# CS6240: Multimedia Analysis

CHUA Tat-Seng (蔡达成)

National University of Singapore

KITHCT Chair Professor

Co-Director, NExT Research Center

# Objectives of this Course

To Introduce:

- The rich and exciting areas of multimedia research
- The background & history to MM Research
- The recent advances in research and applications
- The future trends
- The criteria for a good (multimedia) research

To guide:

- The planning of a good research project
- The writing of a good research paper/report

# Objectives of this Course

- Is this a MM or a Computer Vision course?
- Is it a Machine Learning course?

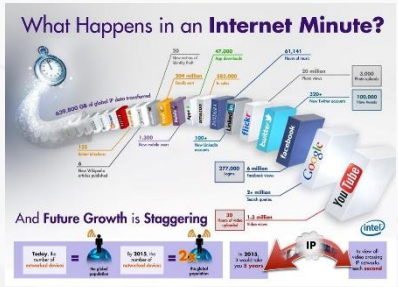
- It is Multimedia as it covers more than visual topics, especially on the integration of language & vision
- It is not ML, as it is primarily about MM techniques and applications, while ML is used as tools to solve MM problems
- The confluence of technologies means that such differentiations are not relevant now.

# OUTLINE

- Trends in AI & MM Research
- History of MM Research
- Key Topics in MM Research to be Covered
- Criteria for Good MM research
- Course Outline and Instructions

# Trend 1: Multi-Channel Data

- Social media sources- live info streams:
  - Spontaneous User-Generated Contents (UGC)
  - Device-Generated Contents (IoT)




- Web sources:


Web Search Engines


Forums

E-commerce Sites

- Knowledge sources:

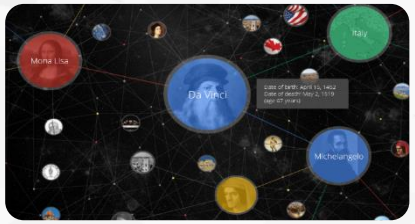

- Other domain sources: offline data, domain data, industry data for vertical domains, such as Fintech..

# Trend 1: Multi Channel Data

## Integration of Multiple Info Sources

Internal Structured Data

External Unstructured (Big) Data



Knowledge Graph

AI Platform

Human (Experts)

Better Augmenting User Decisions

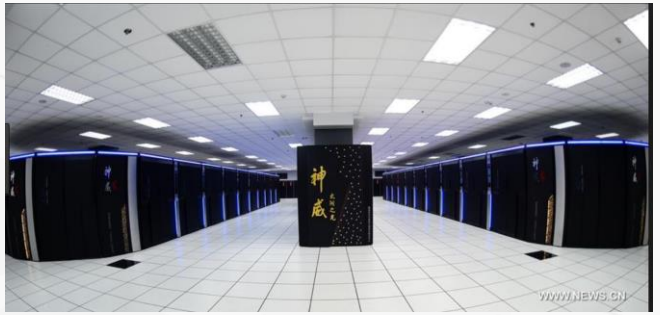
Most new data sources are multimedia in nature

# Trend 2: Emergence of Deep Learning & LLM

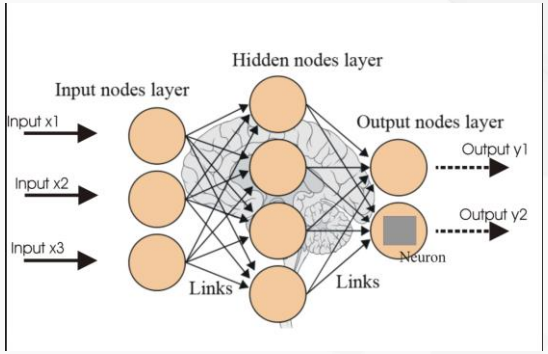
## Conditions are Ripe for the Success of AI

Big Data


Computing Resources



Algorithms




**Knowledge!!**

(Relatively)
Clean Data

Vertical (Narrow) Domain

[Zhang Bo, 2018]

# Trend 2: Emergence of Deep Learning & LLM

## Great Success of AI

**AlphaGo beats Go human champ**

**Deep Net outperforms humans in image classification**

IMAGENET

**Autonomous search-and-rescue drones outperform humans**

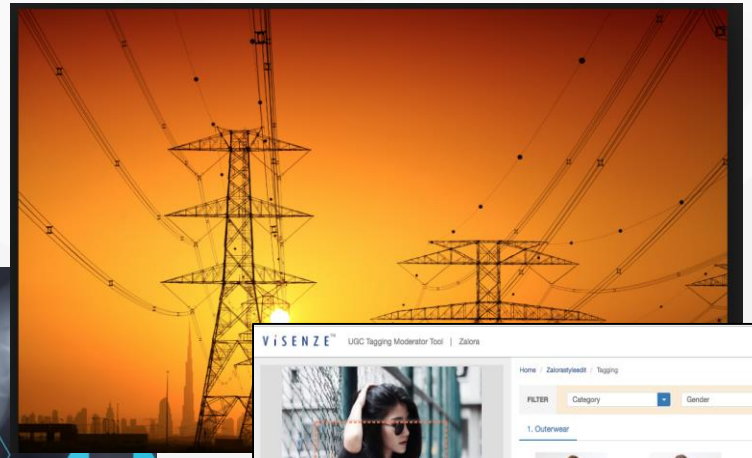
**DeepStack beats professional poker players**

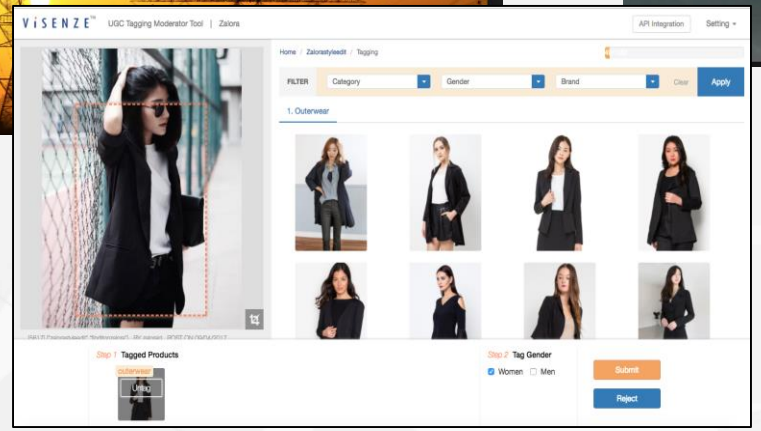
**Computer out-plays humans in "doom"**

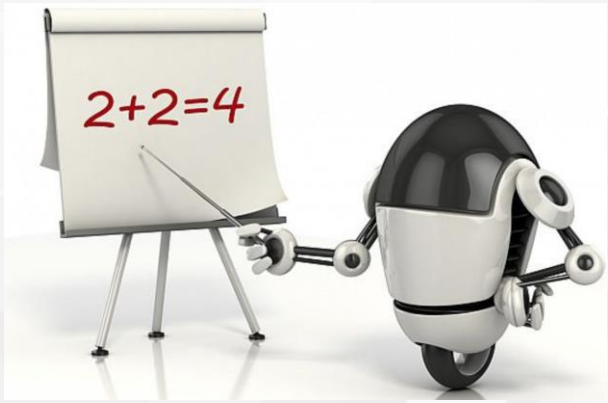

**IBM's Watson destroys humans in jeopardy**


# Trend 2: Emergence of Deep Learning & LLM

## Big AI Opportunity




[Socially Responsible AI, Wendy Hall, 2019]

# Trend 2: Emergence of Deep Learning & LLM

## New Paradigm in Scientific Research

- Big successes in deep-learning data-driven approach evolves into the 4th paradigm in scientific research
- The four paradigms of scientific research are:

  1) Experimental, 2) Theoretical, 3) Computational Sc & 4) Data driven Computing


We are now witnessing the beginning of the Fifth Paradigm – integration of big data and knowledge

Oxford e-Research Centre

# Trend 3: Accountability in Technology (AI)

## Problems and Societal Issues in AI

- The current generation of LLM & DL-based systems:
  - They offer tremendous accuracy and benefits
  - But may make occasional **BIG** mistakes
  - Concerns on their robustness **LIMIT** their applications.
- It brings in issues of accountability:
  - Such as the first accident of self-driving car?
- Big accountability concerns:
  - **Trust** is the key
  - Predictability in results/ performance
  - Accountability of algorithm
  - Data access issues - privacy, security , trust and fairness
  - Ethical and legal issues

### Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian

**Tempe police said car was in autonomous mode at the time of the crash and that the vehicle hit a woman who later died at a hospital**

▲ A car passes the location where a woman pedestrian was struck and killed by an Uber self-driving sport utility vehicle in Tempe, Arizona, on Monday. Photograph: Rick Scuteri/Reuters

An autonomous Uber car killed a woman in the street in Arizona, police said, in what appears to be the first reported fatal crash involving a self-driving vehicle and a pedestrian in the US.

Tempe police said the self-driving car was in autonomous mode at the time of the crash and that the vehicle hit a woman, who was walking outside of the crosswalk and later died at a hospital. There was a vehicle operator inside the car at the time of the crash.

Uber said in a statement on Twitter: “Our hearts go out to the victim’s family. We are fully cooperating with local authorities in their investigation of this incident.”

# Trend 3: Accountability in AI (Technology)

## Legislation Governing the Use of AI

### Composition of the Advisory Council on the Ethical Use of Artificial Intelligence ("AI") and Data

LAST UPDATED 19 DECEMBER 2018

**Singapore**

**The full composition of Singapore's Advisory Council on the Et (Advisory Council) was announced by Minister for Communicat Iswaran at AI Singapore's first year anniversary.**

**SINGAPORE – 30 August, 2018:** The full composition of Singapore's Advisory Council on the Ethical Use o announced by Minister for Communications and Information Mr S Iswaran at AI Singapore's first year anniver June on the establishment of the Advisory Council, to be chaired by former Attorney-General V.K. Rajah SC.

### AI systems should be accountable, explainable, and unbiased, says EU

*The European Union has published new guidelines on developing ethical AI*

By James Vincent | Apr 8, 2019, 12:03pm EDT

The European Union today published a set of guidelines on how companies and governments should develop ethical applications of artificial intelligence.

These rules aren't like Isaac Asimov's "Three Laws of Robotics." They don't offer a snappy, moral framework that will help us control murderous robots. Instead, they address the murky and diffuse problems that will affect society as we integrate AI into sectors like health care, education, and consumer technology.

So, for example, if an AI system diagnoses you with cancer sometime in the future, the EU's guidelines would want to make sure that a number of things take place: that the software wasn't biased by your race or gender, that it didn't override the objections of a human doctor, and that it gave the patient the option to have their diagnosis explained to them.

