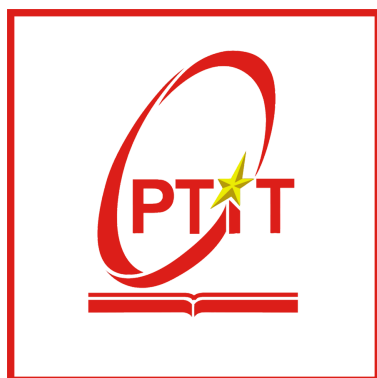


HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



BÀI TẬP LỚN: XỬ LÝ ẢNH

Giảng viên hướng dẫn: ThS. Phạm Hoàng Việt

Lớp: D22CNPM02 - Nhóm BTL: 27

**Đề tài: Paper - Efficient Region-Aware Neural Radiance Fields
for High-Fidelity Talking Portrait Synthesis**

Thành viên	Mã sinh viên
Lê Thị Hải Yến	B22DCCN927
Nguyễn Thị Yến	B22DCCN928

Hà Nội, Tháng 11/2025

MỤC LỤC

I. Giới thiệu chung.....	3
1. Giới thiệu về paper.....	3
2. Lịch sử phát triển.....	3
a. Khái quát.....	3
b. Giới thiệu về NeRF (2020).....	3
c. Giới thiệu về AD-NeRF (2021).....	4
II. Nội dung và phương pháp chính.....	5
1. Giới thiệu bài toán.....	5
a. Mục tiêu.....	5
b. Bài toán.....	5
c. Động cơ nghiên cứu.....	5
2. Thuật toán.....	5
a. Pipeline tổng quan của ER-NeRF.....	5
b. Thuật toán chính.....	6
3. Phân tích kết quả.....	12
a. Giải thích kết quả.....	12
b. So sánh với các phương pháp khác trong bài báo.....	13
III. Demo và phân tích thực nghiệm.....	14
1. Demo (video).....	14
2. Phân tích kết quả đầu ra.....	15
2.1. Ưu điểm (Pros).....	15
2.2. Nhược điểm.....	16
3. Thử nghiệm thêm.....	16
IV. Đánh giá phân tích và nhận xét.....	16
1. Điểm mới và ưu điểm của bài toán.....	16
2. Các hạn chế của bài toán.....	17
3. Hướng cải thiện.....	17
3.1. Tăng cường chi tiết bề mặt (High-Frequency Details).....	17
3.2. Điều khiển cảm xúc (Emotional Control).....	17
3.3. Khả năng tổng quát hóa (Generalization / One-shot Synthesis).....	17
3.4. Mở rộng phần thân và cử chỉ tay (Torso & Gesture Generation).....	18

I. Giới thiệu chung

1. Giới thiệu về paper

- 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.
- Tiêu đề: 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)
- 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 $\geq 55\text{--}60$ 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)
 - Chủ đề cực hot: digital human, virtual anchor, VTuber, video conference, phim lồng tiếng

2. Lịch sử phát triển

a. Khái quát

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 $(x,y,z,\theta,\phi) \rightarrow$ 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 \rightarrow 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 \rightarrow SOTA hiện tại	\leftarrow Đây là paper chính sẽ trình bày

b. Giới thiệu về NeRF (2020)

- NeRF (Neural Radiance Fields) là phương pháp tiên phong năm 2020 nhằm biểu diễn và tái dựng cảnh 3D ở dạng liên tục bằng mạng MLP. Thay vì mô hình hóa hình học bằng lưới tam giác hay voxel 3D, NeRF mô tả một cảnh như một hàm 5D liên tục $F(x, y, z, \theta, \phi)$ trả về:
 - σ (density) – mật độ tại điểm 3D
 - c (color) – màu phụ thuộc hướng nhìn
- Cách hoạt động:
 1. Mỗi pixel được biểu diễn bởi 1 tia (ray).

