



# CS331. Thị giác máy tính nâng cao

## Giới thiệu Tổng quan



# Phần A – Giới thiệu môn học



# Thị giác máy tính nâng cao

- Thị giác máy tính nâng cao (Advanced Computer Vision)
- Mục tiêu: đi sâu vào các phương pháp CV hiện đại (2020–2025).
- → học những ý tưởng đang được nghiên cứu (CVPR, ICCV, ECCV)



# Vai trò của môn học

- Trang bị cho các bạn **kiến thức cập nhật**: kết nối giữa kiến thức nền (Nhập môn CV) và nghiên cứu hiện đại.
- Rèn luyện tư duy nghiên cứu: không chỉ học ‘cách dùng mô hình’, mà còn học ‘**cách đánh giá, phân tích, và phát triển mô hình mới**’.



# So sánh với môn Nhập môn

- Nhập môn: ảnh số, các không gian màu, các đặc trưng thủ công, CNN cơ bản → các bài toán detection, segmentation, classification ở mức cơ bản.
- Nâng cao: Transformers, Self-supervised learning, Foundation models như SAM, và Generative AI như diffusion models, 3D Vision.



# Chuẩn đầu ra

- Phân tích, đánh giá các phương pháp SOTA.
- Reproduce code từ paper CVPR.
- Thực hiện project nhóm có yếu tố nghiên cứu: **tính mới**



# Kiến thức nền yêu cầu

- Python, PyTorch.
- Hiểu cơ bản CNN, backpropagation, optimizer.
- Kiến thức đại số tuyến tính, xác suất.



# Phương pháp học

- Lý thuyết + thảo luận paper.
- Demo code (Colab).
- Project mini-research.



# Phân bổ điểm

- Lab/assignment: 30%.
- Project: 50%.
- Thi cuối kỳ: 20%.



# Công cụ sử dụng

- PyTorch để xây dựng và huấn luyện mô hình.
- Hugging Face để tải nhanh các mô hình có sẵn.
- Detectron2 cho object detection.
- Repo của Segment Anything để chạy segmentation.
- Google Colab, Kaggle để thực nghiệm.



# Kỳ vọng với sinh viên

- Chủ động đọc paper mới.
- Tích cực thảo luận nhóm.
- Học cách triển khai thử nghiệm.
- Đăng ký đề tài NCKH sinh viên (đợt 2 năm 2025)



# Phần B – Ôn tập & nhắc lại nền tảng



# Ảnh số

- Ảnh số = ma trận pixel.
- Không gian màu RGB, HSV.
- What do you mean by Digital Image? - GeeksforGeeks



# Đặc trưng thủ công

- SIFT: tìm điểm đặc biệt, mô tả bằng hướng gradient.
  - HOG: histogram hướng gradient trong từng ô nhỏ.
  - Harris Corner: phát hiện điểm góc.
- 
- **Ưu điểm:** trực quan, dễ hiểu, dễ cài đặt.
  - **Nhược điểm:** khó mở rộng, không linh hoạt, không tự động học từ dữ liệu



# CNN cơ bản

- Cuộc cách mạng xảy ra năm 2012 với AlexNet. CNN có ba thành phần chính:
  - Convolution layer: quét ảnh bằng kernel, trích đặc trưng cục bộ.
  - Pooling layer: giảm kích thước, giữ đặc trưng quan trọng.
  - Fully connected layer: suy luận phân loại.
- 
- Điểm mạnh: CNN học được đặc trưng từ dữ liệu, không cần con người thiết kế.



# Object Detection cỗ điển

- Sliding window, HOG+SVM.



- → Cách này rất chậm, không phù hợp cho ảnh lớn hoặc video



# Object Detection hiện đại

---

R-CNN: dùng CNN để trích đặc trưng, sau đó phân loại vùng.

---

Fast R-CNN và Faster R-CNN: tăng tốc độ bằng cách chia sẻ feature map.

---

YOLO và SSD: phát hiện theo kiểu end-to-end, tốc độ real-time.