So, yes, these guidelines are about stopping AI from running amuck, but on the level of admin and bureaucracy, not Asimov-style murder mysteries.

To help with this goal, the EU convened a group of 52 experts who came up with seven requirements they think future AI systems should meet. They are as follows:

# Trend 4: Towards 3rd Generation AI

- 1st Generation: Knowledge (experience) based
  - Symbolic AI
  - Knowledge-Driven
  - E.g.: Deep Blue that beats the world champ in 1997


- 2nd Generation: Connectionist Model
  - Multi-layer (Deep) Neural Networks
  - Data-Driven
  - E.g.: Alpha-Go beats Go Champ in 2016


- 3rd Generation: Integration of Knowledge + data
  - Symbolic + Big Data
  - Typical application scenario: IBM Watson vs. Ken in Feb 2011
  - In a way, Chat-GPT is the outcome of this integration


# Trend 4: Towards 3rd Generation AI

## Principles of Empowerment: Social Machine

Real life is and must be full of all kinds of social constraint – the very processes from which society arises. Computers can help if we use them to create abstract social machines on the Web: processes in which the people do the creative work and the machine does the administration... The stage is set for an evolutionary growth of new social engines. The ability to create new forms of social process would be given to the world at large, and development would be rapid.

Berners-Lee, *Weaving the Web*, 1999 (pp. 172–175)

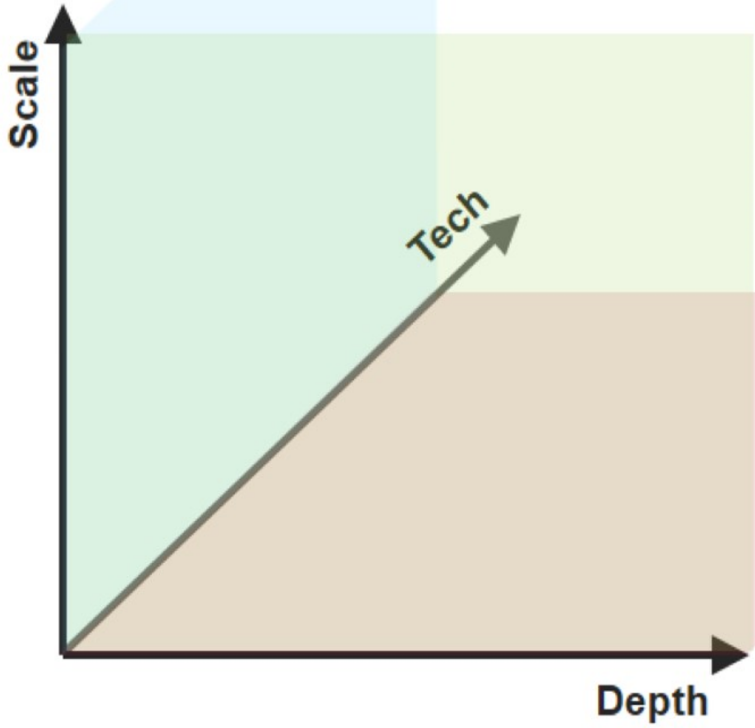
# OUTLINE

- Trends in AI & MM Research
- History of MM Research
- Key Topics in MM Research
- Criteria for Good MM research
- Course Outline and Instructions

# Evolving Multimedia Research

- Multimedia has long been associated with visual media
  - Traditionally it has been visual, visual, visual…
- Text was added to supplement visual analysis with great success
  - Starting with TRECVID Experience on news video retrieval in 2000's
- Social media brings in a new dimension
  - Leveraging noisy tags to annotate image/video
  - Leads to large scale auto (learning) annotation systems
- Experience shows that visual processing alone is inadequate?
  - There is a huge semantic gap
  - The technology is useful only for a few vertical-domain applications
- How to enhance semantics of visual content (the weak link) is of utmost importance

# The History of Visual Processing

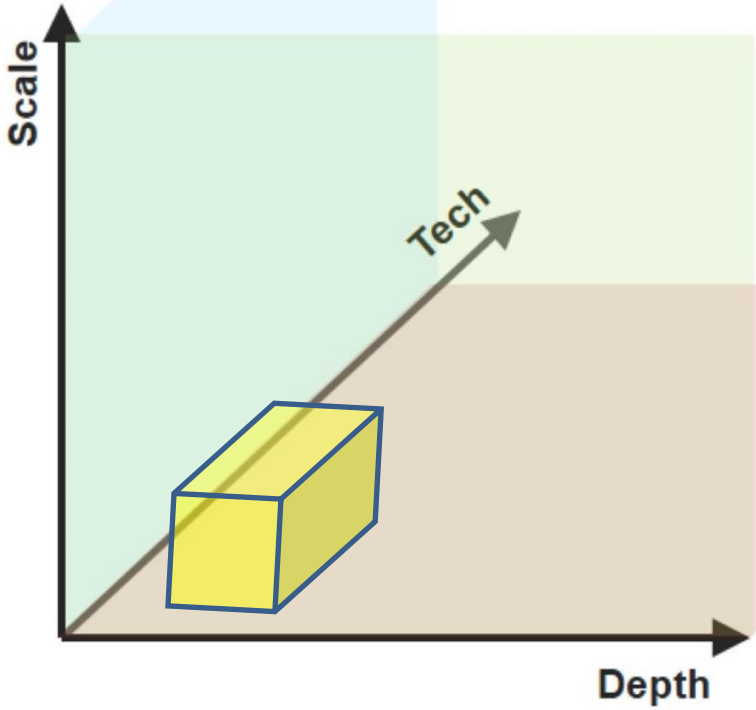



- Scale
  - Amount of Data, # of Concepts
- Depth
  - Content understanding; Features
- Tech
  - Framework; ML Models

# Near 2000: At the end of the early stage

## Small dataset, simple image processing features

Scale: Corel 10s cat. 1,000 s images
Tech: kNN, SVM, color histogram
Depth: Image Level





Smeulders et al. TPAMI'00

# 2000~2005: Fast Progress

## Larger datasets, sophisticated hand-created features

**NUS-WIDE** Chua. ICMR '09

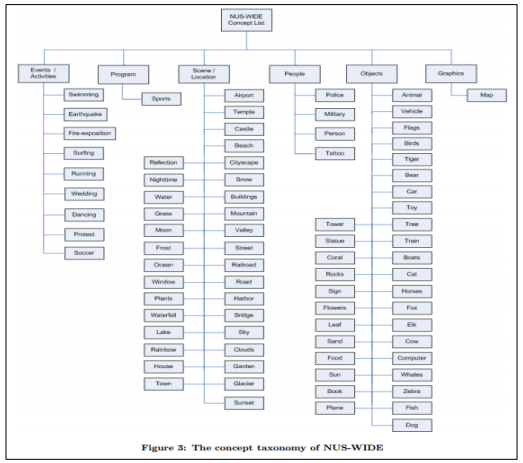
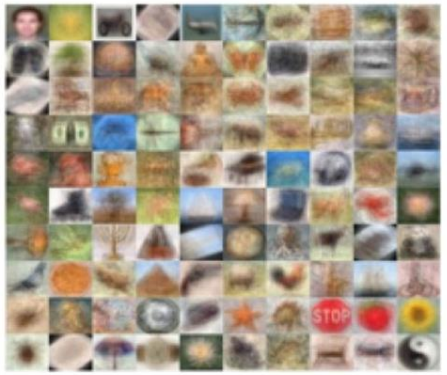
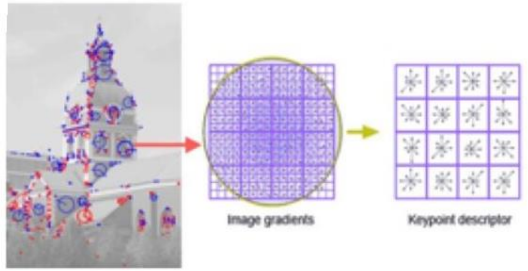


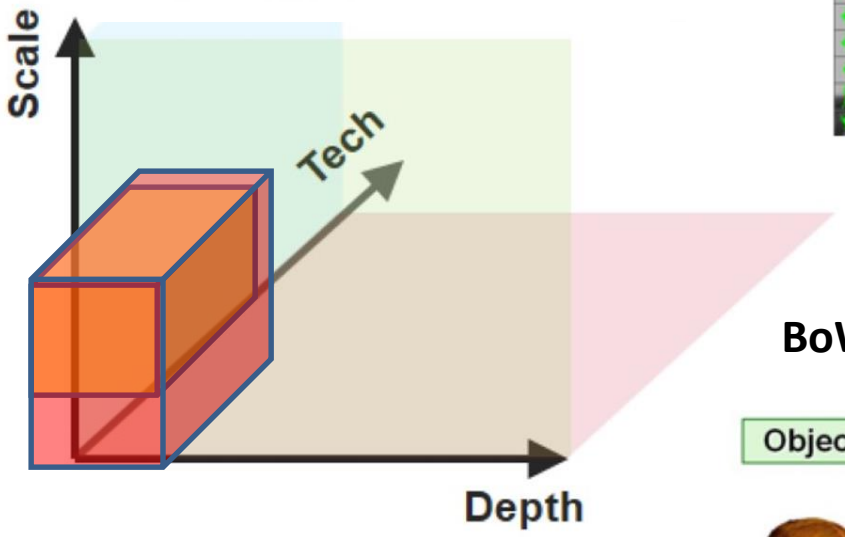

Figure 3: The concept taxonomy of NUS-WIDE

**Caltech101** Fei-Fei. CVPR'05

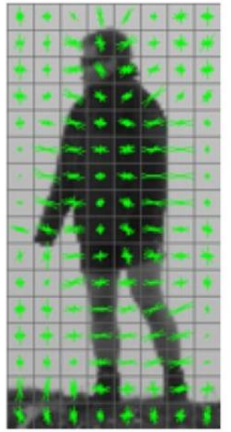



**SIFT** Lowe. IJCV'02

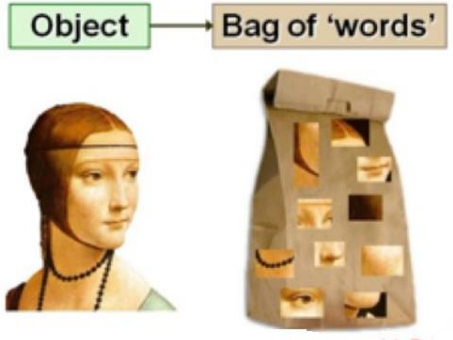



**HoG** Dalal &Triggs. CVPR'05

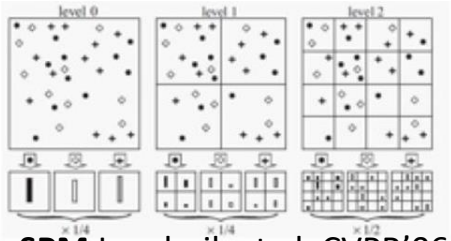


**BoW** Sivic et al. ICCV'05



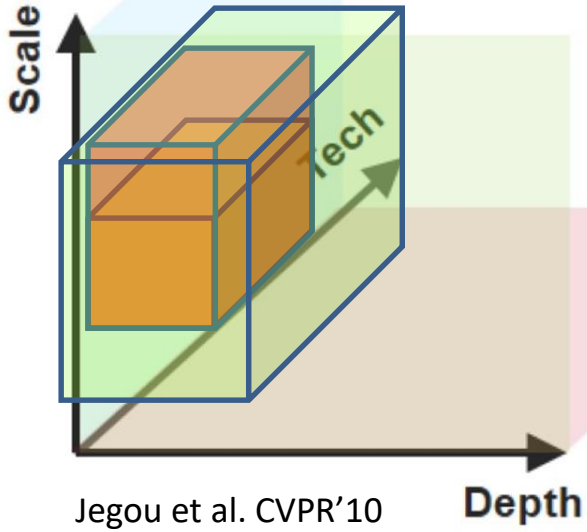
# 2005~2011: The maturity of shallow methods

