Large Vision Models
Importance, Real-World Demand, and Challenges

By
Dr. Mudassar Raza
Professor
Namal University Mianwali

AI and High Performance Computing for Self Employment Track

Dr. Mudassar Raza

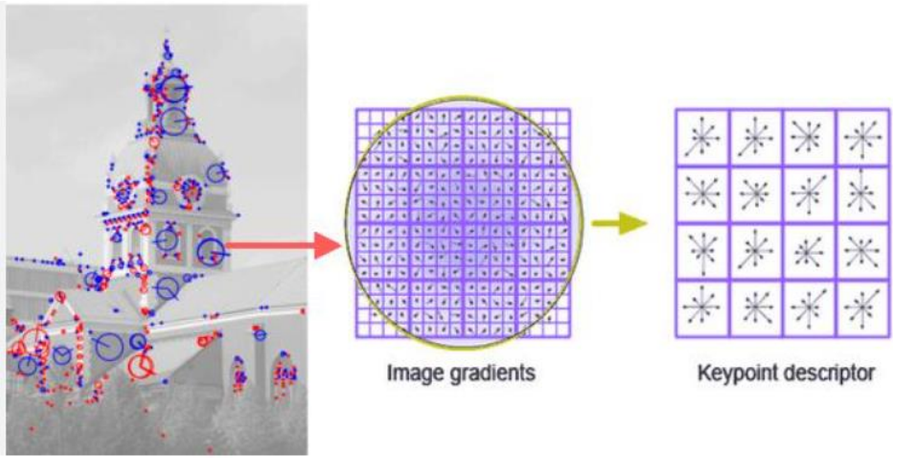
- Senior Member IEEE
- Chair Publications, IEEE Islamabad Section
- Academic Editor PLoS One Journal
- Member National Standards Committee
- Professor

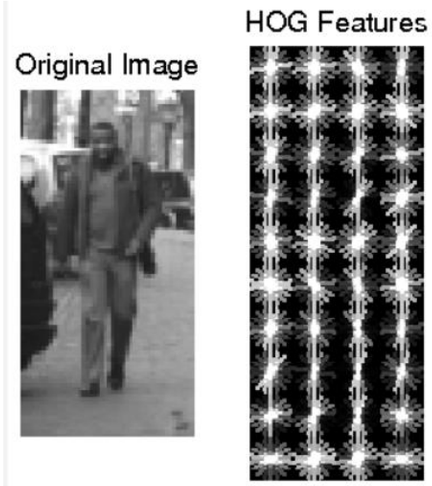
- Over **20 years** of experience (teaching + research)
- Supervised and co-supervised over **50 MS theses**.
- Supervised **3 PhD Theses**
- Supervised more than **100 R&D projects** of undergraduate students.
- **Publications: 150+, Cumulative Impact Factor: 220+, Total Citations: 6650+, H-Index: 45, I-10 Index: 101**
- **Listed as a World top 2% scientist by Elsevier (Published in October 2023):** https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/6
- **Ranked #11 (Computer Science), among the top 3% in Pakistan**
- https://www.adscientificindex.com/scientist/mudassar-raza/419498
- **Research Interest Score is higher than 98% of ResearchGate members**
- **Ranked 15th among the best computer scientists of Pakistan** according to research.com, https://research.com/u/mudassar-raza
- **Highly Ranked Scholar** - Prior Five Years, Ranked top 0.19 percent Scientist worldwide, https://scholargps.com/scholars/19651897791773/mudassar-raza

- **Awardee of the National Youth Award 2008 by the Prime Minister of Pakistan**

Introduction

- Early computer vision relied on handcrafted features (SIFT, HOG, Haar).

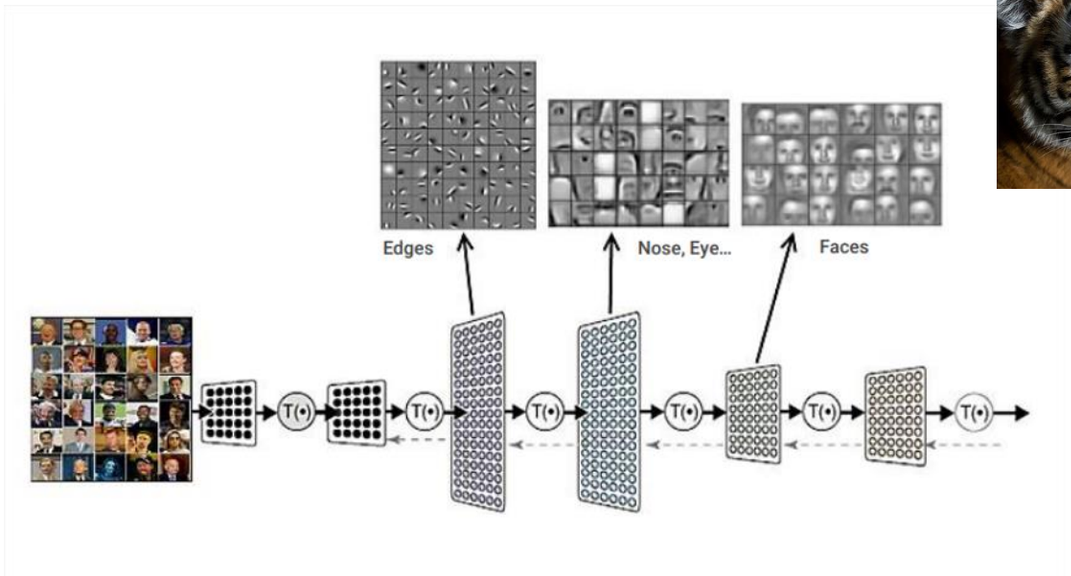


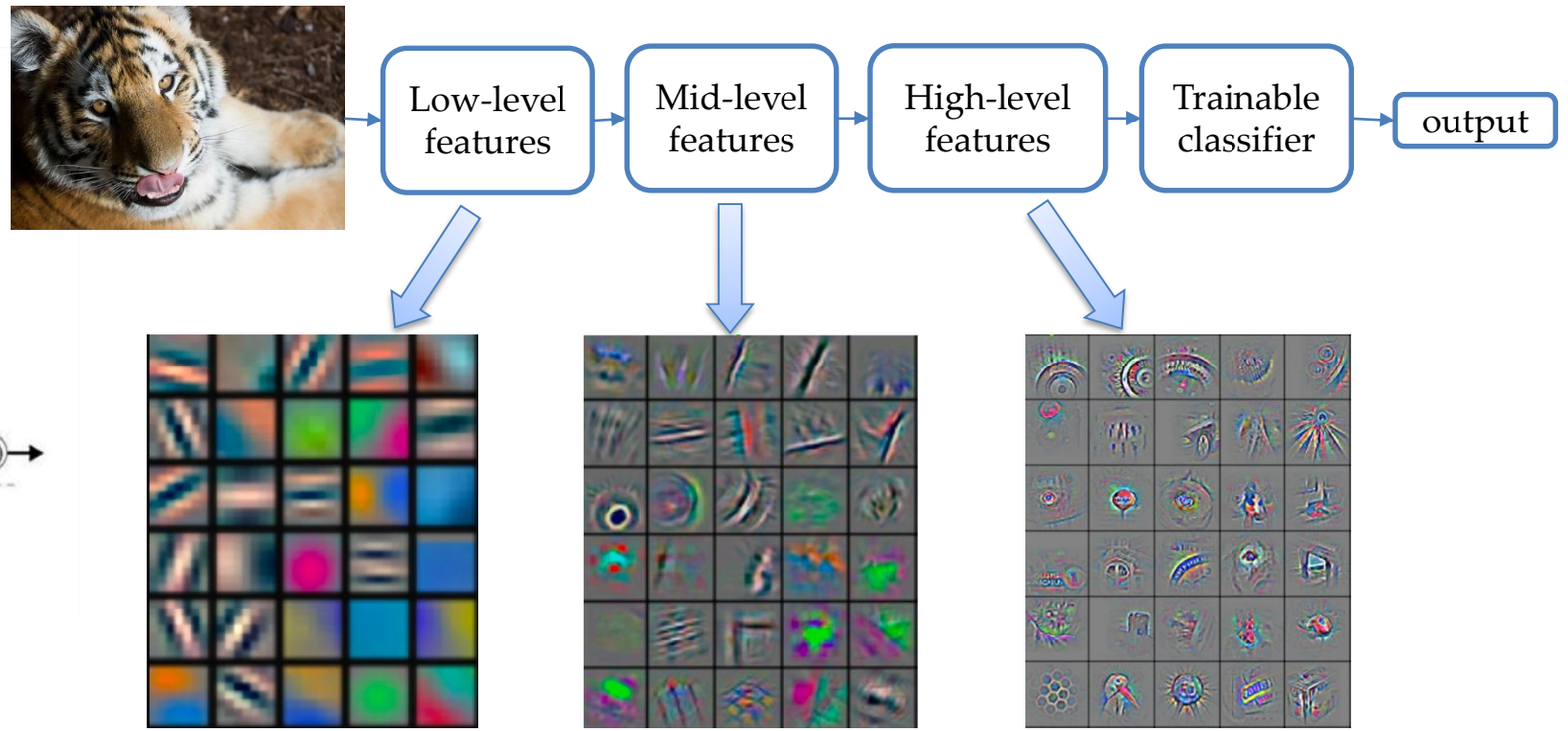


- These features struggled with complex real-world data (illumination, scale, occlusion).