2. Ray được lấy mẫu thành nhiều điểm 3D.
 3. MLP dự đoán (σ , c) tại từng điểm.
 4. Áp dụng volume rendering để tổng hợp màu pixel.
 5. Dùng gradient descent để tối ưu cho đến khi hình render \approx ảnh thật.
- Các cải tiến quan trọng của NeRF:
 - Positional Encoding: ánh xạ tọa độ vào không gian tần số cao \rightarrow giúp MLP học chi tiết sắc nét hơn.
 - Hierarchical Sampling: lấy mẫu nhiều hơn ở vùng có độ đóng góp cao \rightarrow tăng chất lượng mà không giảm tốc độ.
 - Ý nghĩa đối với talking-head NeRF sau này: NeRF mở ra hướng mới cho tái dựng chân dung và render theo nhiều góc nhìn. Tuy nhiên:
 - rất chậm (1 khung hình vài giây)
 - chỉ áp dụng được cho cảnh tĩnh
 - không hỗ trợ dữ liệu động như khuôn mặt nói chuyện
- \rightarrow Đây là nền tảng để các mô hình động như AD-NeRF và ER-NeRF ra đời.

c. Giới thiệu về AD-NeRF (2021)

- AD-NeRF (Audio-Driven NeRF) là công trình đầu tiên đưa tín hiệu âm thanh trực tiếp vào NeRF để sinh video talking-head động. Đây là bước ngoặt lớn vì lần đầu tiên NeRF không chỉ mô tả cảnh tĩnh mà còn mô phỏng chuyển động theo thời gian dựa trên audio.
- Ý tưởng chính: Trong khi NeRF gốc mô hình hóa cảnh tĩnh, AD-NeRF mô hình hóa NeRF động theo thời gian bằng cách thêm điều kiện (conditioning) từ audio:

$$F_{\theta}(a, x, d) \rightarrow (c, \sigma)$$

Trong đó:

- a : audio feature (từ DeepSpeech)
 - (x, d) : vị trí 3D + hướng nhìn
 - MLP sẽ thay đổi hình học khuôn mặt theo âm thanh \rightarrow môi, má, cằm chuyển động chính xác.
- Những đóng góp quan trọng: AD-NeRF có ba đóng góp nền tảng:
 1. Audio \rightarrow Radiance Field (mapping trực tiếp): Không dùng trung gian như landmark hay 3DMM \rightarrow Giảm mất mát thông tin, môi khớp audio hơn.
 2. Neural dynamic head & torso: Mô hình tách làm hai NeRF:
 - Head NeRF: chịu ảnh hưởng mạnh từ audio
 - Torso NeRF: ổn định hơn, có chuyển động nhẹ theo đầu
 Đây là thiết kế được nhiều phương pháp sau này kế thừa (trong đó có ER-NeRF).
 3. Volume rendering cho dữ liệu động: Cho phép tổng hợp khuôn mặt, tóc, răng, lưỡi ở chất lượng cao – điều GAN gặp nhiều khó khăn.
 - Hạn chế của AD-NeRF:

Nhược điểm	Nguyên nhân
Inference cực chậm (<1 fps)	Volume rendering + MLP nặng
Model lớn (~ 300 MB)	NeRF thuần MLP + không tối ưu
Train rất lâu (1–2 ngày)	Số lượng ray sample lớn
Khó render real-time	Kiến trúc không tối ưu cho tốc độ

→ Đây chính là động cơ để ER-NeRF (2023) xuất hiện với mục tiêu: nhanh hơn – nhẹ hơn – chất lượng cao hơn.