**VOC** Everingham et al. IJCV'15



**SPM** Lazebnik et al. CVPR'06

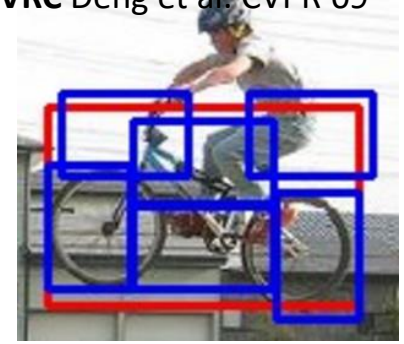

**ScSP** Yang et al. CVPR'09




Jegou et al. CVPR'10

**VLAD**
**FisherVector**
Perronnin et al. ECCV'10

**ILSVRC** Deng et al. CVPR'09



**DPM** Felzenszwalb et al. CVPR'08

**Tricks** Chatfield et al. BMVC'11

# Key Triggers to MM Research
## (Beyond 2010)

- Large-scale media resources available from social media platforms


- Social Media contents becomes more visual – more images & videos

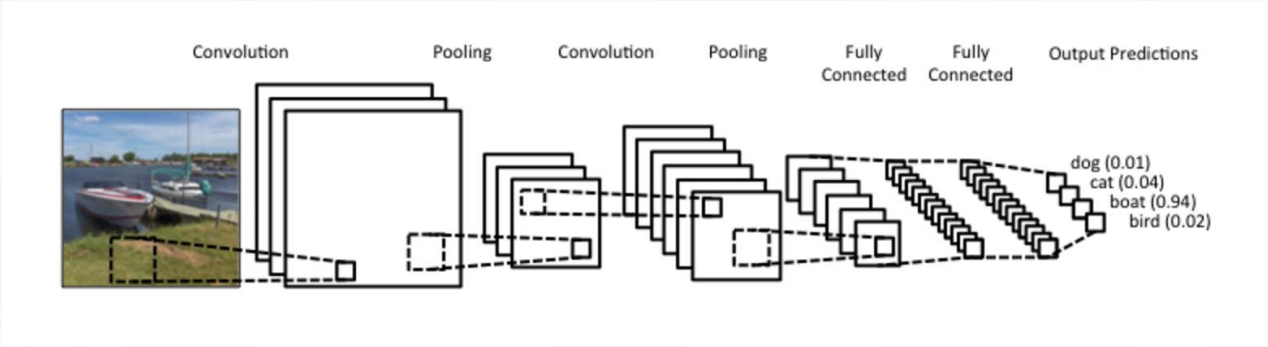

- Research moved towards integrating multimodal data, users and knowledge: in multi-source, multi-task fusion

→ Entering the era of deep learning, improved content representation and multimodal fusion

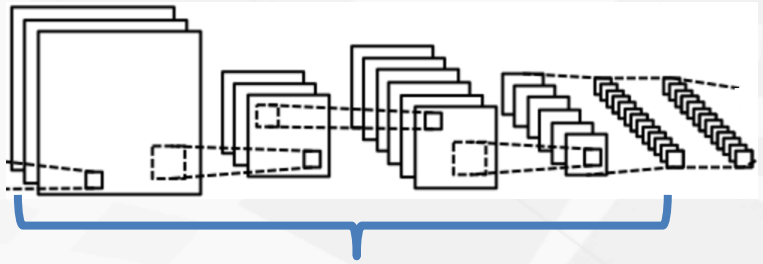
- Emergence of ImageNet
  - 1.3M images, 1K concepts, single-label

# Concept Detection as a Classification Task in Deep Learning Era

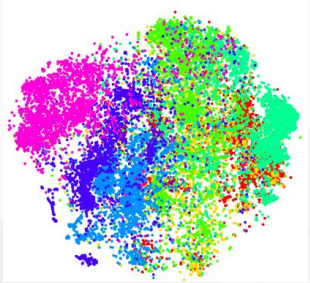
- Map Visual Features to a Fixed Vocabulary (1,000 concepts)




- which Spawns Models and Features for many Other Tasks

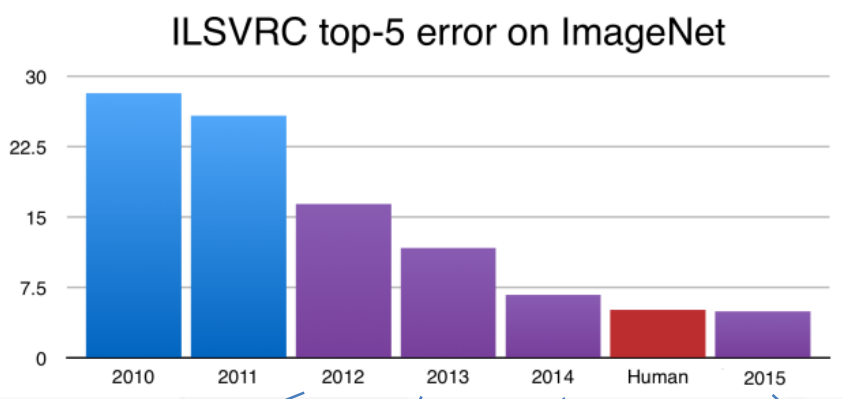


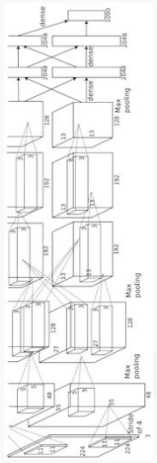
Pre-trained Models for Object Detection



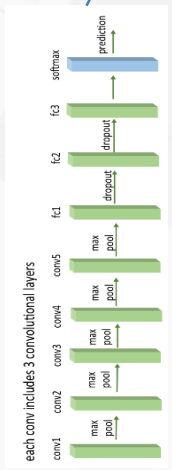
Deep Features for Visual Search

# Classification Performance in Recent Years (2012 - 2015)

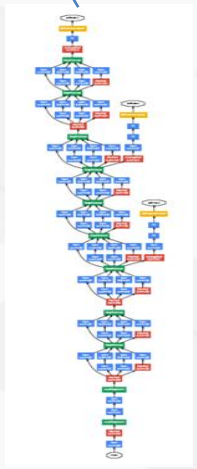


AlexNet
9-layer

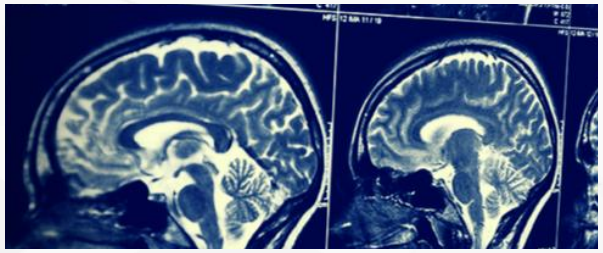



VGG
19-layer



GoogLeNet
22-layer

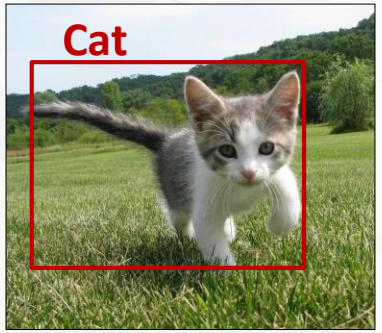
ResNet
152-layer

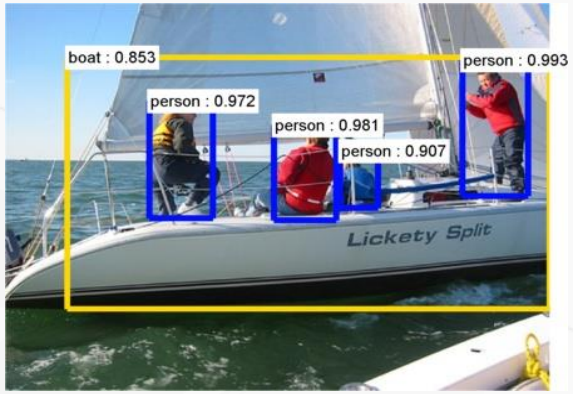


Similarly in
medical
imaging

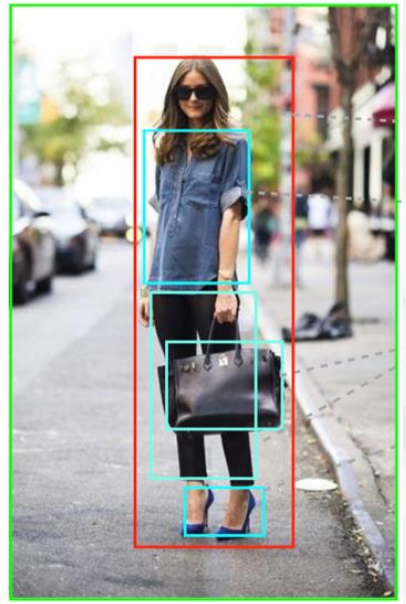
# Beyond Image Classification

1. Fundamental tasks
   - Image classification
   - Object detection (classification + localization)
   - Semantic segmentation
   - Instance segmentation