Introduction

- **Deep learning transformed vision by learning hierarchical representations directly from data. Sets the stage for the rise of CNNs.**





Deep Learning Chronology

1940s–1950s: Foundations

1943 – McCulloch & Pitts introduce the **artificial neuron model**.
1950 – Alan Turing publishes **"Computing Machinery and Intelligence"** (Turing Test).
1958 – Frank Rosenblatt develops the **Perceptron** (first trainable neural network).

1960s–1970s: Early Neural Network Research

1960 – Widrow & Hoff introduce **ADALINE/MADALINE** for adaptive linear networks.
1969 – Minsky & Papert publish **"Perceptrons"**, proving limitations (can't solve XOR) → leads to an **AI Winter**.

1980s: Neural Network Revival

- **1980 –** Fukushima proposes **Neocognitron** (inspiration for CNNs).
- **1986 –** Rumelhart, Hinton & Williams popularize **Backpropagation** → training multilayer networks becomes possible.
- **1989 –** Yann LeCun applies **backpropagation to convolutional networks** for digit recognition (precursor to modern CNNs).

Deep Learning Chronology

1990s: Advances in Training & Applications

1995 – Vapnik introduces **Support Vector Machines (SVMs)**, dominating ML.
1997 – Sepp Hochreiter & Jürgen Schmidhuber introduce **Long Short-Term Memory (LSTM)** networks for sequence learning.
1998 – LeCun develops **LeNet-5 CNN**, successful in digit recognition (MNIST).

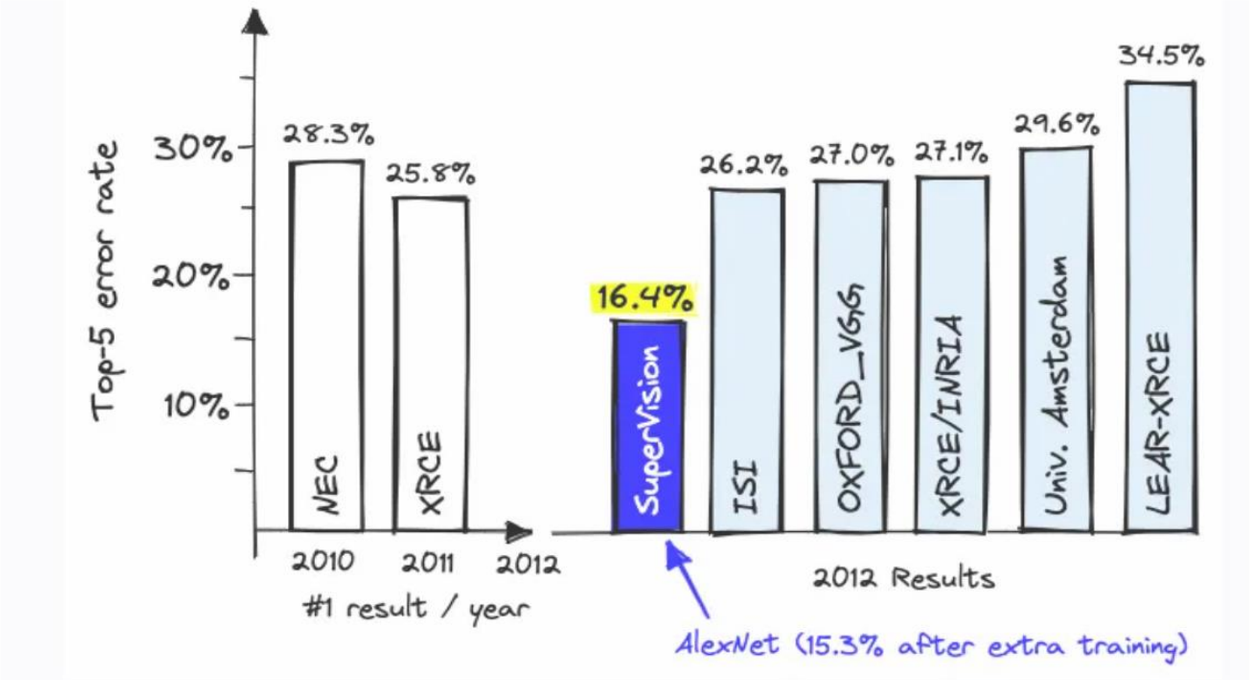
2000s: Slow but Steady Progress

2006 – Geoffrey Hinton introduces **Deep Belief Networks (DBNs)** → starts the modern **Deep Learning era**.
2009 – Fei-Fei Li launches **ImageNet dataset**, a large-scale benchmark for computer vision.

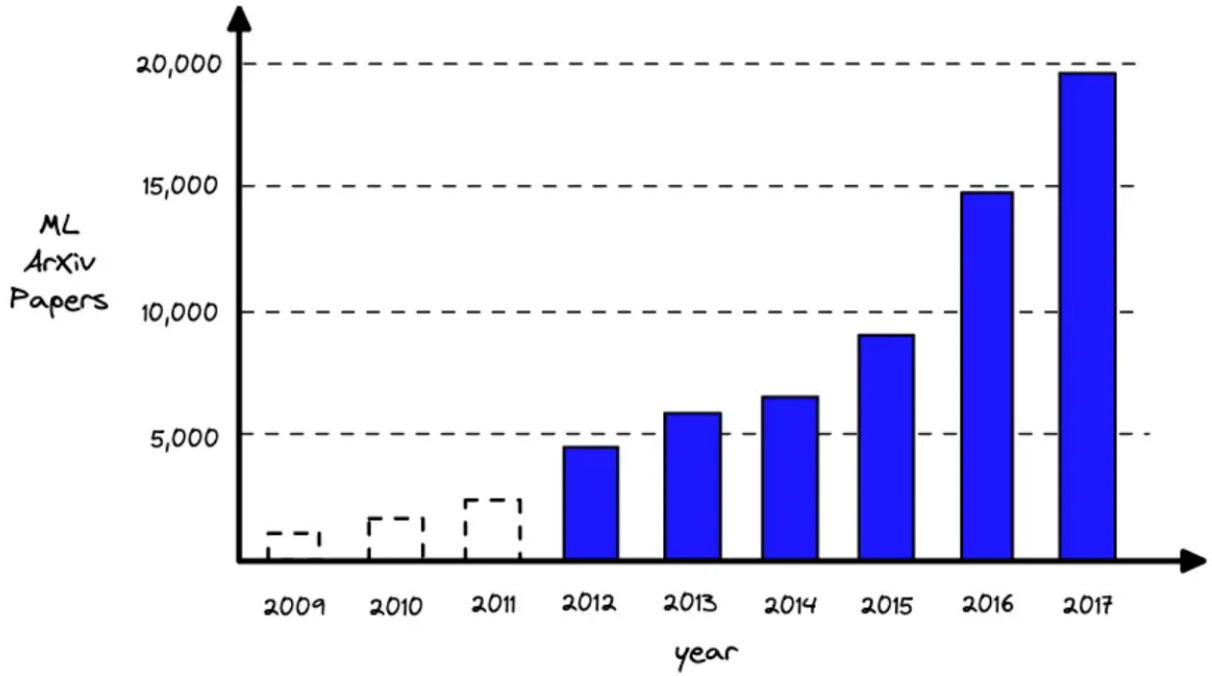
Deep Learning Chronology

2010s: Breakthroughs

2012 – AlexNet (Krizhevsky, Hinton, Sutskever) wins ImageNet challenge with CNNs, sparking the **deep learning revolution**



