Tên đề tài (Tiếng Việt): Nghiên cứu mô hình Học sâu Không gian - Thời gian (Spatiotemporal Deep Learning) phát hiện Cảm xúc vi mô và Trạng thái Bối rối của người học trong môi trường E-learning.

Tên đề tài (Tiếng Anh - Suggested): *Spatiotemporal Micro-Expression Recognition for Learner Confusion Detection using Dual-Stream Vision Transformers and Motion Magnification.*

1. Đặt vấn đề (Problem Statement)

Trong giáo dục, trạng thái "Bối rối" (Confusion) là điểm gãy quan trọng. Nếu giảng viên không phát hiện kịp thời, người học sẽ chuyển sang trạng thái "Chán nản" (Boredom) và bỏ cuộc.

- **Thách thức khoa học:** Các mô hình hiện tại (như VGG16, ResNet trên ảnh tĩnh) hoàn toàn thất bại với MEs vì:
    1. **Cường độ thấp:** Sự thay đổi cơ mặt quá nhỏ, dễ bị nhầm lẫn với nhiễu (noise) hoặc chuyển động đầu tự nhiên.
    2. **Thời lượng ngắn:** Nếu camera quay 30fps, một ME chỉ diễn ra trong khoảng 5-10 frames. Việc bỏ lỡ (missing) rất dễ xảy ra.
- **Câu hỏi nghiên cứu (Research Question):** Làm thế nào để trích xuất và phóng đại các đặc trưng chuyển động cực nhỏ (subtle motions) từ chuỗi video để phân loại chính xác trạng thái "bối rối", bất chấp sự thay đổi về ánh sáng và tư thế đầu?

2. Mục tiêu nghiên cứu (Research Objectives)

1. **Phát triển thuật toán Phóng đại chuyển động (Motion Magnification):** Áp dụng kỹ thuật tiền xử lý để làm rõ các chuyển động cơ mặt nhỏ mà mắt thường khó thấy, nhưng không làm biến dạng khuôn mặt.
2. **Xây dựng kiến trúc mạng lai (Hybrid Network):** Kết hợp trích xuất đặc trưng không gian (Spatial - hình dáng mắt/miệng) và thời gian (Temporal - sự thay đổi theo time-step) sử dụng **3D-CNN** hoặc **Transformers**.
3. **Tối ưu hóa phát hiện đỉnh (Apex Frame Detection):** Tự động tìm ra frame có biểu cảm mạnh nhất trong chuỗi video ngắn để tăng độ chính xác phân loại.

3. Tổng quan tài liệu & Khoảng trống nghiên cứu (Literature Review & Gap Analysis)

3.1. Phương pháp truyền thống (Hand-crafted Features):

- Trước 2018, phương pháp chủ đạo là **LBP-TOP** (Local Binary Patterns on Three Orthogonal Planes). Phương pháp này tính toán sự thay đổi kết cấu trên 3 mặt phẳng không gian-thời gian. Tuy nhiên, nó rất kém khi có sự thay đổi ánh sáng hoặc đầu người chuyển động [1].

3.2. Phương pháp Học sâu hiện đại (Deep Learning Approaches):

- **3D-CNN (C3D, I3D):** Xử lý video như một khối 3D ($W \times H \times Time$) . Tuy nhiên, chi phí tính toán cực lớn [2].
- **Optical Flow & Two-Stream Networks:** Tách video thành 2 luồng: luồng RGB (màu sắc) và luồng Optical Flow (chuyển động). Đây là hướng tiếp cận rất hiệu quả cho MEs.
- **Vision Transformers (ViT) (SOTA 2023-2024):** Các nghiên cứu mới nhất đang áp dụng cơ chế *Self-Attention* để mô hình tập trung vào các vùng động (mắt, khóe miệng) và bỏ qua vùng tĩnh (má, trán) [3].

3.3. Khoảng trống nghiên cứu (The Gap): Hầu hết các nghiên cứu hiện nay chỉ tập trung vào 3 tập dữ liệu phòng thí nghiệm (CASME II, SAMM, SMIC) với các cảm xúc cơ bản (Vui, Ngạc nhiên, Kìm nén).

- **Chưa có mô hình chuyên sâu cho sự "Bối rối" (Confusion):** "Bối rối" là một trạng thái phức tạp, thường là sự kết hợp của *Ngạc nhiên (Surprise)* + *Cau mày (Frowning)*.
- **Thiếu ứng dụng thực tế:** Các mô hình SOTA hiện nay quá nặng để chạy real-time trên web học tập.

4. Tài liệu tham khảo minh chứng (References)

Chọn lọc các bài báo 2-3 năm gần đây (2022-2024) để đảm bảo tính thời sự:

1. **Về Micro-expression SOTA:**
   - *Paper:* "Micro-Expression Recognition Using Transformer with Introduction of Magnification" (IEEE Access, 2023). *Minh chứng cho việc cần dùng Transformer và Magnification.*
   - *Paper:* "Learner Confusion Detection Using Facial Micro-expressions and Eye Gaze" (Computers and Education: Artificial Intelligence, 2024). *Đây là bài báo sát sườn nhất với đề tài này.*
2. **Về kỹ thuật phóng đại (Magnification):**
   - *Paper:* "Learning to Magnify Motion in Video for Micro-Expression Recognition" (CVPR 2023 Workshop).
3. **Về Dataset giáo dục:**
   - *Dataset Reference:* "DAiSEE: A Dataset for Affect in User Interfaces with an Emphasis on Standard and E-Learning Environments" (Dùng để chứng minh sự tồn tại của dữ liệu Confusion).