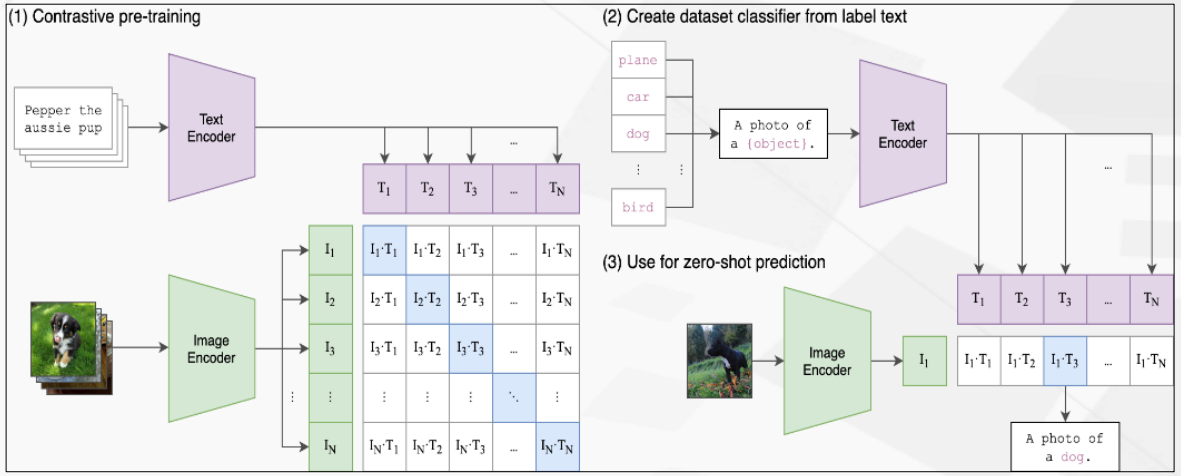

2. Application-driven tasks
   - Face recognition
   - Fashion item recognition
   - ……



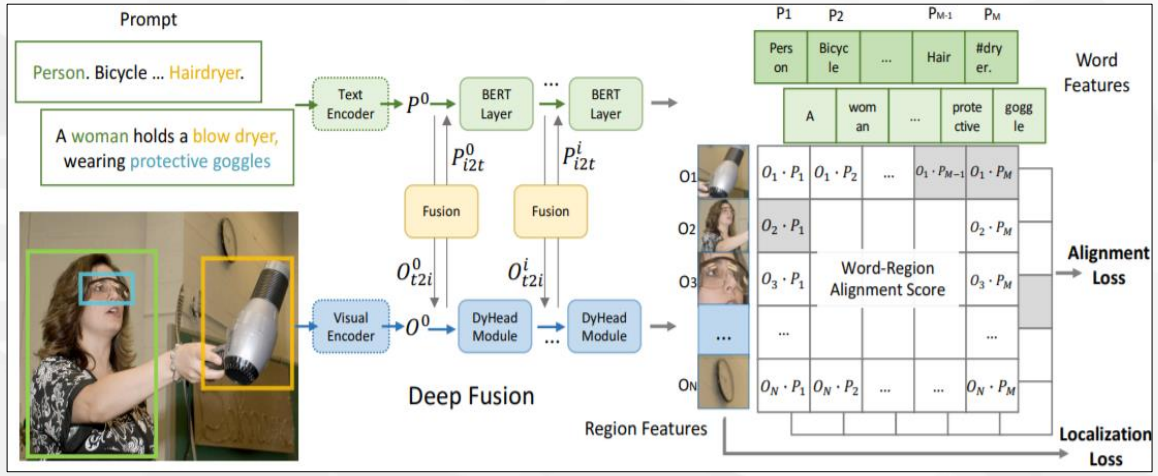
# Beyond 2020: Towards Large Multimodal Foundation Models
## Pre-trained Image-Text Models

### Contrastive Language-Image Pretraining (CLIP)

- Task: image-text matching, support region level visual-text alignment
- Data: 400M image-text + 0.5M text queries (non public)
- Encoder: ResNet-50/ViT for image, BERT for text
- Contrastively learn the cross-model matching between image and text
- Downstream: image recognition, cross-model retrieval, OCR, action recognition…




**CLIP(ICML'21)**




**GLIP(Arxiv)**

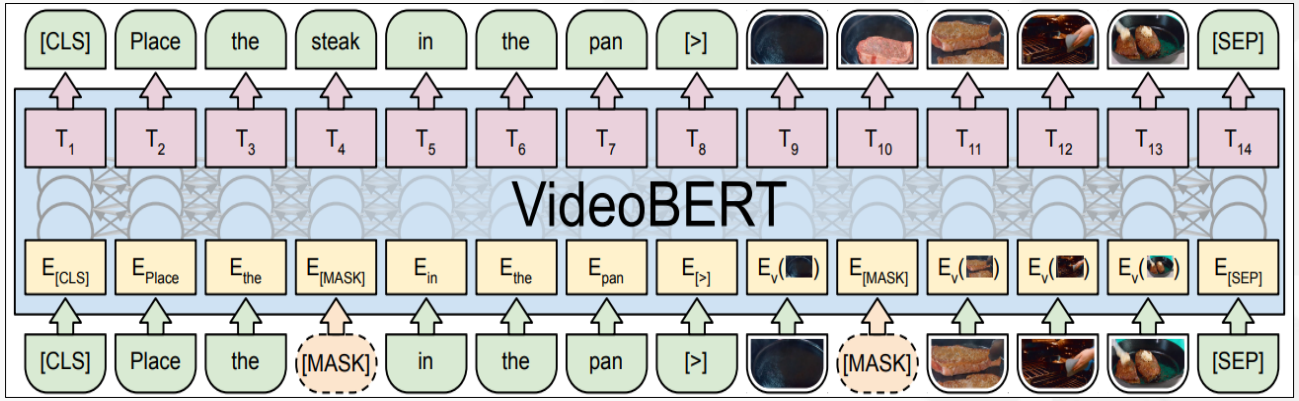
[1] OpenAI. Learning Transferable Visual Models From NL Supervision. ICML'21. [2] MSRA. Grounded Language-Image Pre-training Arxiv.

# Beyond 2020: Towards Large Multimodal Foundation Models
## Pre-trained Video-Text Models

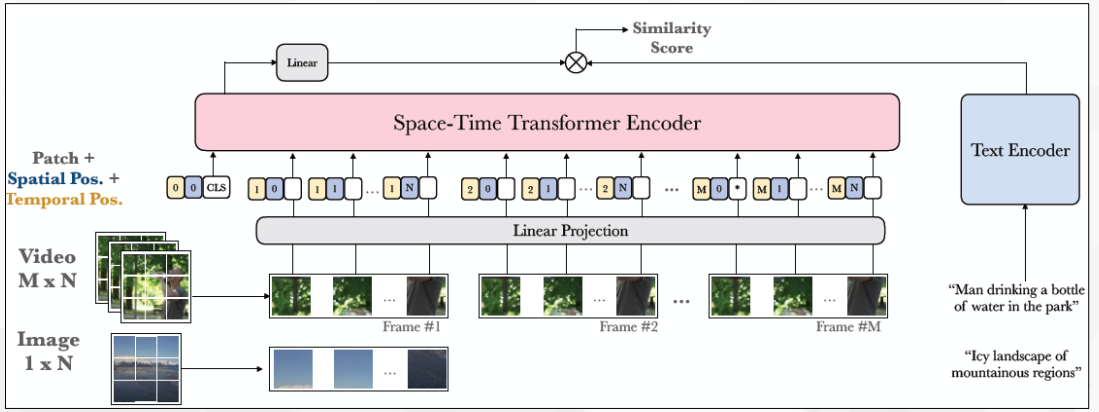
- Many research systems: VideoBERT (by Google), ActBERT (by Baidu), Frozen (OxfordU), videoCLIP (FB), ..

### Video-Text Pretraining (VideoBERT)

- Data: 312K YouTube cooking videos.
- Tasks: text-to-video generation and future forecasting
- Encoder: S3D->Visual Tokens, cross-model Transformer
- Downstream: action classification, video captioning.




**VideoBERT(ICCV'19)**




**Frozen(ICCV'19)**

[1] Google Research. VideoBERT. ICCV'19. [2] VGG OxfordU. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. ICCV'21.

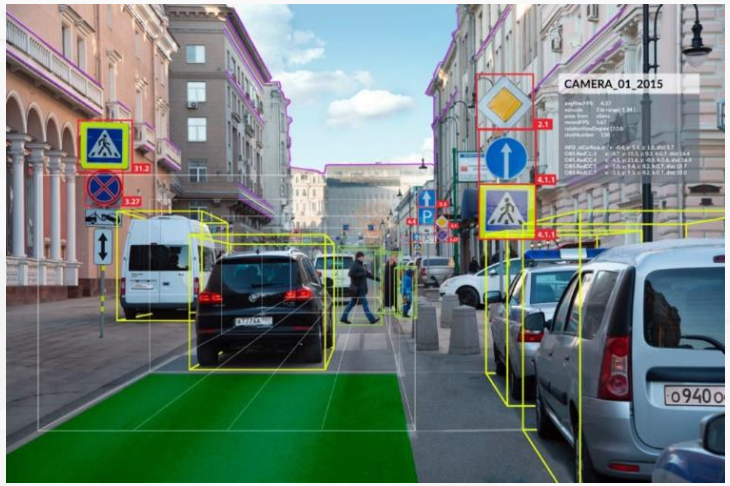
# Visual Processing Ripe for Real-World Applications?