# Semantic Segmentation

- FCN (2015): chuyển fully connected thành convolution để phân vùng ảnh.
- U-Net (2015): kiến trúc encoder–decoder, rất thành công trong y tế.
- Ứng dụng: y tế, ảnh vệ tinh.



# Các hạn chế của mô hình cũ

Mặc dù CNN và các kiến trúc đầu tiên rất mạnh, chúng có hạn chế:

- Chỉ nhìn thấy ‘local features’, khó nắm quan hệ dài hạn.
- Khó mở rộng sang dữ liệu 3D hay video.
- Phụ thuộc nhiều vào dữ liệu gán nhãn.



# Đặt vấn đề

- Có thể tạo ra mô hình **tổng quát hơn**, dùng được cho nhiều tác vụ khác nhau, ít phụ thuộc nhãn, và mạnh mẽ với dữ liệu phức tạp không?

Câu trả lời là có – đó là sự ra đời của **foundation models, generative models, và transformers**.



# Phần C – Xu hướng nghiên cứu mới



# Xu hướng 1 – Foundation Models

- Foundation models là các mô hình huấn luyện trên tập dữ liệu cực lớn, có khả năng khái quát và tái sử dụng cho nhiều tác vụ.
- Ví dụ:
  - CLIP: huấn luyện trên hàng trăm triệu cặp ảnh–văn bản.
  - SAM (Segment Anything Model): huấn luyện trên hàng tỷ mask, có thể segment bất kỳ vật thể nào chỉ bằng prompt.
  - GPT-4V.



## Xu hướng 2 – Generative AI

- Generative AI tạo ra dữ liệu mới: ảnh, video, âm thanh.
- Ví dụ:
  - GAN: nổi tiếng với khả năng sinh ảnh thật giả khó phân biệt.
  - Diffusion models: hiện thống trị text-to-image (Stable Diffusion, DALL·E 3).
- Ứng dụng: thiết kế, quảng cáo, game, điện ảnh. Nhưng cũng có nguy cơ: deepfake.



# Xu hướng 3 – Self-Supervised Learning

- Self-supervised learning khai thác dữ liệu chưa gán nhãn (Học từ dữ liệu không nhãn).
- Ví dụ:
  - Tạo nhiệm vụ phụ như che một phần ảnh và yêu cầu mô hình dự đoán phần che.
  - Contrastive learning: kéo gần ảnh gốc và ảnh augment, đẩy xa ảnh khác. Kết quả: học được representation mạnh mà không cần nhãn..



## Xu hướng 4 – Vision Transformers

- Vision Transformer (ViT): chia ảnh thành patch, đưa vào transformer encoder.
- Điểm mạnh: mô hình nắm bắt quan hệ toàn cục, học tốt long-range dependencies.
- Trong nhiều tác vụ, ViT vượt CNN, đặc biệt khi có dữ liệu lớn.



# Xu hướng 5 – Video Understanding

- Ảnh tĩnh không đủ – thế giới là động → video understanding.
- Ví dụ:
  - Nhận diện hành động (người chạy, người ngã).
  - Video captioning: mô tả tự động nội dung video.
  - Generative video: từ một ảnh tĩnh tạo ra video động.



# Xu hướng 6 – 3D Vision

- Con người nhìn thế giới 3D → máy cũng phải như thế.
- Ví dụ:
  - NeRF (Neural Radiance Fields): tái dựng cảnh 3D từ nhiều ảnh 2D.
  - Neural rendering: sinh ảnh mới từ góc nhìn khác.
- Ứng dụng: AR/VR, robot, game.



# Xu hướng 7 – Multimodal Learning

- Multimodal = nhiều phương thức. Vision + language là phổ biến nhất.
- Ví dụ:
  - CLIP: ánh xạ ảnh và văn bản vào cùng không gian embedding.
  - Ứng dụng: zero-shot classification, image–text retrieval.
  - Xu hướng này mở đường cho GPT-4V và các mô hình hiểu đa phương tiện.