5. Phương pháp nghiên cứu & Kiến trúc đề xuất (Methodology)

NCS cần thiết kế một **Pipeline 3 giai đoạn** chặt chẽ:

Giai đoạn 1: Tiền xử lý & Phóng đại chuyển động (Preprocessing & Magnification)

Micro-expressions quá nhỏ, nên nếu đưa trực tiếp vào CNN, mạng sẽ coi đó là nhiễu.

- **Giải pháp:** Sử dụng **Eulerian Video Magnification (EVM)** hoặc mô hình học sâu **MagNet**.
- **Nguyên lý:** Phân tách video thành các tần số không gian (spatial frequencies). Giữ lại các tần số thấp (background) và khuyếch đại các tần số thay đổi theo thời gian (temporal variations) trong dải tần số của chuyển động cơ mặt ($0.5 - 2Hz$).

$$I'(x, y, t) = I(x, y, t) + \alpha \cdot B(x, y, t)$$

(α là hệ số phóng đại, giúp cái nhíu mày trở nên rõ ràng hơn).

Giai đoạn 2: Trích xuất đặc trưng Optical Flow (Optical Flow Extraction)

Để loại bỏ sự ảnh hưởng của màu da hay ánh sáng, chúng ta sử dụng Optical Flow (TV-L1 hoặc DeepFlow) để tạo ra bản đồ chuyển động.

- Bản đồ này chỉ chứa thông tin: "Pixel nào đang di chuyển và di chuyển hướng nào?". Điều này giúp loại bỏ thông tin thừa về danh tính người học.

Giai đoạn 3: Mạng Dual-Stream Attention (Kiến trúc lõi)

Đề xuất mô hình 2 nhánh song song:

1. **Spatial Stream (ResNet-18):** Đầu vào là **Apex Frame** (frame có biểu cảm rõ nhất). Nhánh này học đặc trưng hình học (ví dụ: lông mày đang ở vị trí thấp).
2. **Temporal Stream (3D-CNN hoặc LSTM):** Đầu vào là chuỗi **Optical Flow**. Nhánh này học động lực học (ví dụ: tốc độ nhíu mày nhanh hay chậm).
3. **Fusion Layer:** Kết hợp 2 nhánh bằng cơ chế **Attention**, cho phép mạng tự trọng số hóa xem thông tin nào quan trọng hơn tại thời điểm đó.

6. Kế hoạch thực nghiệm (Implementation)

6.1. Dữ liệu (Datasets)

NCS cần làm việc với các bộ dữ liệu sau:

- **CASME II & SAMM (Training cơ bản):** Đây là benchmark bắt buộc cho mọi nghiên cứu về MEs. Nó chứa các nhãn: *Positive, Negative, Surprise, Others*.
- **DAiSEE (Fine-tuning):** Bộ dữ liệu lớn nhất về trạng thái người học (Boredom, Confusion, Engagement, Frustration). Đây là chìa khóa để áp dụng vào Education.

6.2. Metrics đánh giá

- Tuyệt đối không dùng Accuracy vì dữ liệu MEs rất mất cân bằng (biểu cảm Neutral chiếm 90%).

- Sử dụng **UF1 (Unweighted F1-score)** và **UAR (Unweighted Average Recall)**.

7. Tài nguyên & Công cụ (Resources)

Đây là những "vũ khí" tốt nhất để sinh viên bắt đầu:

A. GitHub Repositories (Code SOTA):

1. **Micro-Expression-Recognition (Apex):**
   - Keyword: Micro-Expression APEX GitHub.
   - Mô tả: Chứa các code cơ bản để phát hiện frame đỉnh (Apex frame) - bước quan trọng nhất.
2. **RCN-A (Recurrent Convolutional Network for MEs):**
   - Tìm kiếm: RCN-A micro expression github.
   - Đây là kiến trúc chuẩn mực kết hợp CNN và LSTM cho bài toán này.
3. **EVM (Eulerian Video Magnification):**
   - Code MATLAB/Python của MIT CSAIL. Dùng để thực hiện bước phóng đại chuyển động.

B. Sách & Giáo trình chuyên sâu:

- *Handbook of Face Recognition (3rd Edition)* - Stan Z. Li & Anil K. Jain. (Chương về Facial Expression Analysis).
- *Advanced Methods and Deep Learning in Computer Vision* (Tài liệu đã có trong đề cương) - Chương về 3D CNNs.

Lời khuyên:

- **Spotting vs. Recognition:** Hãy thu hẹp phạm vi.
  - *Spotting:* Chỉ tìm xem *khi nào* có ME xảy ra (từ giây thứ 3.1 đến 3.5).
  - *Recognition:* Đã biết có ME, phân loại nó là gì.
  - *Lời khuyên:* Làm **Recognition** trước vì dễ hơn. Spotting là bài toán rất khó (hardcore).
- **Cross-database Validation:** Đừng train và test trên cùng một tập dữ liệu (dễ bị overfitting). Hãy Train trên CASME II và Test trên SAMM để chứng minh mô hình có khả năng tổng quát hóa (Generalization).