The best ImageNet challenge results in 2010 and 2011, compared against all results in 2012, including AlexNet [2]



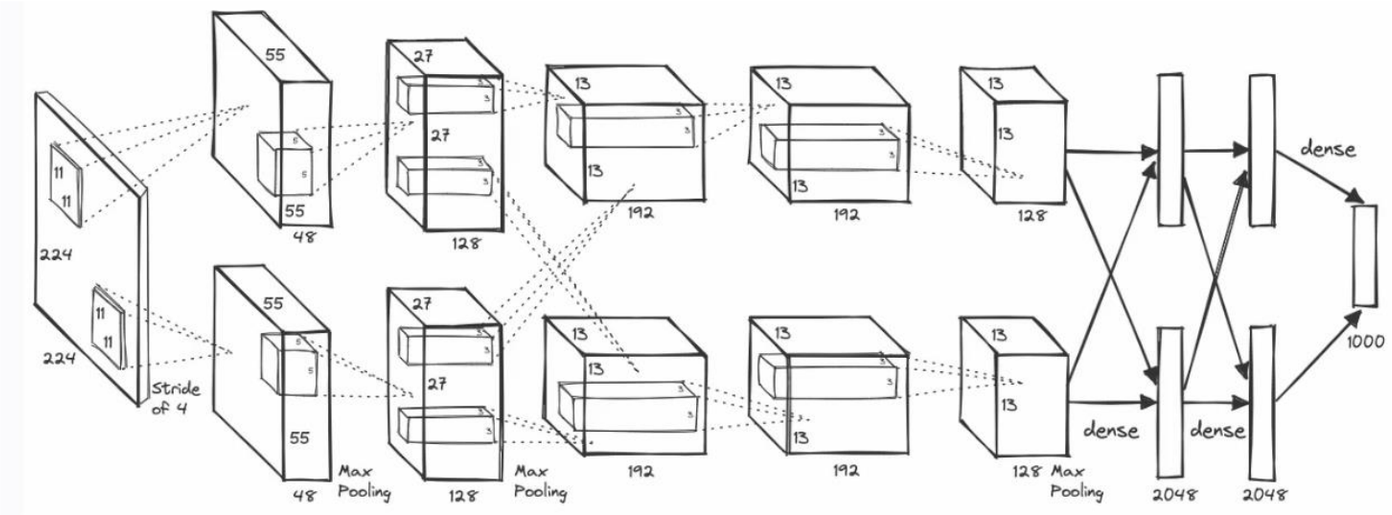
ImageNet **L**arge **S**cale **V**isual **R**ecognition **C**hallenge (ILSVRC)

- Used deep CNNs + GPUs for acceleration.
- Pioneered the deep learning revolution in computer vision.
- ImageNet became the "**CERN moment**" for AI vision.

Deep Learning Chronology

2010s: Breakthroughs

2012 – AlexNet (Krizhevsky, Hinton, Sutskever) wins ImageNet challenge with CNNs, sparking the **deep learning revolution**



Network architecture of AlexNet [1].

AlexNet was distributed across two GPUs. Each GPU handled one-half of AlexNet. The two halves would communicate in specific layers to ensure they were not training two separate models.

Today's deep learning revolution traces back to the 30th of September, 2012. On this day, a **C**onvolutional **N**eural **N**etwork (CNN) called AlexNet won the ImageNet 2012 challenge [1].
AlexNet didn't just win;
it *dominated*.

Deep Learning Chronology

2010s: Breakthroughs

2012 – AlexNet (Krizhevsky, Hinton, Sutskever) wins ImageNet challenge with CNNs, sparking the **deep learning revolution**

The **founder and main author of MatConvNet (2014) is Andrea Vedaldi, a computer vision researcher and professor at the** University of Oxford.

Deep Learning Chronology

2010s: Breakthroughs & Explosion

- **2012 – AlexNet (Krizhevsky, Hinton, Sutskever)** wins ImageNet challenge with CNNs, sparking the **deep learning revolution**.
- **2013 –** Word2Vec (Mikolov et al.) introduces distributed word embeddings for NLP.
- **2014 –** Ian Goodfellow proposes **Generative Adversarial Networks (GANs)**.
- **2015 –** ResNet (He et al.) introduces **skip connections**, winning ImageNet with human-level accuracy.
- **2016 – AlphaGo (DeepMind)** defeats world champion in Go using deep reinforcement learning.
- **2017 –** Vaswani et al. introduce the **Transformer architecture** → revolutionizes NLP.
- **2018 –** Google releases **BERT**, achieving state-of-the-art NLP results.
- **2019 –** OpenAI releases **GPT-2**, showing large language models' potential.

Deep Learning Chronology

2020s: Era of Large Models

2020 – OpenAI releases **GPT-3 (175B parameters)** → milestone in generative AI.
2021 – DALL·E, CLIP, Codex introduced → multimodal & code generation breakthroughs.
2022 – Stable Diffusion & MidJourney popularize **AI image generation**; **ChatGPT (based on GPT-3.5)** launches.
2023 – OpenAI releases **GPT-4** with multimodal capabilities; **PaLM, LLaMA, Claude** push open-source & competition.
2024–2025 – Rise of **Large Vision Models (LVMs), agentic AI,** and **multimodal deep learning** integrating vision, text, speech, and actions.

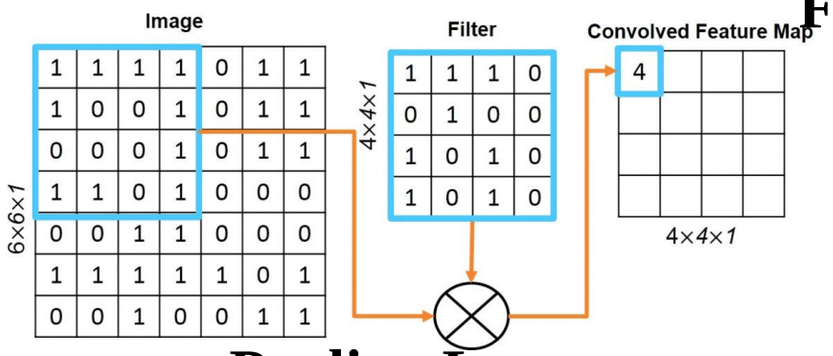
Why CNNs Revolutionized Vision

- Learn features automatically instead of handcrafted.
- Invariance to many features like scale, rotation, etc.
- Layers build progressively:
  - Edges → Shapes → Objects.
- Enabled breakthroughs in
  - classification,
  - detection,
  - segmentation.

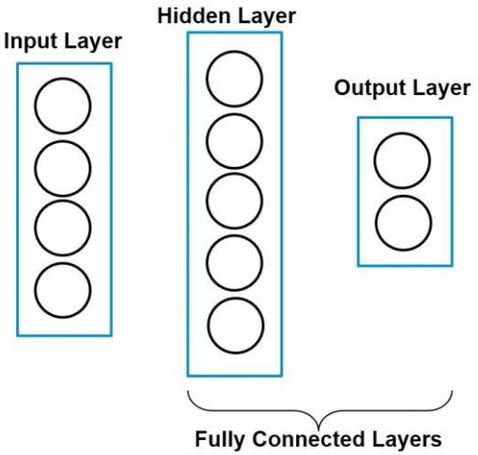
Building Blocks of Deep Learning Models

- Neurons
- Activation functions (ReLU, sigmoid)
- Convolutional layers (feature extraction)
- Pooling layers (downsampling)
- Fully connected layers (decision making).
- Backpropagation (Gradient Descent)