- With rapid progress on object annotations, is visual processing ready for real-world applications?
  - Technologies for deep feature leaning and hashing are maturing
  - Cloud technologies are also well developed with affordable costs
- However, successful applications of visual processing fall only in several vertical domains, such as home security, entertainment and e-commerce

# Home Security

- Built on human/ face and object recognition research
- Leads to multiple unicorns in China:
  - SenseTime (**商汤科技**): USD 7.5B
  - Megvii (**旷视科技**): USD 4B
  - CouldWalk (**云从科技**): 5.5B
  - Yitu (**亿图** ): USD 2.4B 2020


Person Re-targeting

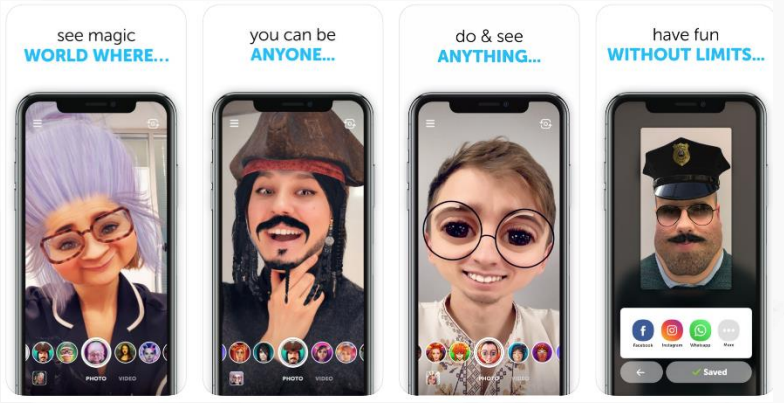


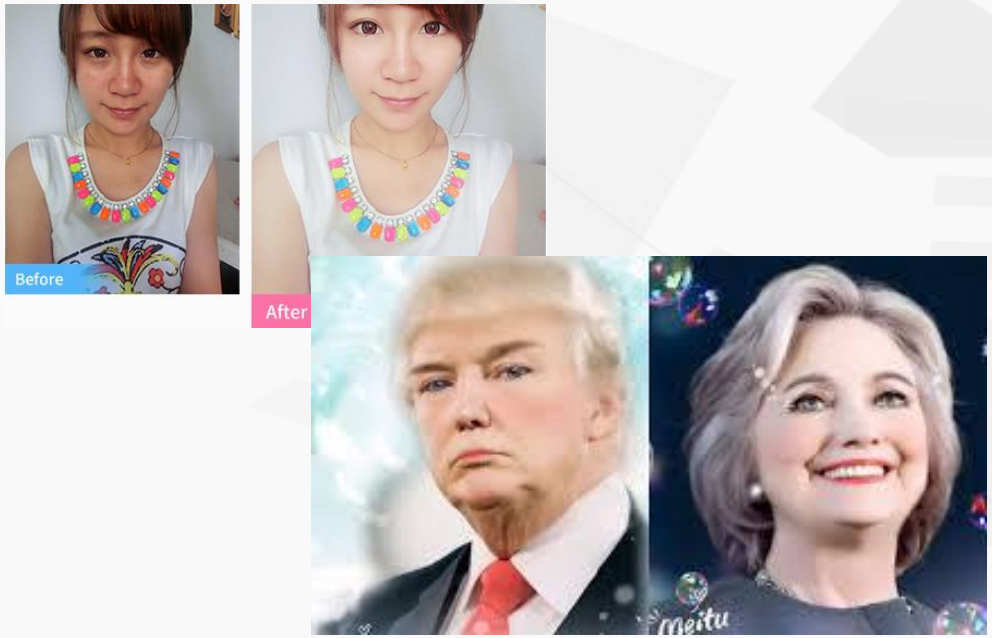

Vehicle Recognition


Human Recognition

# Entertainment

- Built on face recognition and morphing,  and various facial manipulation tools

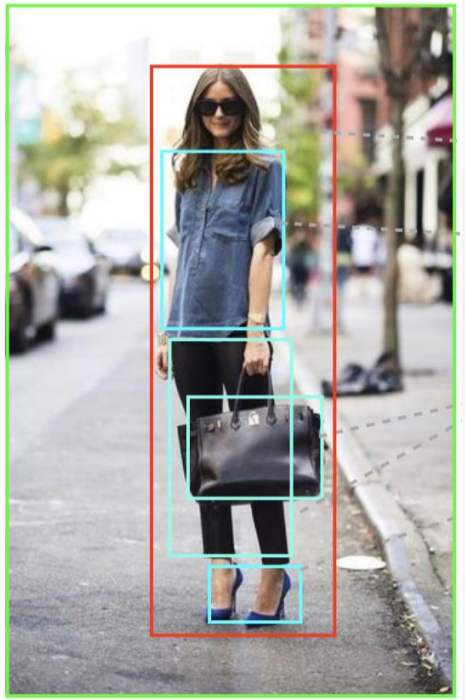


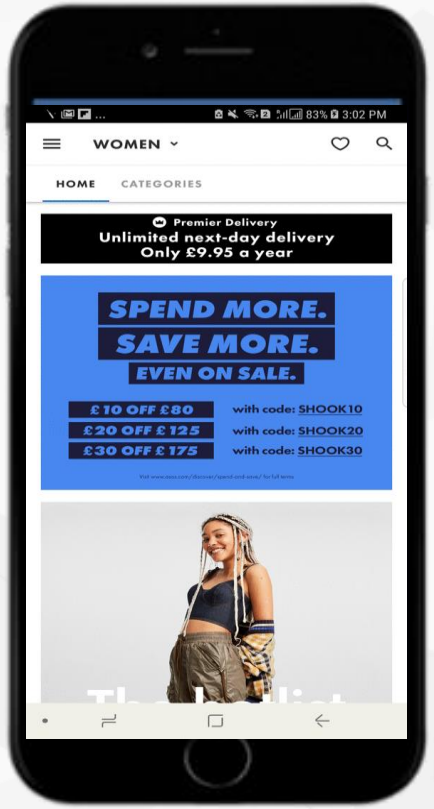

# E-Commerce

- Visual product search: fashion, home furnishing, items, etc.

Human, fashion Item & Background Recognition



Mobile Visual Search




asos UNIQLO H&M

VR: Edge Computing

# OUTLINE

- Trends in AI & MM Research
- History of MM Research
- Key Topics in MM Research to be Covered
- Criteria for Good MM research
- Course Outline and Instructions

# Key Topics to be covered in this Course -1

- Selection of topics and papers is based only on my perspectives

**A. Core Visual & ML Techniques:**

- Object Detection, Recognition & Vision-Language Models
- Semantic and Temporal Segmentation and Relation Grounding
- Cross-modal Alignment and Multimodal Scene Graph
- Few-Shot, Meta and Causal Learning

**B. Generative MM:**

- Condition-based Diffusion Models for MM Generation
- Large Multimodal Foundation Model (LMFM)
- NExT-GPT & Multimodal Dialogues

**C. Responsible AI:**

- Responsible AI: Trust, Safety, Privacy & Biased in MM

# Key Topics to be covered in this Course -2

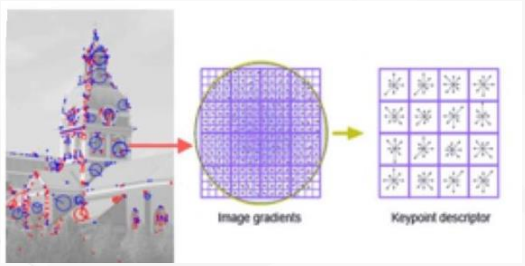
**D. Downstream Techniques & Applications:**

- Image/ Video QA, and Reasoning
- Multimodal Event (& Fashion) Detection & Forecasting
- MM Recommendation

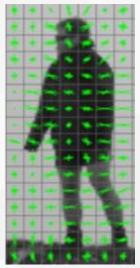
# A) Core Visual & ML Techniques
## From Hand-Crafted to Auto Features

- Disappearing hand-crafted features

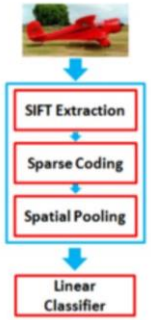
**SIFT**
(Lowe. IJCV'02)



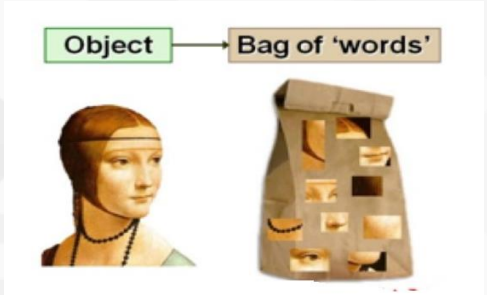

**HoG**
(Dalal &Triggs. CVPR'05)



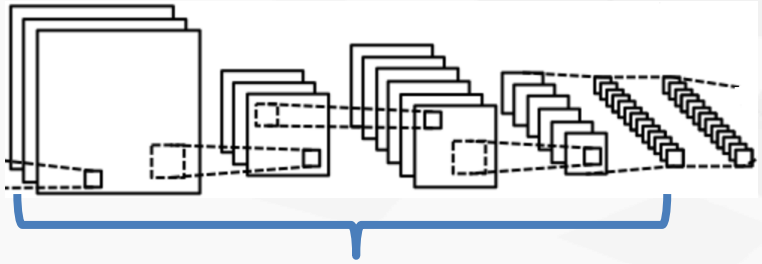
**VLAD Fisher Vector**
(Perronnin et al. ECCV'10)



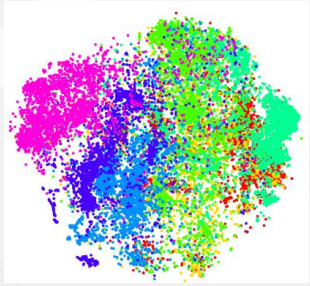

**BoW**
(Sivic et al. ICCV'05)




- Towards deep features



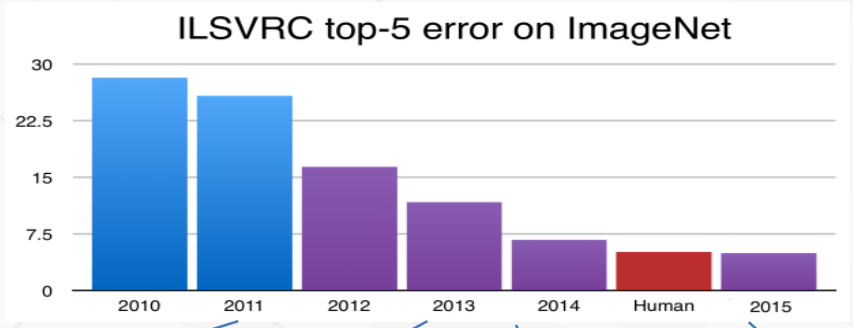
Pre-trained Models for Object Detection

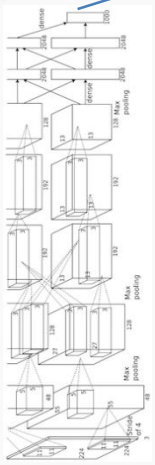


Deep Features for Visual Search

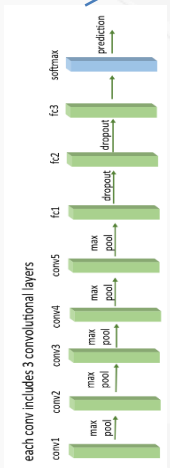
# A) Core Visual & ML Techniques

## Deeper Networks and Bigger Data

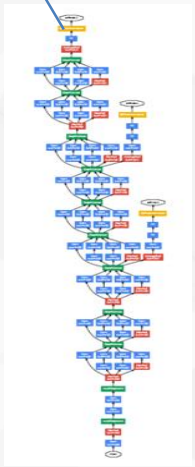


AlexNet
9-layer




VGG
19-layer



GoogLeNet
22-layer

ResNet
152-layer

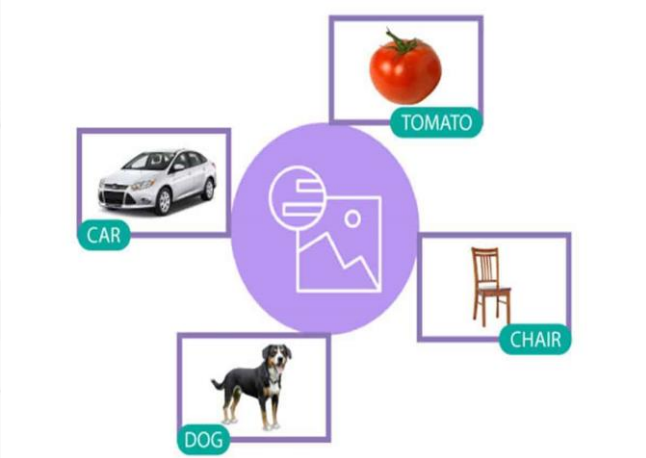
- Towards deeper network and bigger data for better accuracy
- Evolved into Large pre-trained model and Large Foundation Models (LFMs)