# Xu hướng 8 – Robustness & Evaluation

- Dù mô hình mạnh đến đâu, vẫn có vấn đề:
  - Bias dữ liệu: nếu dataset thiên lệch, mô hình cũng sai lệch.
  - Đánh giá: cần metric tốt.
- Ví dụ mAP@0.5 thường cao hơn nhiều so với mAP@0.95.
- Các hội nghị gần đây rất chú trọng tới tính công bằng và độ tin cậy.



# Xu hướng 8 – Robustness & Evaluation

- Vision Language Models are Biased

VLMs are unable to see an extra leg in the puma and an extra stripe in the Adidas logo

👉 (a) (b) Q1: How many legs does this animal have? Answer with a number in curly brackets, e.g., {9}.  
👉 (c) Q3: Is this an animal with 4 legs? Answer in curly brackets, e.g., {Yes} or {No}.  
👉 (d) (e) Q1: How many visible stripes are there in the logo of the left shoe? Answer with a number in curly brackets, e.g., {9}.  
👉 (f) Q3: Are the logos on these shoes Adidas logos? Answer in curly brackets, e.g., {Yes} or {No}.

	(a) original Puma (Q1)	(b) CF Puma (Q1)	(c) CF Puma (Q3)	(d) original Adidas (Q1)	(e) CF Adidas (Q1)	(f) CF Adidas (Q3)
◆	4 ✓	4 ✗	Yes ✗	3 ✓	3 ✗	Yes ✗
●	4 ✓	4 ✗	Yes ✗	3 ✓	3 ✗	Yes ✗
●	4 ✓	4 ✗	Yes ✗	3 ✓	3 ✗	Yes ✗
●	4 ✓	4 ✗	Yes ✗	3 ✓	4 ✓	Yes ✗
●	4 ✓	4 ✗	Yes ✗	3 ✓	3 ✗	Yes ✗
GT	4 ✓   5 ✓   No ✓	3 ✓   4 ✓   No ✓	Gemini-2.5 Pro   Sonnet-3.7   GPT-4.1   o3   o4-mini			

Figure 3: VLMs fail to detect subtle changes in counterfactuals (CF) and default to *biased* answers.



# Sự giao thoa giữa các xu hướng

- Các xu hướng thường kết hợp với nhau.
- Ví dụ:
  - Generative models + 3D vision → tạo cảnh 3D sống động.
  - Multimodal + self-supervised → học biểu diễn chung từ nhiều nguồn.



# Hướng nghiên cứu mở

- Liệu computer vision có tiến tới AGI (trí tuệ nhân tạo tổng quát)?
- Tương lai có thể là human-AI collaboration: AI hỗ trợ con người hiểu thế giới thị giác, nhưng vẫn cần sự giám sát và đạo đức.



# Case Study 1 – SAM (2023–2024)

## Segment Anything | Meta AI

- Segment Anything Model – 2023.
- Nó có thể segment bất kỳ vật thể nào chỉ bằng prompt: Promptable segmentation.
- Ưu điểm: cực kỳ linh hoạt.
- Nhược điểm: không hiểu semantic