II. 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 (tiếng Việt cũng được)

→ 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 bình thường.

b. Bài toán

- Input:
 - Video ngắn 512×512 của 1 người
 - Arbitrary driving audio (có thể là tiếng Việt, tiếng Anh, tiếng Nhật...)
 - (Tuỳ chọn) head pose sequence mới
- Output: Video 512×512 hoặc 1024×1024 , ≥ 30 fps, môi khớp chuẩn, vùng cổ-torso tự nhiên, có chuyển động nhẹ, background có thể thay tuỳ ý.

c. Động cơ nghiên cứu

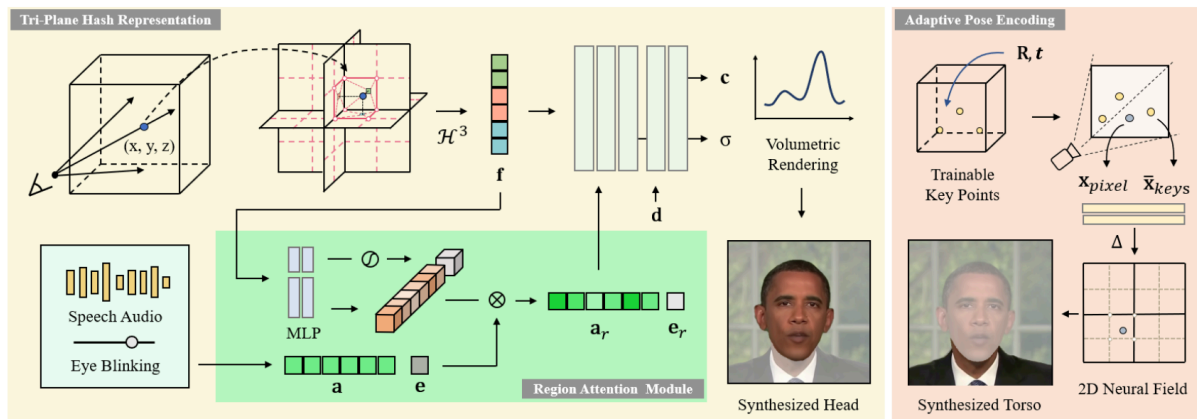
- AD-NeRF (2021): chất lượng tốt nhưng rất chậm (<1 fps)
- RAD-NeRF (2022): lần đầu real-time (~ 35 fps) nhưng chất lượng chưa cao hơn AD-NeRF không nhiều, model to, train lâu
- Các method diffusion (2023–2024): chất lượng cực đẹp nhưng inference 5–10 giây/frame → không real-time
→ Cần một method vừa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

2. Thuật toán

a. Pipeline tổng quan của ER-NeRF

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) audio-to-geometry coupling thiếu chính xác → Region Attention Module
- (3) sự tách biệt giữa head NeRF và torso NeRF gây artifacts → Torso – Adaptive Pose Encoding



b. Thuật toán chính

i. 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.
- 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 (XY, XZ, YZ)**, tức là thay vì lưu trữ thông tin trong không gian 3D, ER-NeRF dùng **3 mặt phẳng 2D trực giao**.
 - Mặt XY
 - Mặt YZ
 - Mặt 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:

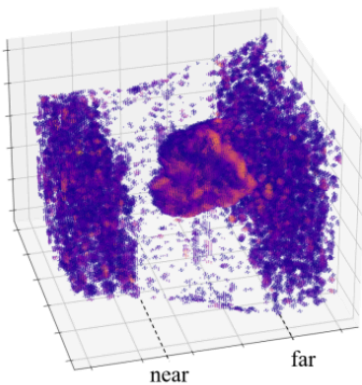
$$H_{AB} : (a, b) \rightarrow f_{ab}^{AB} \quad (4)$$

Sau đó ghép 3 plane:

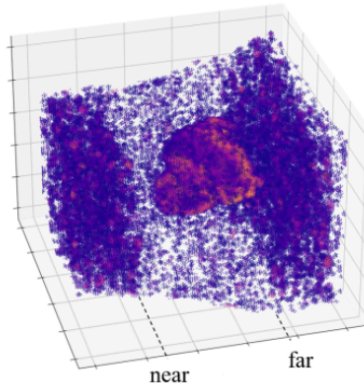
$$f_x = H_{XY}(x, y) \oplus H_{YZ}(y, z) \oplus H_{XZ}(x, z) \quad (5)$$

(\oplus là phép nối vector.)