- Difficult for academia to make meaningful contributions on large-scale LFMs
- Need to focus on small scale LFMs (<10B) for vertical domains, for most commercial applications where a combination of accuracy, trust and privacy is important

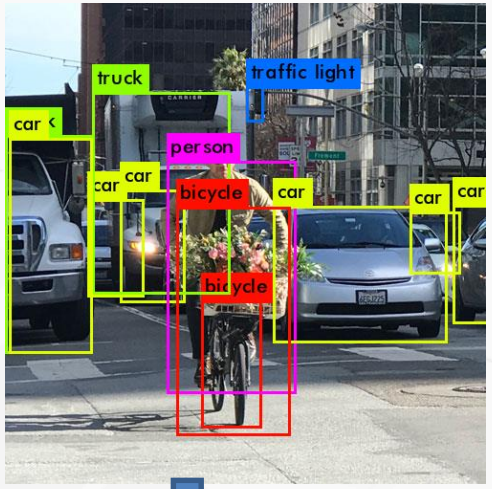
# A) Core Visual & ML Techniques

## More Complex Visual Analysis and Alignment Tasks

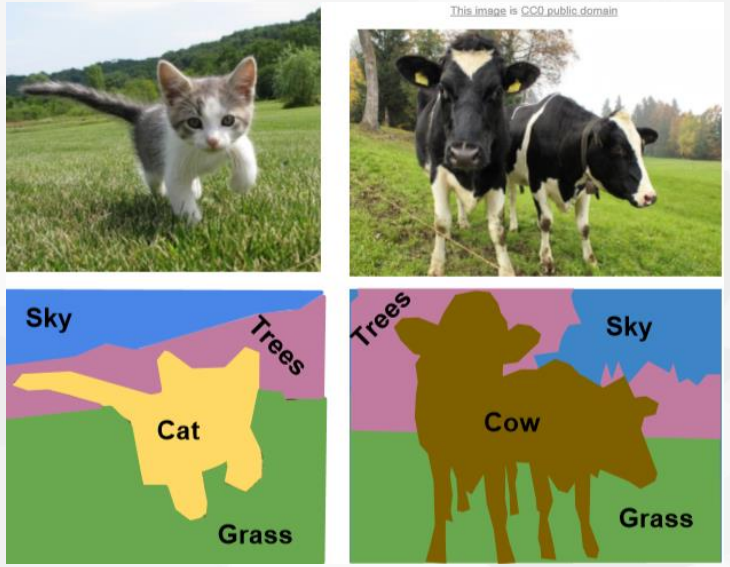
Image Classification: whole image to label (ResNet)




Object Detection: classification + position (R-CNN)




Segmentation: pixel-level (Mask R-CNN)



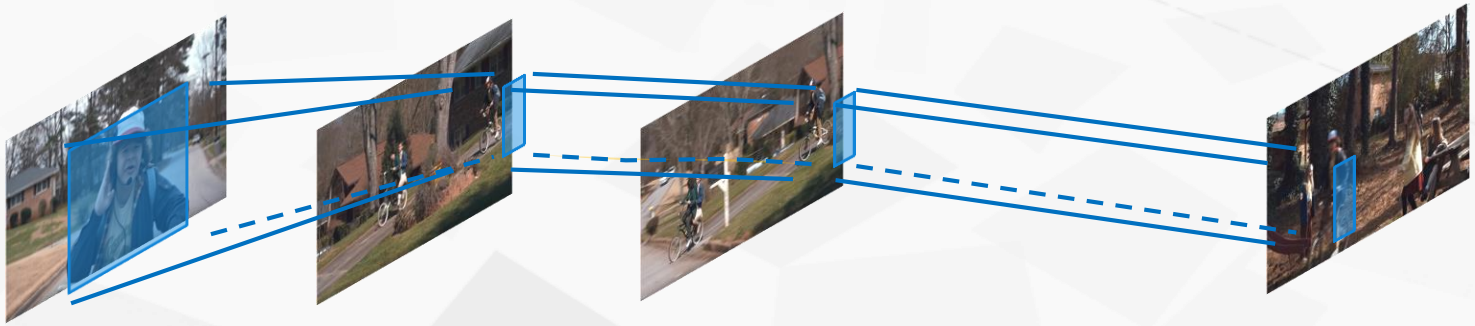

Generalized to various alignment tasks:

- Spatial-temporal segmentation, relation grounding
- Text-vision alignment
- Alignment of visual tokens and text tokens in LFMs
- ……

# A) Core Visual & ML Techniques

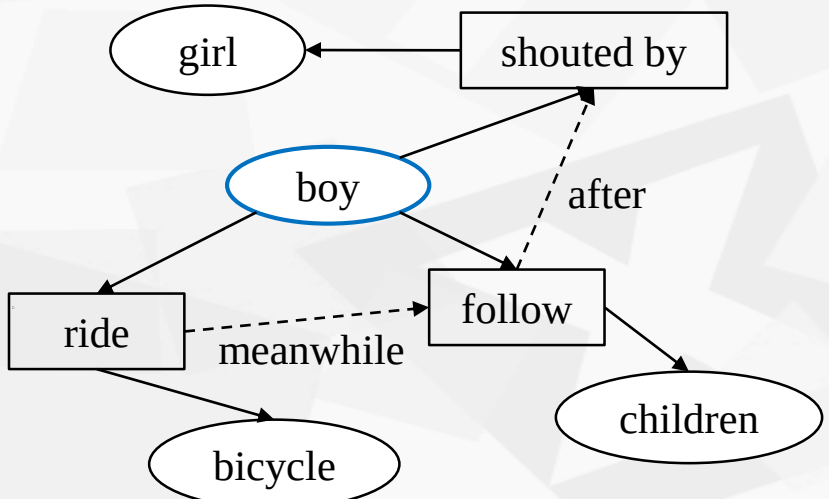
## (Dynamic) Scene Relation Graph from Video

*Spatio-Temporal Grounding*



- Uniquely refer an object in video using some of its involved visual relations and temporal relations of the visual relations
- Should be extended to multimodal scene graphs – towards visual-language integration

***Scene Relation Graph***



# Key Topics to be covered in this Course

A. **Core Visual & ML Techniques:**
   - Object Detection, Recognition & Vision-Language Models
   - Semantic and Temporal Segmentation and Relation Grounding
   - Cross-modal Alignment and Multimodal Scene Graph
   - Few-Shot, Meta and Causal Learning

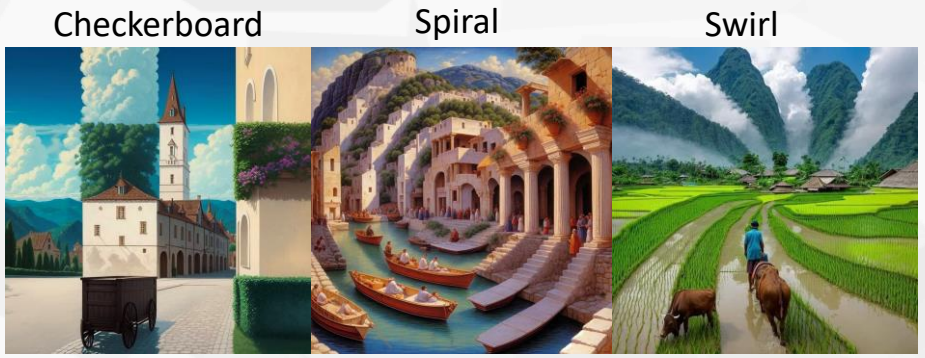
B. **Generative MM:**
   - Condition-based Diffusion Models for MM Generation
   - Large Multimodal Foundation Model (LMFM)
   - NExT-GPT & Multimodal Dialogues

C. **Responsible AI:**
   - Responsible AI: Trust, Safety, Privacy & Biased in MM

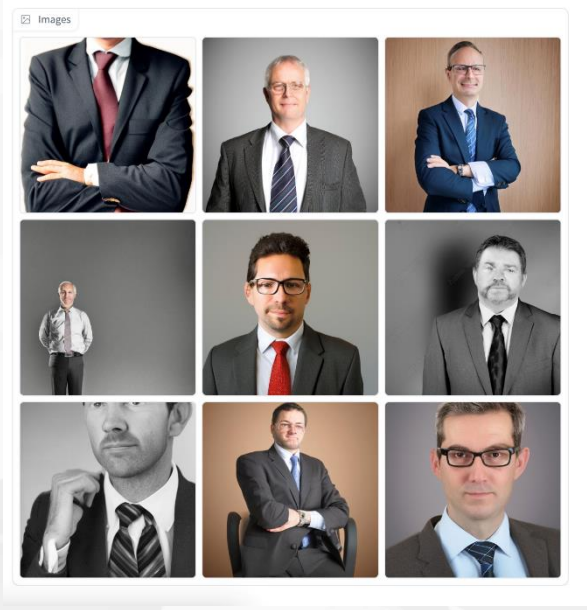
# B) Generative MM

- Great improvements in diffusion-based models for image generation
  - DALLE-2 to DALLE-3: quality has improved tremendously but still has problem of quality and biased
  - Need better “balanced” training data, or conditioning to alleviate these problems
  - Biased: (1) “people in positions of authority” such as “ceo”, “manager”, “doctor” -> white & male; (2) while nurse” “emotional”, “sensitive” -> female) [1]