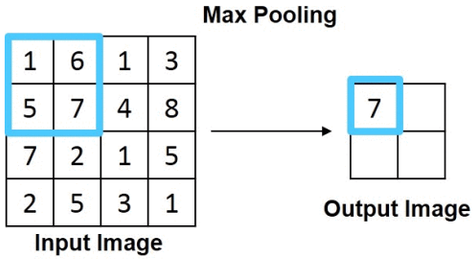
Convolutional Layer



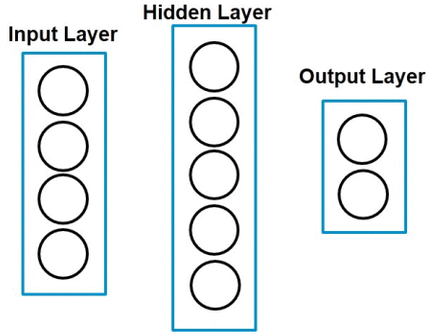
Fully connected Layer



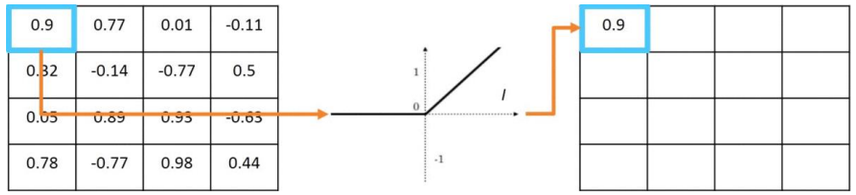
Pooling Layer



Dropout Layer



Rectified Linear Units (ReLUs) Layer

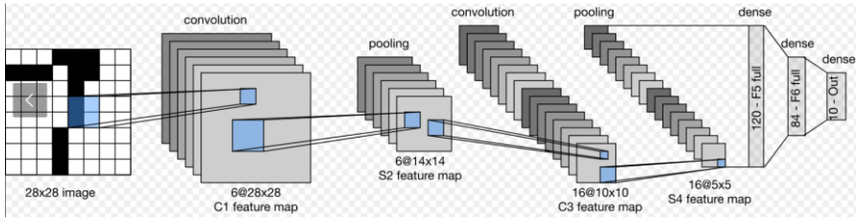


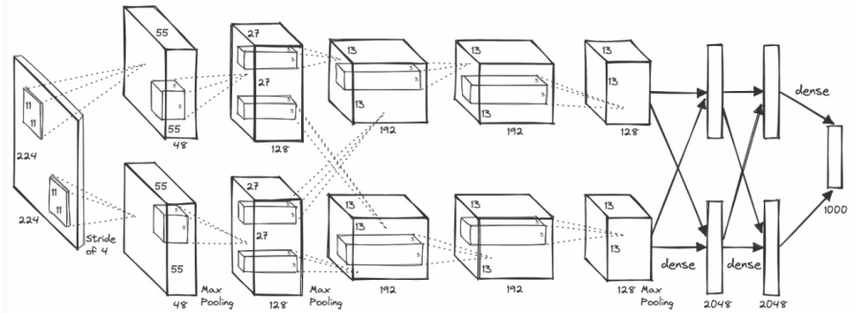
Core Frameworks

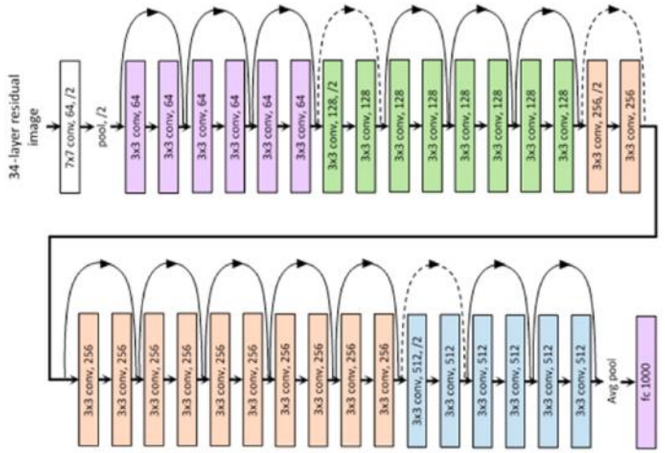
- TensorFlow (Google), PyTorch (Meta), Keras.
- Enabled global adoption in academia and industry.

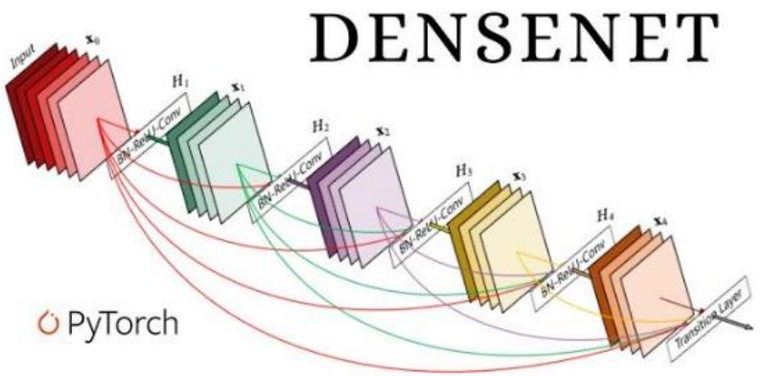
CNN Architectures

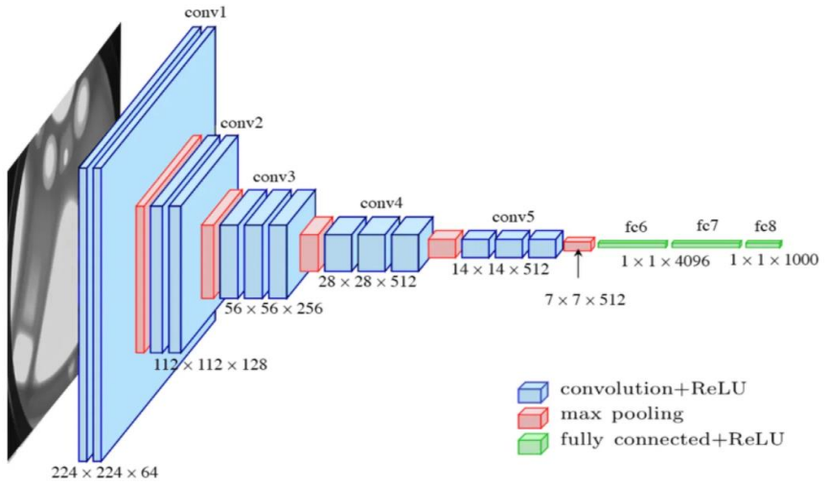
- LeNet (1998) – Early CNN for digit recognition.
- AlexNet (2012) – Breakthrough on ImageNet.
- VGG (2014) – Deeper CNNs with small kernels.
- ResNet (2015) – Skip connections, 100+ layers.
- DenseNet (2017) – Dense connectivity.
- EfficientNet (2019) – Compound scaling.