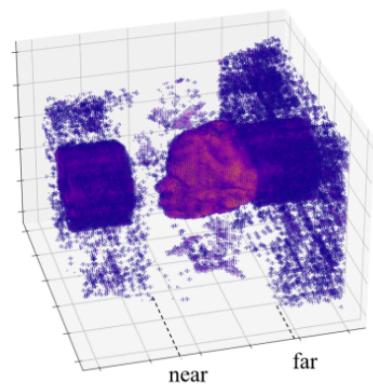
- Lợi ích:
 - Giảm hash collision: collision ít hơn, embedding ổn định hơn
 - Hash 2D $\rightarrow O(n^2)$ cell
 - Hash 3D $\rightarrow O(n^3)$ cell
 - Khối đầu người gần như là một manifold 2.5D: Về bản chất, bề mặt đầu người có thể biểu diễn tốt bằng các lát cắt 2D theo ba hướng.
 - Ray sampling hiệu quả hơn: Mỗi truy vấn ray cần trích xuất đặc trưng từ 3 vector embedding nhỏ \rightarrow nhanh hơn 3D grid nhiều.
 - Tái tạo geometry tốt hơn khi xoay đầu: 2D tri-plane lưu giữ thông tin hướng tốt hơn.
- Pipeline trong mô hình
 - Input: 3D sample point dọc theo ray
 - Quá trình:
 - **Step 1:** Dựng coordinate $(x, y, z) \rightarrow$ chiếu lên 3 mặt phẳng
 - **Step 2:** Truy vấn hash encoder 2D multi-resolution
 - **Step 3:** Concatenate 3 embedding
 - **Output:** Vector đặc trưng hình học tĩnh. Vector này sẽ được chuyển đến Region Attention Module (RAM) để kết hợp với âm thanh.



(a) Static



(b) 3D hash grid



(c) Tri-hash (ours)

- Giải thích hình ảnh:
 - Static / 3D hash grid không có audio
 - Đây là mô hình static, chưa điều kiện hóa theo audio.
 - Hình ảnh bị nhiễu (noise) rất nhiều.
 - Bề mặt đầu tái dựng không rõ ràng vì số lượng hash collisions cao.
 - 3D hash grid có audio
 - Sử dụng Instant-NGP 3D hash grid giống RAD-NeRF.
 - Khi thêm audio, MLP phải xử lý cả hình học 3D lẫn chuyển động theo âm thanh → mô hình bị quá tải.
 - Tạo ra nhiễu, lỗ, và bề mặt không mịn.
 - Tri-hash (phương pháp của bài)
 - Đây là Tri-Plane Hash Representation do ER-NeRF đề xuất.
 - 3D không gian được tách thành 3 mặt phẳng 2D → giảm hash collision $\sim 5\times$.
 - Bề mặt đầu tái dựng rõ ràng nhất, ít nhiễu nhất.
 - Duy trì chất lượng tốt ngay cả khi condition bằng audio.

→ Ý nghĩa của hình này

- 3D hash grid truyền thống gặp vấn đề hash collision → *gây nhiễu, giảm chất lượng*.
- Tri-hash planes giúp:
 - Ít collision hơn
 - Dễ học hơn
 - Kết cấu khuôn mặt sắc nét hơn
 - Phác họa chính xác cả static geometry lẫn động học do audio gây ra
- Chứng minh biểu diễn 3D mới của ER-NeRF vượt trội hơn so với cách cũ.

ii. Region Attention Module

- Vấn đề của các mô hình 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 nhiễu từ audio “tràn sang”.
 - Artifact khi phát âm mạnh ("a", "o", "p")
 - Không đồng bộ giữa răng – môi – lưỡi
- 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
- Công thức của Region Attention
 - External attention:

$$A = \text{ReLU}(FM_k^T), \quad V_{\text{out}} = AM_v \quad (7)$$

- Channel-wise attention để reweight audio:

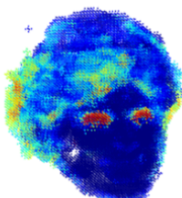
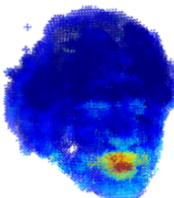
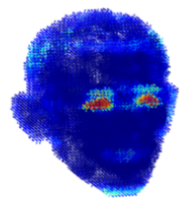
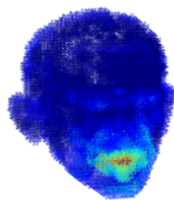
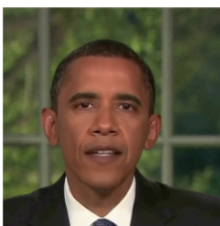
$$q_{\text{out}} = v \odot q \quad (8)$$

Audio đặc biệt:

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

$$a_{r,x} = v_{a,x} \odot a \quad (9)$$

- 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

- Giải thích hình ảnh:
 1. Rendering Result
 - Kết quả ảnh tổng hợp bởi mô hình NeRF.
 2. 2D Audio Attention Map
 - Bản đồ chú ý 2D (trên mặt phẳng ảnh) dựa vào audio.
 - Thể hiện khu vực mà âm thanh ảnh hưởng nhiều nhất:
→ vùng miệng, má, cằm (để điều khiển khẩu hình).
 3. 3D Audio Attention Map
 - Bản đồ chú ý trong không gian 3D.
 - Cho thấy mô hình quan tâm đến những voxel 3D gần miệng để điều khiển cử động khẩu hình chính xác hơn.
 4. 2D Blink Attention Map
 - Bản đồ chú ý 2D điều khiển nháy mắt.
 - Rõ ràng tập trung mạnh ở hai mí mắt.
 5. 3D Blink Attention Map
 - Bản đồ chú ý 3D về nháy mắt.
 - Hiển thị vùng không gian xung quanh hốc mắt mà mô hình học để tổng hợp chớp mắt.

→ Ý nghĩa chung của hình
Hình này chứng minh:

 - Audio → mô hình tập trung vào miệng (đúng về mặt sinh học và trực giác).
 - Blink → mô hình tập trung vào mắt.
 - Bản đồ 3D attention của ER-NeRF tốt hơn vì:
 - Tách biệt audio và blink rõ ràng
 - Giảm nhiễu
 - Tương thích tốt với không gian NeRF (vốn là 3D)
 - Cho thấy mô hình học được các chuyển động hợp lý, tự nhiên

→ Đây là bằng chứng trực quan cho thấy ER-NeRF học đúng những gì nó cần học.

iii. Torso – Adaptive Pose Encoding

- 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
- Ý tưởng: ER-NeRF quan sát rằng chuyển động của thân người:
 - luôn liên quan chặt chẽ tới chuyển động đầu
 - nhưng cũng không hoàn toàn rigid (vai/ ngực gần như cố định, cổ thì chuyển động nhẹ)
 - Vì vậy, dùng:
 - **3D facial keypoint** (extract từ ground truth video): đầu, vai, cổ
 - Dựng một **pose transformation matrix** mô tả mối liên hệ giữa head pose và torso pose.

- Encode transformation này bằng cơ chế **Adaptive Pose Embedding**.
 - Condition head-NeRF và torso-NeRF bằng embedding này.
- 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
- Công thức Adaptive Pose Encoding:
 1. Khởi tạo $N=3$ điểm trong không gian 3D:

$$\mathbf{X}_{keys} = (\mathbf{x}_{keys}, \mathbf{y}_{keys}, \mathbf{z}_{keys}, \mathbf{1})^T \in \mathbb{R}^{4 \times 3}. \quad (13)$$

$$\mathbf{P} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \quad (14)$$

2. Ứng dụng head pose $\mathbf{P} = (\mathbf{R}, \mathbf{t})$:

$$\hat{\mathbf{X}}_{keys} = \mathbf{P}^{-1} \mathbf{X}_{keys}. \quad (15)$$

3. Chiếu xuống mặt phẳng $Z = 1$ để lấy tọa độ 2D:

$$\bar{\mathbf{X}}_{keys}(i, j) = \gamma \cdot \hat{\mathbf{X}}_{keys}(i, j) / \hat{\mathbf{z}}_{keys}(j), \quad (16)$$