    One work tackle it as a distributional alignment problem
  - Expression of artistic contents
  - Improvement of quality (e.g. symmetry)
- Issues in LMFM
  - Fine-grained concept alignment
  - Alignment of linguistic and visual tokens
  - Usual issues in LLM



Artistic Conditioning using DALL-E 3

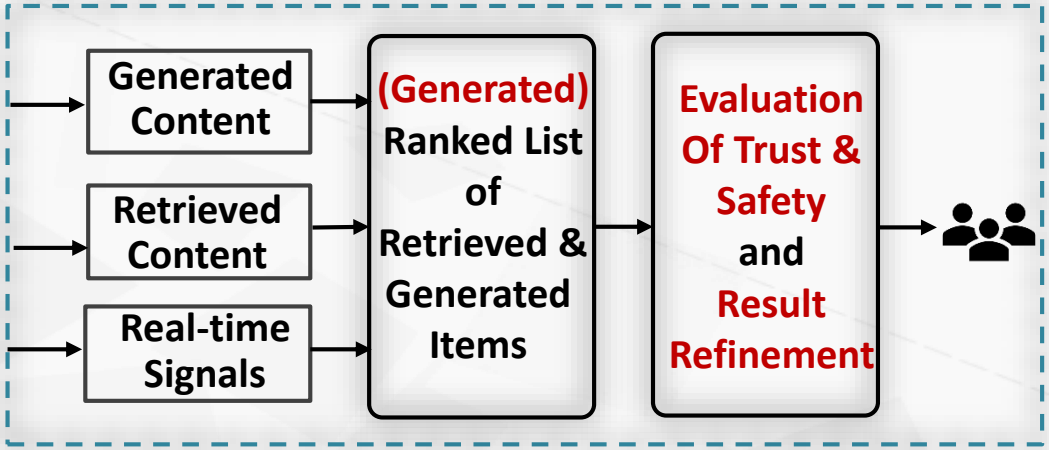


Images of “Managers” by Stable Diffusion

[1] Melissa M: These new tools let you see for yourself how biased AI image models are. MIT Technology Review, Mar 2023.

# C) Responsible AI

- **Trust and Safety of AIGC**
  - Problems of hallucination and erroneous reasoning are inherent part of generative AI: need a combination of external resources and self evaluation to tackle this problem
  - Key issues: Trust & Authenticity; Bias & fairness; Privacy; Safety
  - Legal compliance: new laws and regulations
  - Identifiability: digital watermark, detection, and tracking




- **Need new approaches for evaluation**
  - **Item-side evaluation**. *e.g.,* Fréchet Video Distance (FVD) for measuring micro-video quality
  - **User-side evaluation**. Online evaluation by traditional metrics, *e.g.,* CTR, dwell time

- **Audit LLMs from multiple perspectives (measures)**
  - **Performance measures:** speed, accuracy, diversity ..
  - **Content measures:** commonsense knowledge, domain knowledge, weaknesses ..
  - **Trust measures:** trust & authenticity; bias & fairness; privacy, safety, legal compliance …

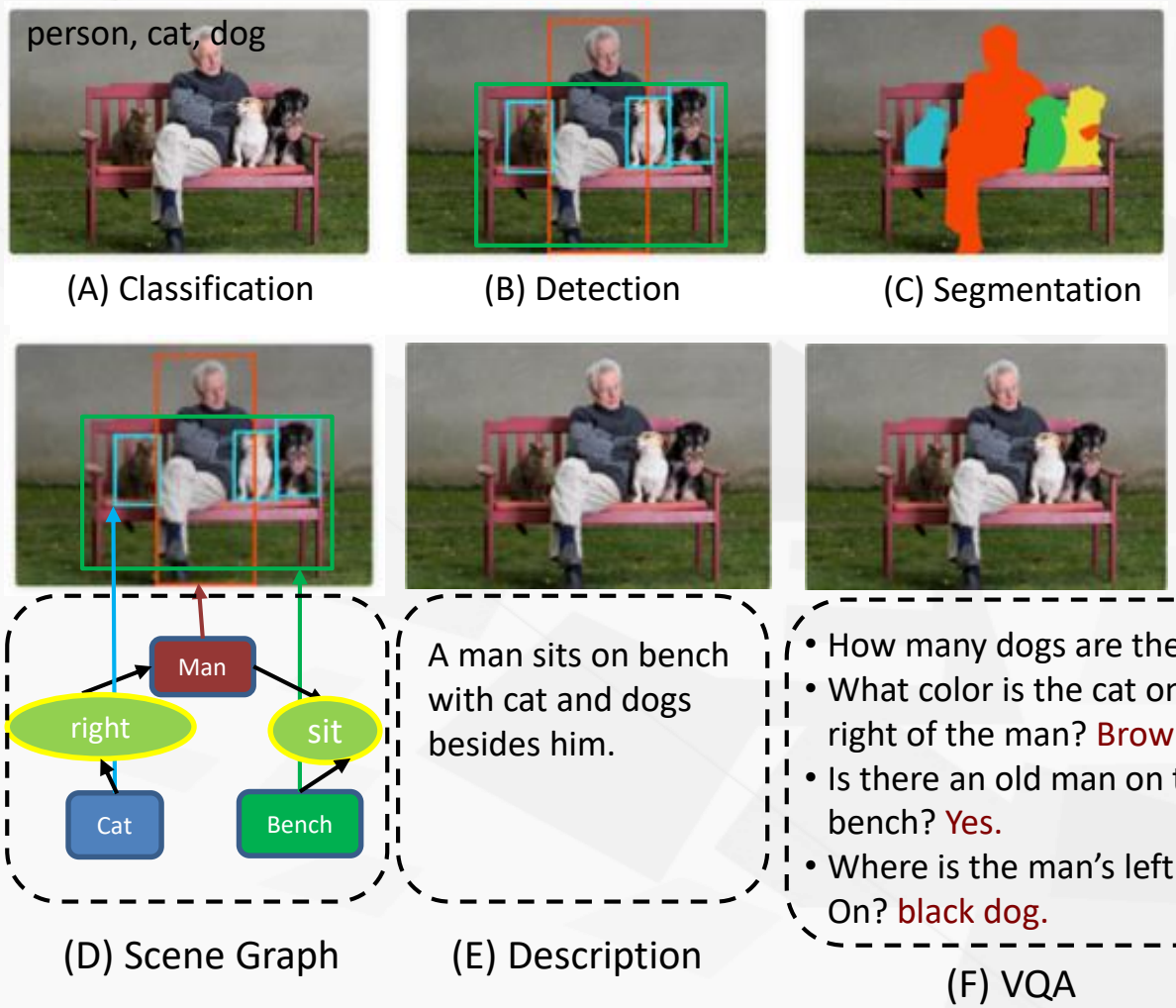
# Key Topics to be covered in this Course

**D. Downstream Techniques & Applications:**

- Image/ Video QA, and Reasoning
- Multimodal Event (& Fashion) Detection & Forecasting
- MM Recommendation

# D) Downstream Techniques & Applications:
## From Visual Classification to Captioning and QA

- Towards AI-compete visual understanding



# D) Downstream Techniques & Applications:
## Multimodal Dialogue and Fashion Search


**Motivation:**

Existing dialogue systems only utilize textual information, which is not enough for full understanding of the dialogue.

- What are “these”?
- What is “it”?

Note: both User utterances and Agent utterances can be of text and image modality

- Key research: How to integrate text and visual model in conversation:
  How do they reinforce each other to express the multimodal intent precisely?

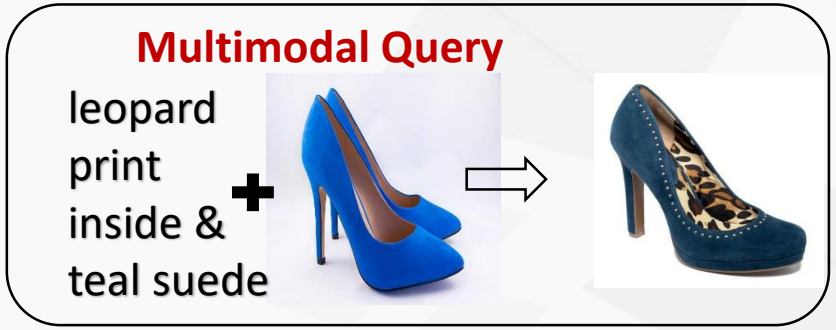
Liao et al.: Knowledge-aware Multimodal Dialogue Systems, ACM Multimedia 2018

# D) Downstream Techniques & Applications:
## Query Generation for Conversational Search

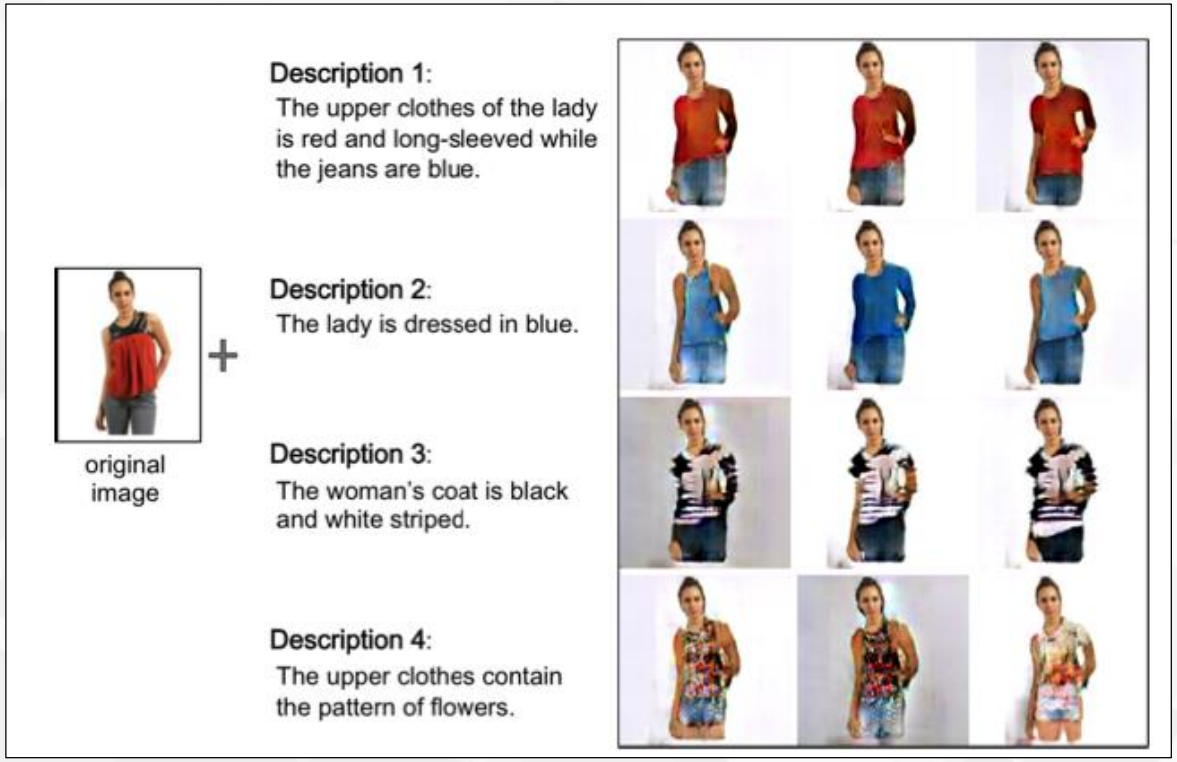
- More advanced multimodal techniques can be used to model and generate complex queries that are hard to express

### Motivation:

Multimodal query is easy and feasible for users to express their intention, since users are involved to refine the query.