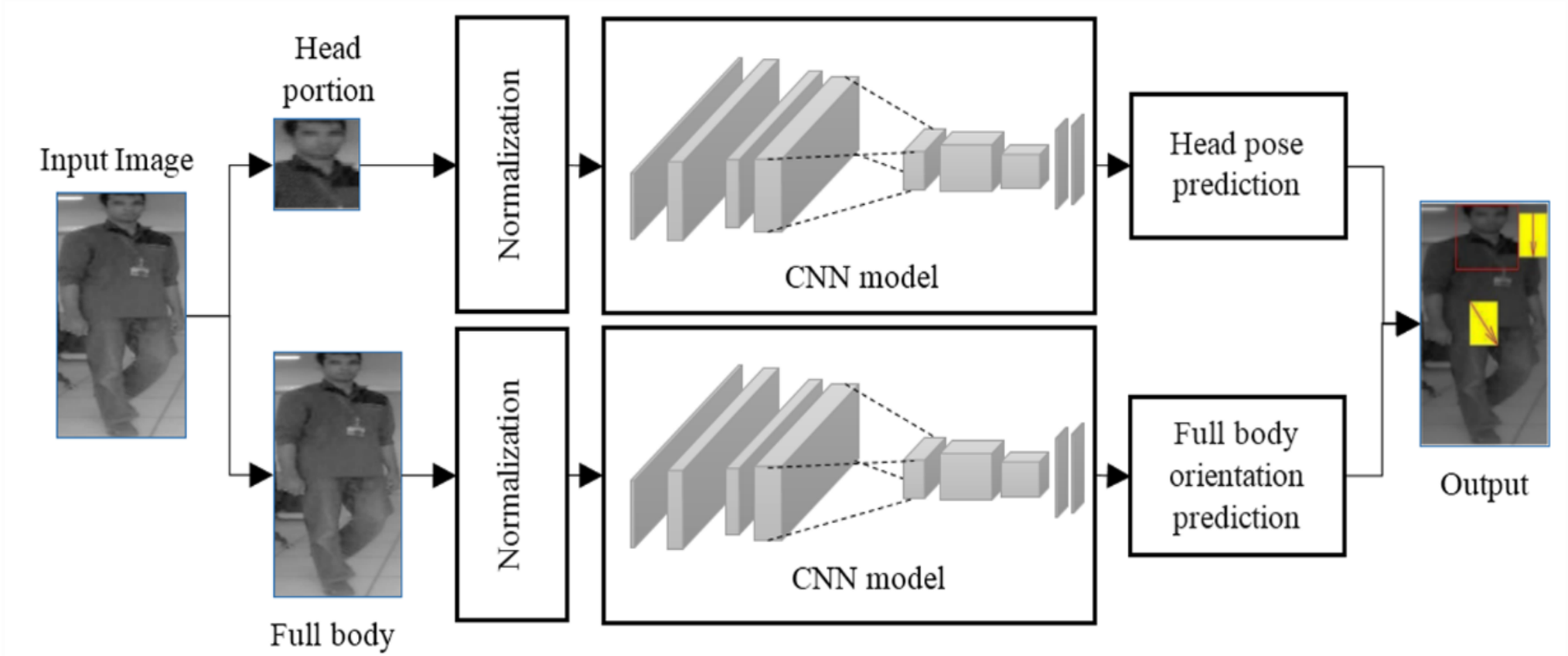






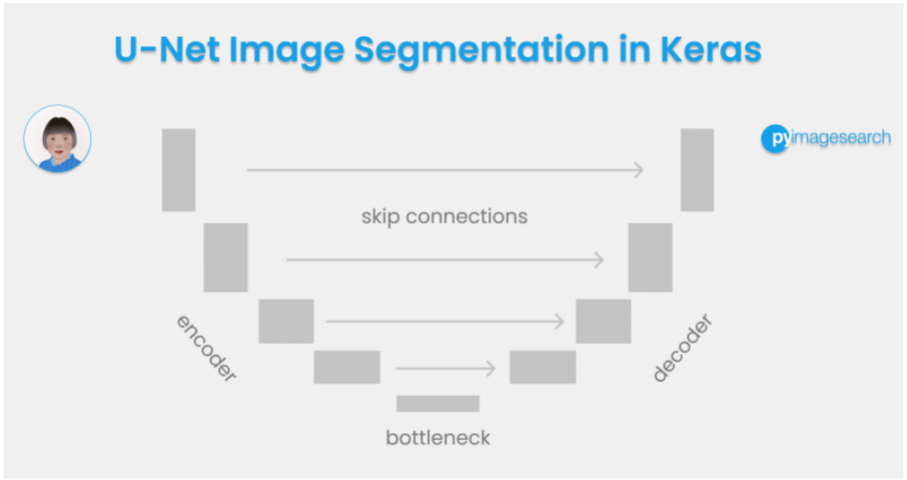
Uses of Vision Models

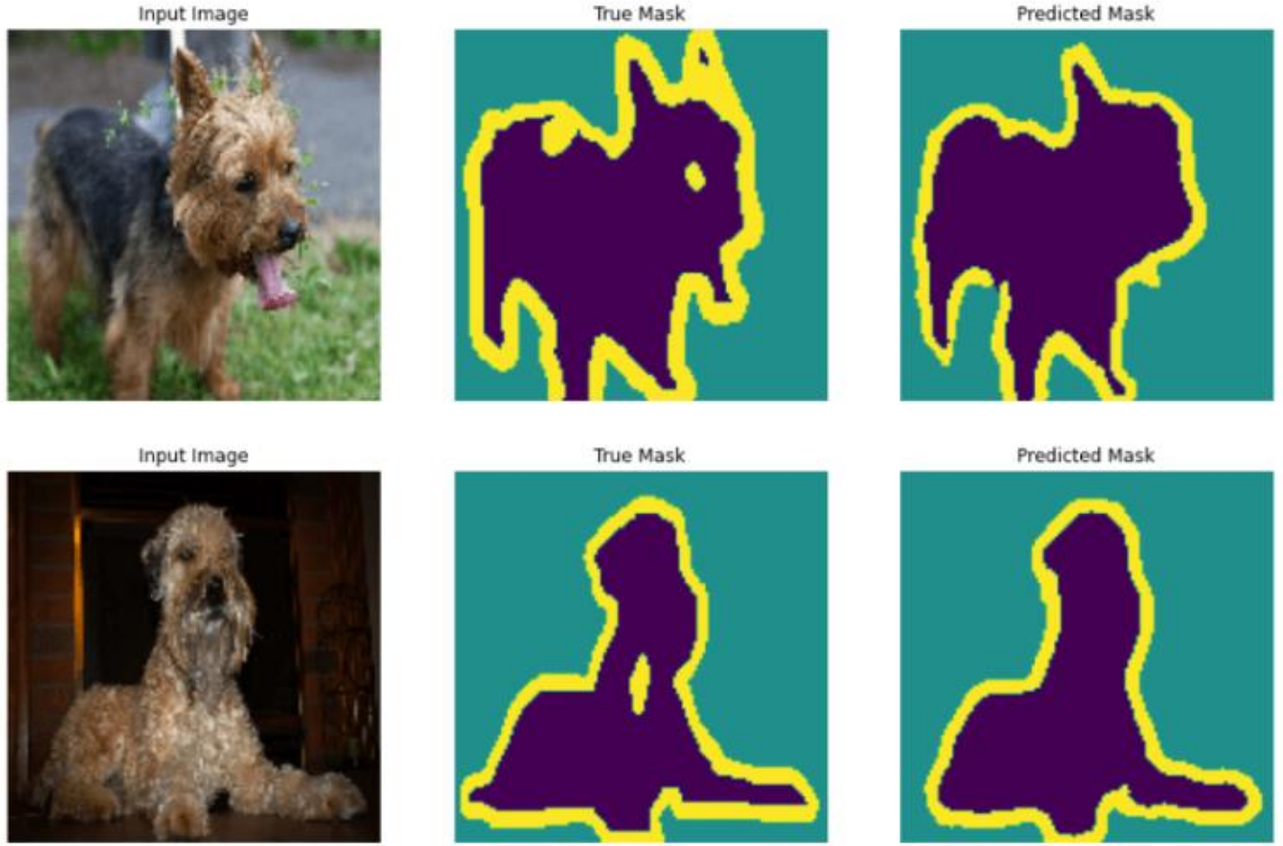
- **Image Classification**



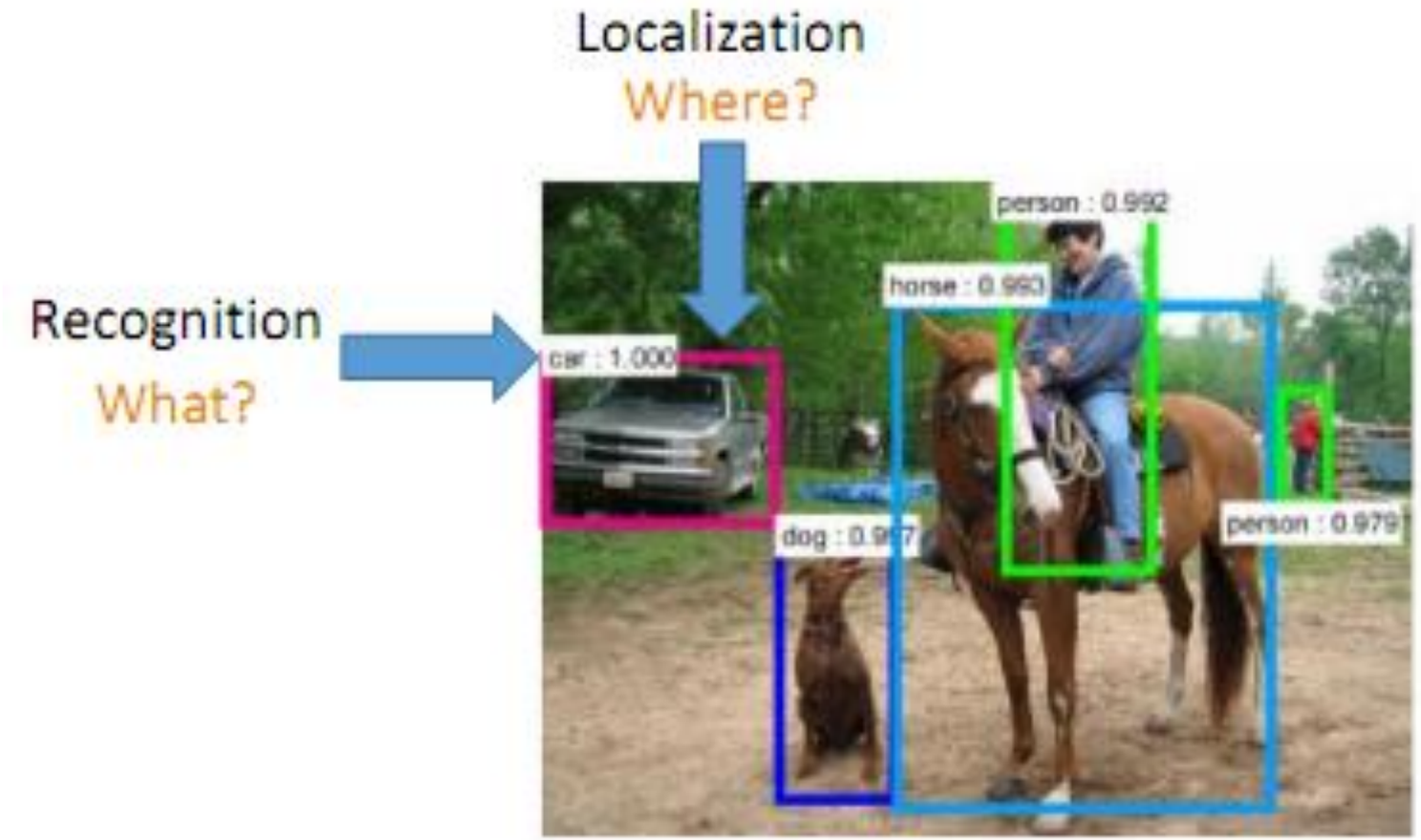
Uses of Vision Models

- Image Semantic Segmentation
- UNET