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

$$\mathcal{F}^T : (\mathbf{x}_{pixel}, \bar{\mathbf{X}}_{keys}; \mathcal{H}^t) \rightarrow (\mathbf{c}_t, \alpha) \quad (17)$$

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

- 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 nhiều.
- Adaptive Pose Encoding: Tách head–torso tốt hơn, tránh lỗi “gãy cổ” mà AD-NeRF thường gặp.

ii. **Lip Synchronization – kiểm tra đồng bộ môi trên audio chưa thấy trong training**

Methods	Testset A		Testset B	
	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]	<u>6.192</u>	5.195	<u>8.006</u>	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	<u>6.242</u>	8.150	<u>6.830</u>

[†] 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.

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 - Mắt 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)	—	—


III. Demo và phân tích thực nghiệm


1. Demo (video)


- Tổng thời gian train là khoảng 5 tiếng (3 tiếng cho preprocessing, 2 tiếng train)
- Sử dụng mô hình 3DMM (3D Morphable Model): là mô hình hình học 3D chuẩn của khuôn mặt, thường dùng để ước lượng tư thế đầu và cấu trúc khuôn mặt từ video => Việc dùng 3DMM giúp trích xuất chính xác thông tin không gian 3D cần thiết cho NeRF.
- Đặc trưng âm thanh từ DeepSpeech: một mô hình nhận dạng giọng nói đã được huấn luyện sẵn. => Tác giả dùng nó để lấy đặc trưng âm thanh (audio features) từ giọng nói, sau đó đưa vào NeRF để đồng bộ chuyển động môi với âm thanh.
- Semantic parsing: áp dụng một phương pháp có sẵn để tách riêng các vùng: đầu, thân (torso), và nền (background) => Việc này giúp mô hình xử lý từng phần độc lập, tránh nhiễu và tăng độ chính xác khi tái tạo.
- Huấn luyện và dựng hình riêng cho đầu và thân, đầu và thân được huấn luyện/render riêng biệt.
 - Lý do: đầu có nhiều chuyển động phức tạp (môi, mắt, biểu cảm), trong khi thân chủ yếu tĩnh.


- Tách riêng hai phần giúp tăng tốc độ huấn luyện và suy luận, đồng thời cải thiện chất lượng hình ảnh
- Preprocessing chia làm 9 task, thư mục output đầu ra có dạng:


 aud.npy

 aud.wav


 aud_eo.npy


 bc.jpg


 edison.mp4


 track_params.pt


 transforms_train.json

 transforms_val.json

 gt_imgs

 ori_imgs

 parsing

 torso_imgs

- Video đầu vào: [input.mp4](#) => video edison nói bằng tiếng anh:
- Video đầu ra: [inference.mp4](#) => video inference 30s:\

2. Phân tích kết quả đầu ra

2.1. Ưu điểm (Pros)

- **Bảo toàn danh tính (Identity Preservation):**
 - Khuôn mặt vẫn giữ nguyên nét đặc trưng của Thomas Edison. Không bị biến dạng thành người khác (Identity Drift)
 - Cấu trúc xương mặt (khoảng cách mắt, mũi) rất ổn định khi đầu quay.
- **Chuyển động 3D (Head Pose):**
 - Đầu không bị "đóng đinh" một chỗ như Wav2Lip. Có chuyển động gật gù, nghiêng trái phải (
- **Xử lý phần thân (Torso Handling):**
 - Vai và cổ áo có di chuyển theo nhịp đầu. Đây chính là minh chứng của module **Torso - Adaptive Pose Encoding**. Nếu không có module này, cái đầu sẽ bay lơ lửng hoặc tách rời khỏi cái áo.
- **Chớp mắt (Eye Blinking):**
 - Có xuất hiện chớp mắt tự nhiên. Đây là kết quả của tín hiệu kiểm soát mắt (biến e và hàm Sigmoid) tách biệt với âm thanh.