[Wu et al.: The Fashion IQ Dataset: Retrieving Images by Combining Side Information and Relative Natural Language Feedback, CVPR 2021]

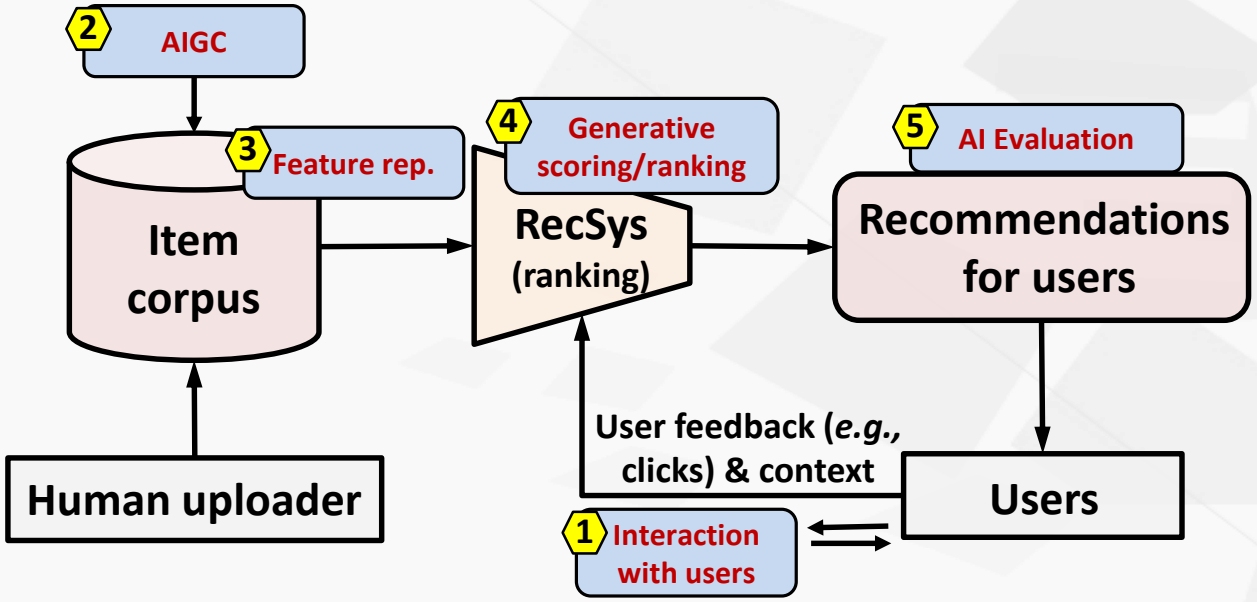



[Zhu et al.: Be Your Own Prada: Fashion Synthesis with Structural Coherence, ICCV 2017]

# D) Downstream Techniques & Applications:
## Multimodal Recommendation

- Trends in recommendation
  - Integrating collaborative filtering with content based
  - Rick interactions to better understand user intents and feedbacks
  - Towards personalized recommendation
  - More emphasis on discovery

- Towards generative recommendation




1. **Interaction with users**
   - LLMs for flexible user interactions and feedbacks
2. **Content generation**
   - Generative AI for content repurposing and creation
3. **Feature representation**
   - Feature encoder & data augmentation
4. **Generative scoring/ranking**
   - LLM-based recommendation
5. **Evaluation**
   - Agent for user simulation & LLMs for content auditing

# OUTLINE

- Trends in AI & MM Research
- History of in MM Research
- Key Topics of MM Research
- Criteria for Good MM research
- Course Outline and Instructions

# What is Multimedia Research?

- Deep Learning and LFM are here to stay
  - How to do accountable research in Big Data Deep Learning Era?
- The basics in multimedia research
  - Should involve more than one media
  - Consider using other forms of media and knowledge to supplement the low-semantic level of visual data
  - Solution should work in a noisy, ood & small sample environment
- But there should be more..
  - More than just fusion of multimodality
  - More than just accuracy..
  - More than a technical problem, should consider user and social issues

# Criteria for Good Multimedia Research

- Should involve multiple modalities
- Should go beyond accuracy, but consider problems of trust, robustness and scalability
- and, as appropriate, explainability, fairness & privacy
- Should support users, & offer insights & solutions

- The same for the writing good papers

# OUTLINE

- Trends in AI & MM Research
- History of in MM Research
- Key Topics of MM Research
- Criteria for Good MM research
- Course Outline and Instructions

# Objectives of this Course

To Introduce

- The rich and exciting areas of multimedia research
- The background & history to MM Research
- The recent advances in research and applications
- The future trends
- The criteria for a good multimedia research

Aim of Assessment:

- Aware of recent development: Paper reading, presentation and QA (22%)
- Aware of recent issues: writing of short idea/opinion articles (2x15 = 30%)
- High-quality paper writing: writing of paper (of publishable quality) (48%)

# Course Assessment

- **Presentation & QA** (22%)
  - Presentation of recent papers in assigned topics
  - Critiques of papers via QA and report
  - Presentation of issues/ideals of recent topics
- **Short Idea/Opinion Articles** (2x15 = 30%)
  - To write short article on 2 selected topics/ issues
  - The article should cover the background, issues, analysis and insights
  - The article should be within 4 pages
  - Deadlines: 16 Feb & 8 Mar
- **Brave-New Idea Paper** (48%)
  - To identify research with original ideas & vision that points to a novel direction
  - The papers should offer: (i) novel, exploratory solutions with sufficient evidence of proof-of-concept; (ii) visions describing a new or open problem in multimedia research; and (iii) a novel perspective on existing multimedia research
  - The paper should be within 5 pages (in ACM 2-column format)
  - Presentation & System Demo

# Course Schedule (1)

| Wk | Date | Lecture/Tutorial Topics | Remarks |
|---|---|---|---|
| 1. | 16 Jan | L1: Overview and Intro<br>(Background, history, featured-based vs. DL-based vs. pre-trained models, LMFM, safety & privacy, etc.) | Outline paper reading, presentation, article and BNI paper requirements |
| 2. | 23 Jan | L2: Visual Object Recognition, Detection and Vision-Language Models | Guest Lecture: Ji Wei<br>Release assignment of paper presentation list (26 Jan) |
| 3. | 30 Jan | L3: Semantic & Temporal Segmentation and Relation Grounding | Group 1 (3 Pax) |
| 4. | 6 Feb | L4: Cross-modal Alignments and Multimodal Scene Graphs | Group 2 (3 Pax) |
| 5. | 13 Feb | L5: Few Shot, Meta and Causal Learning | Group 3 (3 Pax)<br>Submission of Article 1<br>(16 Feb @1700) |
| 6. | 20 Feb | L6: Image/ Video QA, and Reasoning<br>Presentation of Idea Article 1 | Group 4 (2 Pax)<br>Present Article 1 (2 Pax)<br>Meeting: Idea for BNI Papers |
| | 27 Feb | **Recess Week** | |

# Course Schedule (2)

| | | | |
|---|---|---|---|
| 7. | 5 Mar | L7: Condition-based Diffusion Models for MM Generation | Guest Lecture: Jin Zhe<br>Group 5 (2 Pax)<br>Submission of Article 2<br>(8 Mar @1700) |
| 8. | 12 Mar | L8: Large Multimodal Foundation Model (LMFM) | Group 6 (3 Pax) |
| 9. | 19 Mar | L9: NExT-GPT & Multimodal Dialogues | Guest Lecture: Wu SQ<br>Present Article 2 (2 Pax) |
| 10. | 26 Mar | L10: Responsible AI: Trust, Safety, Privacy & Biased in MM | Group 7 (3 Pax) |
| 11. | 2 Apr | L11: Multimodal Event (& Fashion) Detection & Forecasting | Lecture by: Ma Yunshan<br>Group 8 (2 Pax)<br>Submission of BNI Papers<br>(5 Apr @1700) |
| 12. | 9 Apr | L12: MM Recommendation<br>Presentation of BNI Papers -1 | Group 9 (2 Pax)<br>Present BNI Papers |
| 13. | 16 Apr | L13: Presentation of BNI Papers -2 | Present BNI Papers |
| 14. | 23 Apr | **End of Course** | To return all graded course works & assignments. |

# Next Week

## L2: Visual Object Recognition, Detection & Vision-Language Model:

(to be presented by Ji Wei)

### P2-1: Image Classification:

(SOTA): Swin transformer: Hierarchical vision transformer ... ICCV 2021 (Best Paper)

(Must-Read) Deep Residual Learning for Image Recognition. CVPR 2016.

(To-Read) CoAtNet: Marrying Convolution & Attention for All Data Sizes. NeurIPS 2021.

### P2-2: Object Detection:

(SOTA) Swin Transformer V2: Scaling Up Capacity and Resolution. CVPR2022.

(To-Read): End-to-End Object Detection with Transformers. ECCV 2020

(Must-Read): Faster R-CNN: … PAMI 2016.

### P2-3: Vision-Language Models:

(SOTA) Multi-Grained Vision Language Pre-Training: ... ArXiv 2021.

Must-Read) VisualBERT: A Simple & Performant Baseline for V&L. arXiv 2019.

(Must-Read) Multimodal Convolutional Neural Networks for Matching Image and Sentence. ICCV 2015.

# Any Questions?

- From Week 3 onwards, the selected topics will be presented by you all
- I will get guest lectures for some of the topics
- For each lecture: I will give 10-15 mins overview of the topic, optionally a guest lecture, and follow by the presentation of up to 45 mins by a group of 2 or 3 students

- I need 3 volunteers for presentation on Week 3
- I will (randomly) assign presenters for topics from Lecture 4.