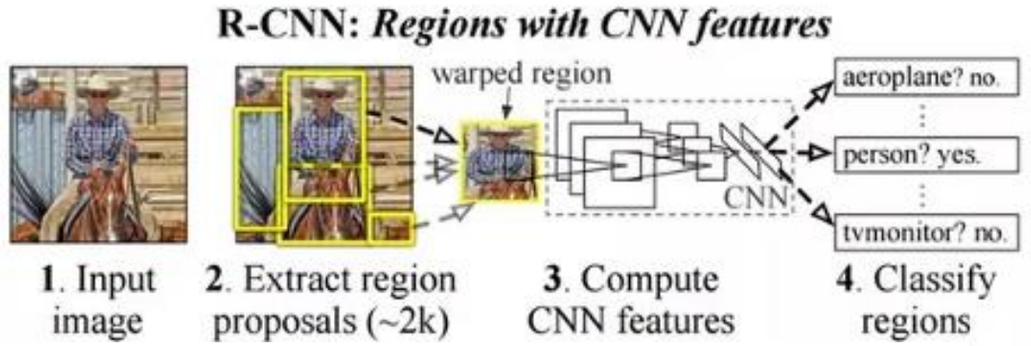


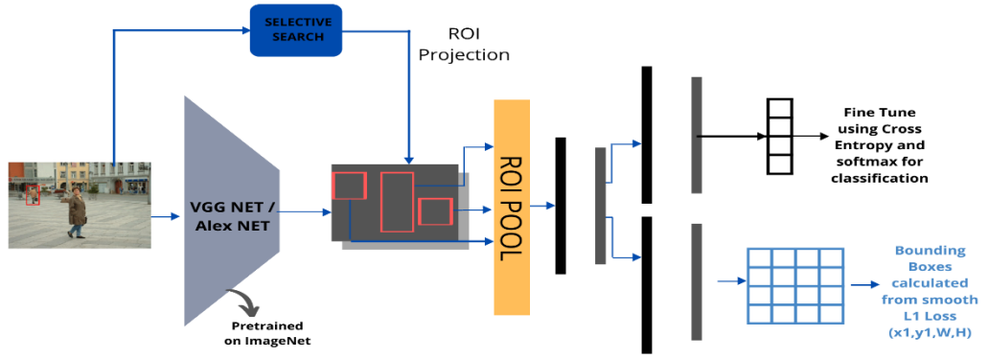
Object Detection

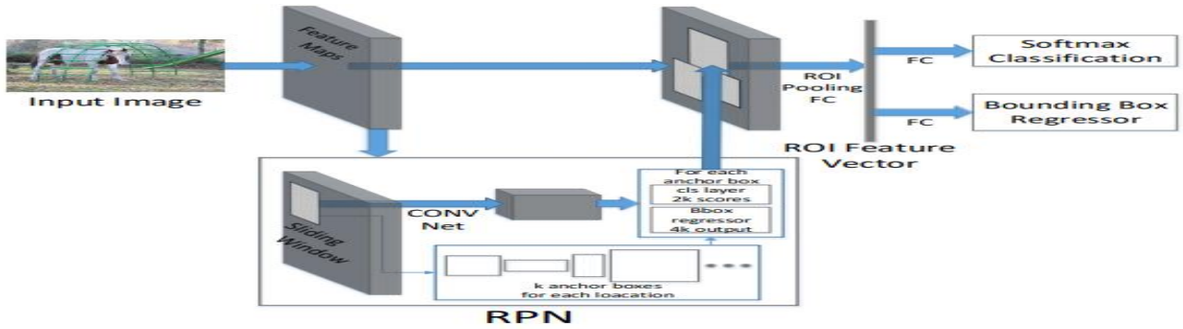


Uses of Vision Models

- RCNN
- Fast RCNN
- Faster RCNN







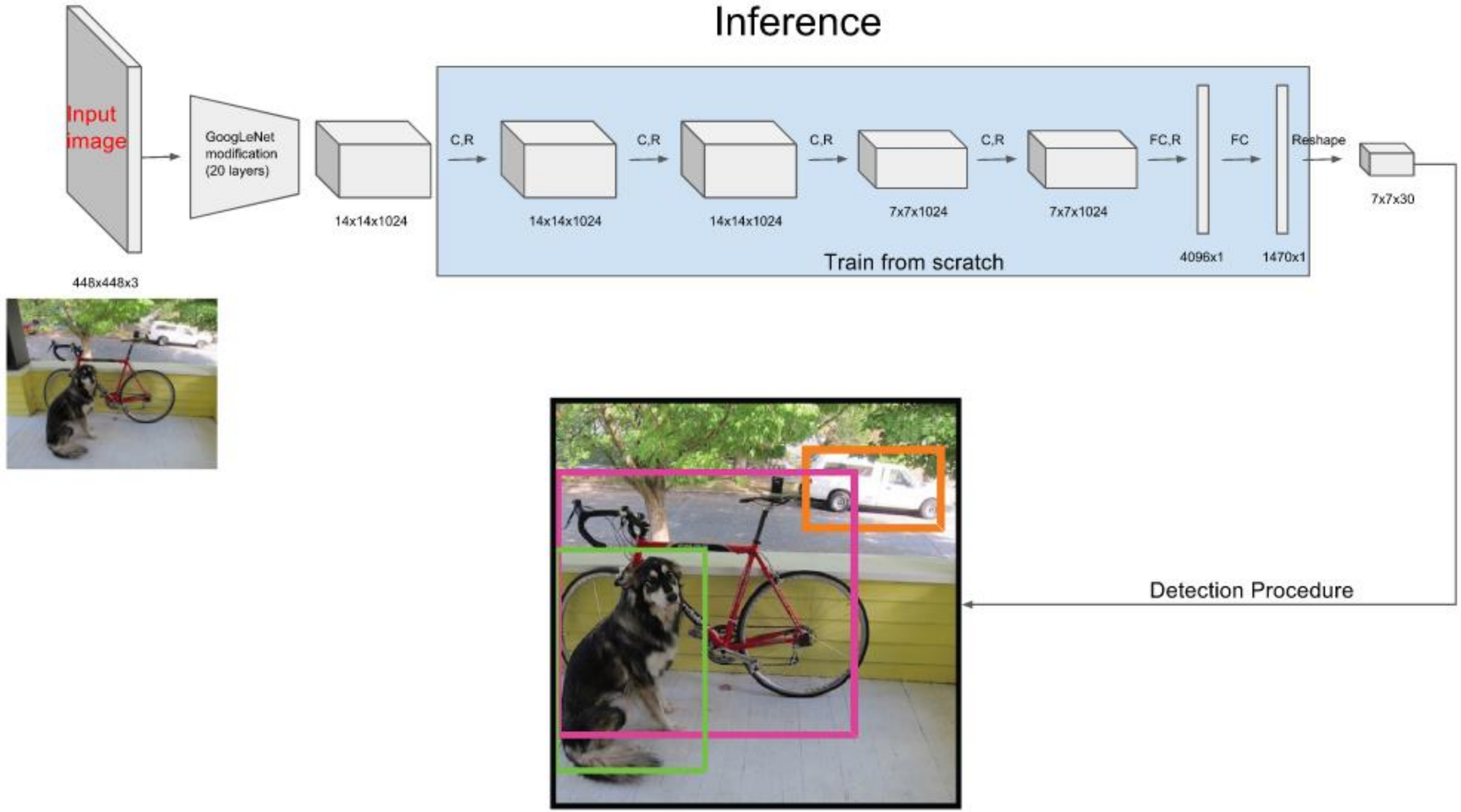
YOLO

You Only Look Once

(A Modern Object Detection Algorithm)