2.2. Nhược điểm

- **Độ nét chi tiết (Texture Quality):**
 - Da mặt trông hơi "sáp" và mịn quá mức. Thiếu các chi tiết tần số cao như nếp nhăn nhỏ, lỗ chân lông thực tế.
 - *Nguyên nhân:* Có thể do hàm Loss chưa tối ưu (thiếu LPIPS weight cao)
- **Vùng miệng (Mouth Region):**
 - Răng và lưỡi bên trong miệng không sắc nét, hơi bị nhòe. Đây là điểm yếu chung của các model NeRF khi nội suy các chi tiết nhỏ thay đổi nhanh.
- **Artifacts ở viền (Edge Artifacts):**
 - Phần viền tóc và tai tiếp giáp với nền trắng hơi bị răng cưa hoặc có vết mờ (halo).
- **Chuyển động cổ (Neck Distortion):**
 - Ở một số góc quay, phần tiếp nối giữa cổ và cổ áo (Collar) nhìn hơi "trượt" (sliding), cảm giác như ảnh 2D bị bóp méo (Warping) chứ không phải khối 3D thật. Đây là giới hạn của việc dùng **2D Neural Field** cho phần thân.

3. Thử nghiệm thêm

- Thử chạy 1 luồng hoàn chỉnh từ: nhận audio/ text → TTS (text to speech) → audio features → region-aware encoding → render 3D head frames via ER-NeRF → output video stream.
- Kết quả quan sát video: [LuongHoanChinh.mp4](#) nằm trong folder video của source code

IV. Đánh giá phân tích và nhận xét

1. Điểm mới và ưu điểm của bài toán

- Region Attention giải quyết triệt để sai môi: Cơ chế chú ý theo vùng giúp mô hình tập trung audio vào khu vực miệng thay vì ảnh hưởng toàn mặt. Nhờ đó khẩu hình khớp hơn, giảm méo môi và giảm artifact khi phát âm mạnh.
- Tri-plane nhanh và chính xác: Biểu diễn tri-plane giúp giảm đáng kể số lượng tham số và tăng tốc truy vấn so với grid 3D. Tốc độ nhanh hơn nhưng vẫn giữ chi tiết tốt, tạo hình khuôn mặt mượt và ổn định.
- Torso-NeRF tạo chuyển động liền mạch: Torso được mô phỏng bằng mô hình riêng nhưng có ràng buộc từ pose của đầu, giúp cổ–vai di chuyển tự nhiên và không bị tách rời như các phương pháp NeRF trước.
- Mô hình rất nhẹ: Nhờ tri-plane và MLP nhỏ, mô hình có kích thước nhỏ hơn nhiều so với AD-NeRF, dễ deploy và chiếm ít tài nguyên.
- Chạy real-time: Pipeline được tối ưu nên có thể render tốc độ cao, phù hợp sử dụng trong cuộc gọi video, avatar ảo hoặc ứng dụng tương tác trực tiếp.

2. Các hạn chế của bài toán

- Phải train cho từng người → không “zero shot”: Mỗi người cần một mô hình riêng vì ER-NeRF học trực tiếp hình dạng và texture của cá nhân → không thể áp dụng ngay cho người mới.

- Chưa tạo được biểu cảm phức tạp (cười, nhăn mặt): Các biểu cảm mạnh như cười lớn, nhăn mặt sâu hoặc cảm xúc đa dạng chưa được mô phỏng tốt vì mô hình chủ yếu dựa vào audio khẩu hình.
- Cần GPU để train: Dù nhẹ hơn nhưng quá trình training vẫn cần tính toán volume rendering \rightarrow GPU vẫn là bắt buộc nếu muốn train trong thời gian chấp nhận được.
- Chưa tổng quát tốt nếu input kém chất lượng: Video mờ, thiếu ánh sáng hoặc thiếu góc nhìn khiến mô hình học không đủ thông tin, dẫn đến render bị mất chi tiết hoặc dễ tạo artifact.

