự nhiên.
  - Các vùng tĩnh (ví dụ tóc, nền) được loại bỏ nhiễu.
- Adaptive Pose Encoding: Tách head–torso tốt hơn, tránh lỗi "gãy cổ" mà AD-NeRF thường gặp.

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

3. Phân tích kết quả

a. Giải thích kết quả

2. Lip Synchronization – kiểm tra đồng bộ môi trên audio chưa thấy trong training:

	Testset A		Testset B	
Methods	LMD ↓	Sync ↑	LMD ↓	Sync ↑
Ground Truth	0	6.701	0	7.309
Wav2Lip [32]	6.221	**8.378**	**7.393**	**8.966**
PC-AVS [57]	7.112	8.087	7.722	8.565
SynObama [39]	6.540	6.802	-	-
NVP [41]	-	-	7.954	4.313
LSP [29]	**5.905**	4.287	8.122	5.843
AD-NeRF [23]	6.192	5.195	8.006	4.316
SSP-NeRF [28]	6.332	5.422	-	-
RAD-NeRF [40]	6.357	6.186	8.332	6.680
RAD-NeRF†	6.339	6.119	8.355	6.392
Ours	6.254	6.242	8.150	6.830

† using AU45 and overall LPIPS finetune.

- ER-NeRF đạt Sync cao nhất trong nhóm NeRF-based trên cả Testset A và B.
- LMD thấp → môi khớp tốt với âm thanh ngay cả audio ngoài tập huấn luyện.
- Chứng tỏ mô hình generalize tốt, không bị overfit vào audio gốc.

2. NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH

3. Phân tích kết quả

b. So sánh với các phương pháp khác trong bài báo

Phương pháp	Đặc điểm chính	Nhược điểm	So sánh với ER-NeRF
AD-NeRF (2021)	- Dùng MLP lớn - Tái tạo 3D head theo audio	- Rất chậm: 18 giờ train- FPS chỉ 0.13, không real-time - Lỗi “gãy cổ”, lệch head–torso	- ER-NeRF nhanh hơn ~140× - Đồng bộ môi tốt hơn- Ảnh sắc nét hơn
RAD-NeRF (2022)	- Dùng 3D hash grid - Hỗ trợ tái tạo đầu tốt hơn AD-NeRF	- Hash collision cao → ảnh mờ, méo cấu trúc - Train chậm: 5 giờ- Model size lớn: 11.8 MB	- ER-NeRF dùng Tri-Plane Hash → collision ↓ 5×- Train nhanh hơn: 2 giờ - Model size nhỏ hơn: 2.51 MB- FPS tương đương nhưng ổn định hơn
One-shot models (Wav2Lip, PC-AVS)	- Không cần train riêng cho từng người - Lip-sync tốt	- Hình ảnh kém thật - Mất đặc trưng gương mặt - Chỉ xuất video 2D (không 3D head modeling)	- ER-NeRF giữ được đặc trưng khuôn mặt 3D - Sync tốt (nhưng thua Wav2Lip vì one-shot)
ER-NeRF (proposed)	- Tri-Plane Hash + Region Attention- Pose Encoding mới - Đồng bộ môi tốt- Chạy real-time (34 FPS)	—	—

3. DEMO VÀ PHÂN TÍCH THỰC NGHIỆM

Original video

Inference video

3. DEMO VÀ PHÂN TÍCH THỰC NGHIỆM

4. ĐÁNH GIÁ, PHÂN TÍCH VÀ NHẬN XÉT

1. Điểm mới và ưu điểm của bài toán

- Region Attention chưa giải quyết triệt để sai môi lắm nhưng tốt hơn so với các model trước đó
- Tri-plane nhanh & chính xác
- Torso-NeRF tạo chuyển động liền mạch
- Mô hình rất nhẹ
- Chạy real-time

2. Các hạn chế của bài toàn

- Phải train cho từng người → không “zero shot”
- Chưa tạo được biểu cảm phức tạp (cười, nhăn mặt)
- Cần GPU để train
- Chưa tổng quát tốt nếu input kém chất lượng