Note: The slides in this presentation are copied from different existing sources. This presentation is just for education purpose

Uses of Vision Models

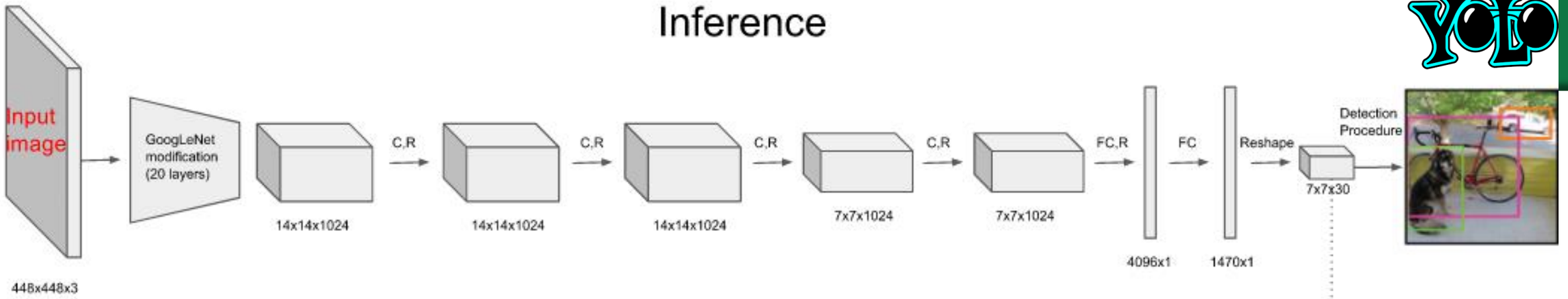


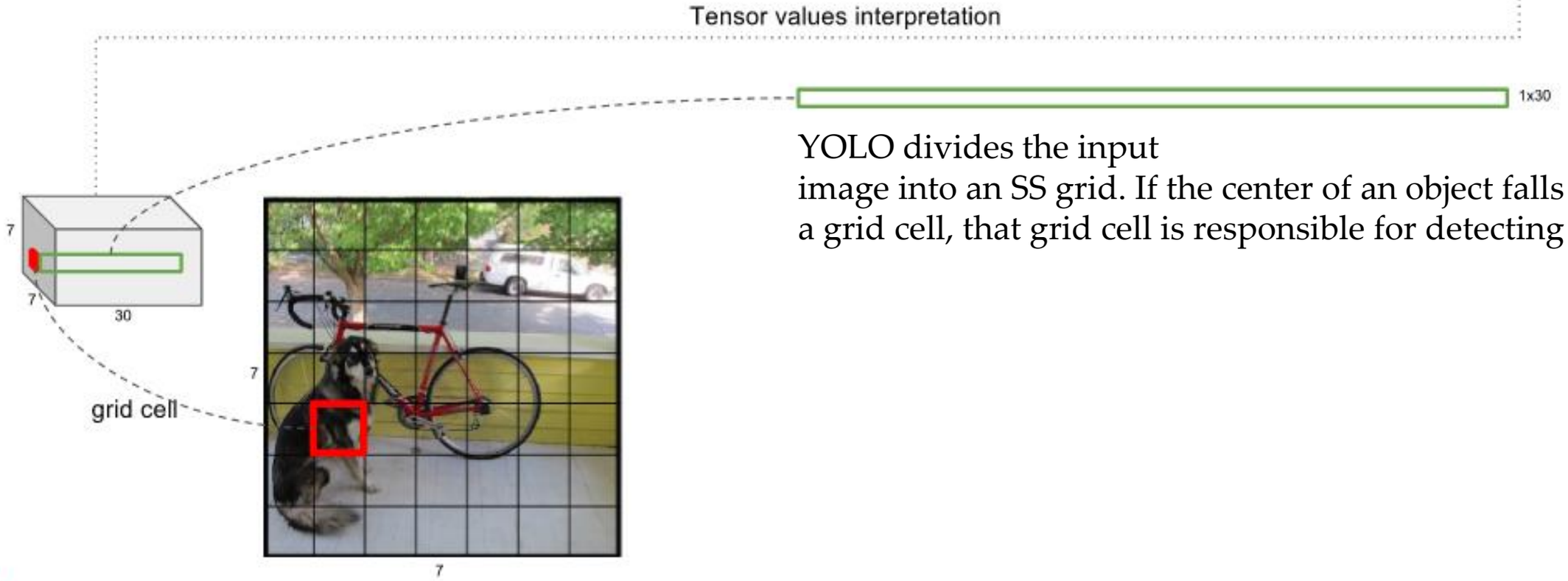
deepsystems.io

YOLO


www.BANDICAM.com
RUNNING...
Stop
Deploy
Running on: CPU
Configuration
Select Video Input Source
Upload Video File
Dahua Camera (RTSP)
Upload Cricket Video
Drag and drop file here
Limit 200MB per file • MP4, AVI, MOV, MPEG4
Browse files
100.mp4
5.0MB
{
"Filename" : "100.mp4"
IP PTZ Camera
Crease Ball Detection
Real-time Detection Performance
clideo.com

Inference

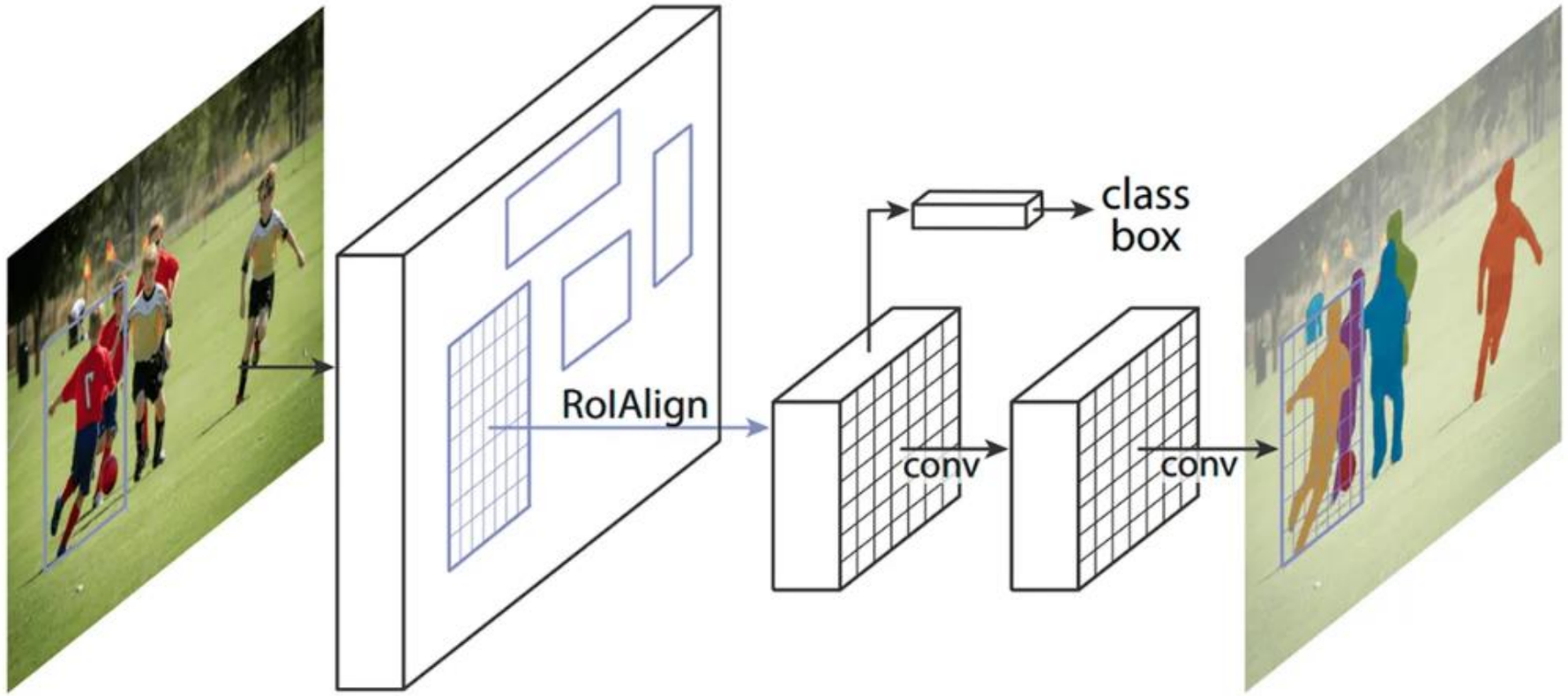




YOLO divides the input
image into an SS grid. If the center of an object falls into
a grid cell, that grid cell is responsible for detecting that object.