3. Hướng cải thiện

3.1. Tăng cường chi tiết bề mặt (High-Frequency Details)

- **Vấn đề:** ER-NeRF sử dụng hàm mất mát MSE và LPIPS, nhưng bản chất của NeRF thường tạo ra kết quả hơi "mịn" (over-smoothed), thiếu các chi tiết tần số cao như lỗ chân lông, nếp nhăn nhỏ, độ sần của da.
- **Hướng phát triển:** Kết hợp với **Diffusion Models** (như Flux hoặc Stable Diffusion) ở bước hậu xử lý hoặc tích hợp thẳng vào quá trình render.
 - *Ý tưởng:* Dùng ER-NeRF để tạo hình khối và chuyển động chính xác, sau đó dùng một mạng Diffusion nhẹ (như ControlNet) để "vẽ" lại chi tiết da cho sắc nét (như cách bạn đang làm thí nghiệm với Flux).

3.2. Điều khiển cảm xúc (Emotional Control)

- **Vấn đề:** ER-NeRF hiện tại chủ yếu tập trung vào việc **khớp khẩu hình (Lip-sync)**. Khuôn mặt thường giữ một biểu cảm trung tính hoặc chỉ bắt chước biểu cảm ngẫu nhiên từ dữ liệu train. Nó chưa thể hiện được: *Giận dữ, Vui vẻ, Buồn bã* theo ý muốn.
- **Hướng phát triển:** Tách biệt (Disentangle) yếu tố **Cảm xúc (Emotion)** khỏi yếu tố **Nội dung nói (Content)**.
 - *Ý tưởng:* Thêm một vector điều kiện cảm xúc (Emotion vector) vào mô hình RAM. Ví dụ: Input = Audio + "Happy" \rightarrow Môi cười khi nói.

3.3. Khả năng tổng quát hóa (Generalization / One-shot Synthesis)

- **Vấn đề lớn nhất:** ER-NeRF là mô hình **Person-Specific**. Muốn tạo video cho ông Obama, bạn phải train lại từ đầu bằng video của ông Obama. Muốn làm cho bạn, phải train lại bằng video của bạn.
- **Hướng phát triển:** Xây dựng mô hình **One-shot Talking Head**.
 - *Ý tưởng:* Train một mô hình cực lớn trên hàng nghìn khuôn mặt (như cách VASA-1 hay EMO làm). Sau đó, chỉ cần đưa **1 bức ảnh tĩnh bất kỳ** vào là mô hình có thể tạo video nói chuyện ngay lập tức mà không cần train lại (Zero-shot/One-shot).

3.4. Mở rộng phần thân và cử chỉ tay (Torso & Gesture Generation)

- **Vấn đề:** ER-NeRF hiện tại dùng kỹ thuật 2D warping cho phần thân, chỉ lược bỏ nhẹ. Nó không thể tạo ra các chuyển động phức tạp như: vung tay, chỉ tay, gật đầu mạnh.
- **Hướng phát triển:** Kết hợp với các mô hình sinh chuyển động cơ thể (Body Motion Generation).
 - *Ý tưởng:* Dùng Audio để dự đoán không chỉ chuyển động môi mà cả chuyển động tay (Gesture), sau đó dùng mô hình 3D Body (như SMPL) để điều khiển phần thân thay vì chỉ warping ảnh 2D.

Tài liệu tham khảo

- | | |
|--|---|
| [1] ER-NeRF paper | https://arxiv.org/abs/2307.09323 |
| [2] RAD-NeRF paper | https://arxiv.org/abs/2211.12368 |
| [3] AD-NeRF paper | https://arxiv.org/abs/2103.11078 |
| [4] LiveTalking luồng cho video facetime | https://github.com/lipku/LiveTalking |