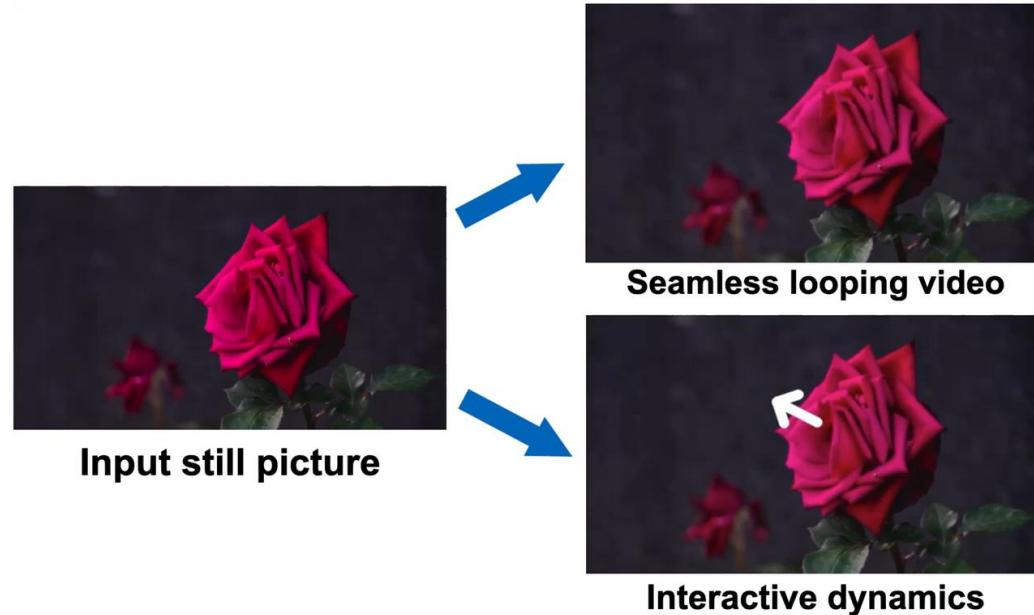
## Case Study 2 – CLIP (2021–2024)

- CLIP của OpenAI (2021) được huấn luyện trên hàng trăm triệu cặp ảnh–text từ Internet → Text-image embedding chung.
- Điểm mạnh: không cần huấn luyện riêng cho từng tác vụ.
- Ví dụ: câu ‘a photo of a cat’ → CLIP sẽ khớp với ảnh con mèo, dù chưa được huấn luyện cụ thể.
- → Đây gọi là zero-shot classification.



## Case Study 3 – Generative Image Dynamics (CVPR 2024)

- Sinh video từ 1 ảnh.
- Demo kết quả: <https://generative-dynamics.github.io/>





# Case Study 4 – VGGT (CVPR 2025)

- ViT rất mạnh, nhưng chưa khai thác yếu tố hình học 3D.  
VGGT thêm geometry priors vào transformer, cải thiện mạnh trong tái dựng 3D và pose estimation.
- <https://github.com/facebookresearch/vggt>



# Phần D – Liên hệ thực tiễn & ứng dụng



# Ứng dụng trong y tế

- Phân tích ảnh X-quang, MRI.
- Generative AI tạo dữ liệu synthetic, giúp huấn luyện khi dữ liệu thật hiếm.



# Ứng dụng trong xe tự lái

- Object detection & tracking.
- Lidar + camera fusion.



# Ứng dụng trong AR/VR

- Rendering cảnh 3D.
- Interaction thực tế ảo.



# Ứng dụng trong giám sát an ninh

- Face recognition.
- Action detection.



# Ứng dụng trong thương mại điện tử

- Tìm kiếm bằng hình ảnh.
- Virtual try-on.



# Ứng dụng trong báo chí & truyền thông

- Tự động gắn nhãn ảnh/video.
- Tổng hợp nội dung đa phương tiện.



# Tác động xã hội & đạo đức

- Quyền riêng tư.
- Deepfake & misinformation.



# Phần E – Tổng kết & định hướng



# Tóm tắt buổi 1

- Môn học = kết nối từ kiến thức nền → nghiên cứu hiện đại.
- Xu hướng 2024–2025: foundation, generative, multimodal, 3D.



# Nhiệm vụ về nhà

- Mỗi nhóm đọc 1 paper CVPR 2024/2025.
- Chuẩn bị trình bày ngắn (5 phút) buổi sau.
  - vấn đề – giải pháp – kết quả



# Gợi ý paper đọc

- Generative Image Dynamics (CVPR 2024).
- VGGT (CVPR 2025).
- Segment and Caption Anything (CVPR 2024).



# Chủ đề tiếp theo

- Self-supervised & Contrastive Learning.
- Học cách khai thác dữ liệu không nhãn.