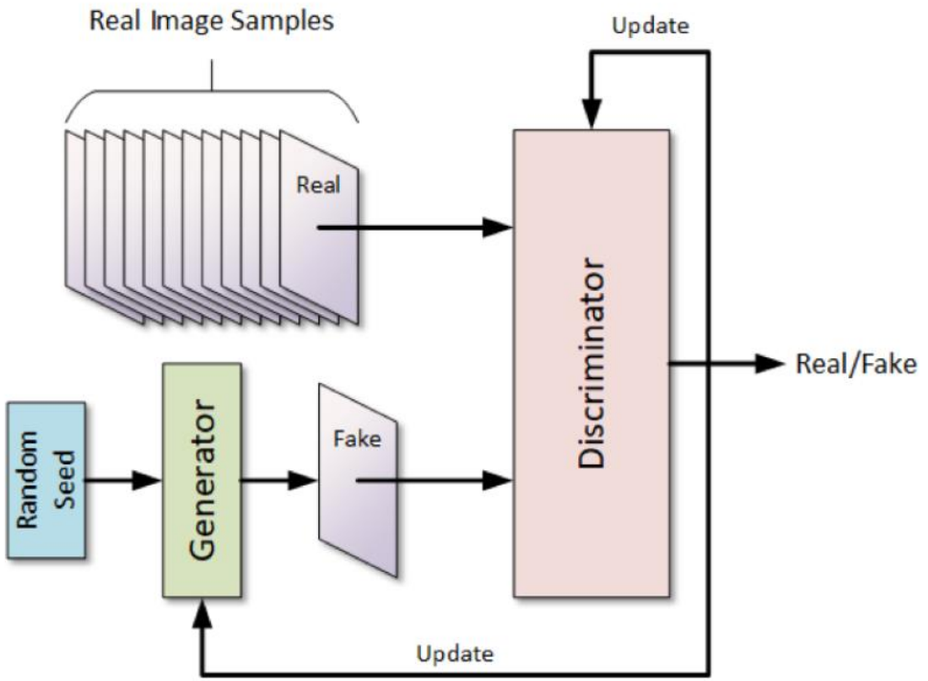
Uses of Vision Models

- Instance Segmentation
- Masked RCNN

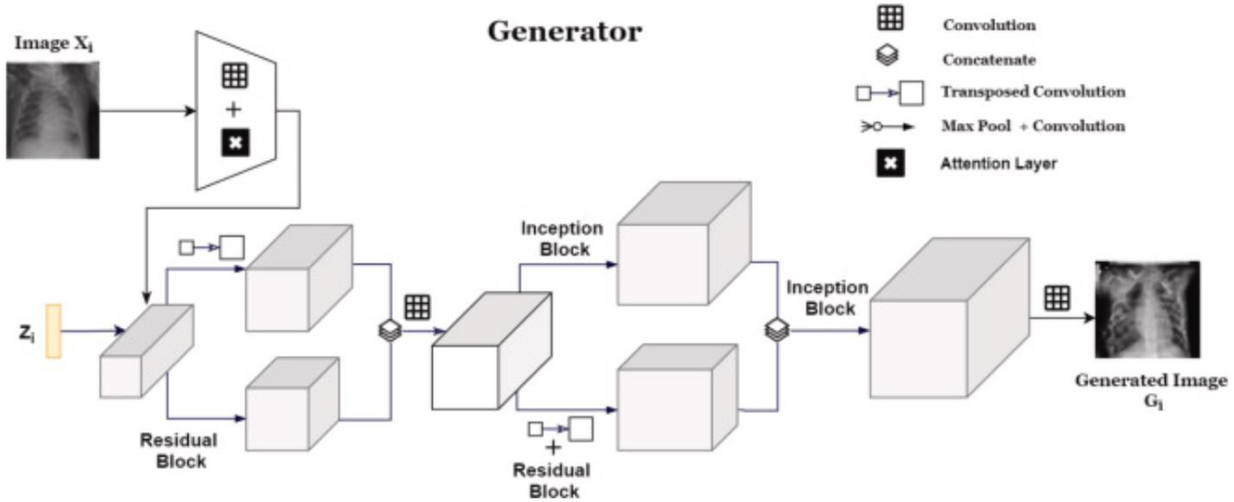


Uses of Vision Models

- **Generative Vision Models**
- GANs
- Generate realistic synthetic
- Used in art, design, data augmentation.



Basic GAN architecture. The generator creates an image from a random seed. The discriminator evaluates the image based on its training to see if it can tell real from fake. The result goes back to the generator and discriminator so that they improve. Source: Bryon Moyer/Semiconductor Engineering



Uses of Vision Models

- **Vision Language Models or Large Vision Models (LVMs)**

- **Large Vision Models = Foundation-scale models for vision.**
- **Trained on billions of images with billions of parameters.**
- **Analogy: LLMs in language → LVMs in vision.**

Vision Language Models (VLMs) represent one of the most exciting frontiers in artificial intelligence, combining the power of computer vision and natural language processing. These models can understand and generate content that involves both visual and textual information, opening up countless possibilities for AI applications.

What are Vision Language Models?

- **Vision Language Models (VLMs) are artificial intelligence systems that can process and understand both visual content (images, videos) and textual content simultaneously.**
- **Unlike traditional AI models that handle only one type of data, VLMs can:**
  - Analyze images and describe them in natural language
  - Answer questions about visual content
  - Generate images from text descriptions
  - Perform visual reasoning tasks
  - Translate between visual and textual representations

What are Vision Language Models?

- **Key Characteristics:**
- **Multimodal Processing:** Handle multiple types of input (vision + language)
- **Cross-modal Understanding:** Connect visual concepts with linguistic descriptions
- **Bidirectional Communication:** Can go from text to image or image to text
- **Contextual Awareness:** Understand relationships between visual and textualelements

What are Vision Language Models?

- **Why Do We Need VLMs?**
- **Bridging the Communication Gap (Visual+Linguistics)**
- Enhanced AI Capabilities
- Real-World Applications
- Accessibility and Inclusion
  - Describe images for visually impaired users
  - Generate visual content for text-based descriptions
  - Provide alternative ways to interact with technology

What are Vision Language Models?

- **Why Do We Need VLMs?**
- **Bridging the Communication Gap (Visual+Linguistics)**
- Enhanced AI Capabilities
- Real-World Applications
- Accessibility and Inclusion
  - Describe images for visually impaired users
  - Generate visual content for text-based descriptions
  - Provide alternative ways to interact with technology

What are Vision Language Models?

Key Processing Steps:

[Image] → Vision Encoder → Visual Features

↓

Multimodal Fusion ← Text Features

↓

[Text] ← Language Decoder ← Shared Representation

What are Vision Language Models?

How VLMs Work

- **Input Processing**
  - Visual input is processed through a vision encoder
  - Text input is processed through a language encoder
  - Both are converted to numerical representations (embeddings)

Vision Encoder

- **Purpose:** Convert images to numerical representations
- **Common Architectures:**
  - Convolutional Neural Networks (CNNs)
  - Vision Transformers (ViTs)

Language Encoder/Decoder

- **Purpose:** Process and generate text
- **Common Architectures:**
  - Transformer models (BERT, GPT variants)
  - LSTM/GRU networks (less common in modern VLMs)

What are Vision Language Models?

How VLMs Work

- **Feature Alignment**
  - Visual and textual features are aligned in a shared embedding space
  - The model learns to associate visual patterns with textual concepts

What are Vision Language Models?

How VLMs Work

- **Multimodal Reasoning**
  - The model processes both modalities together
  - Cross-attention mechanisms help relate visual and textual elements

Fusion Module

- **Purpose**: Combine visual and textual information
- **Methods**:
  - Cross-attention mechanisms
  - Concatenation and projection layers
  - Gated fusion techniques

What are Vision Language Models?

How VLMs Work

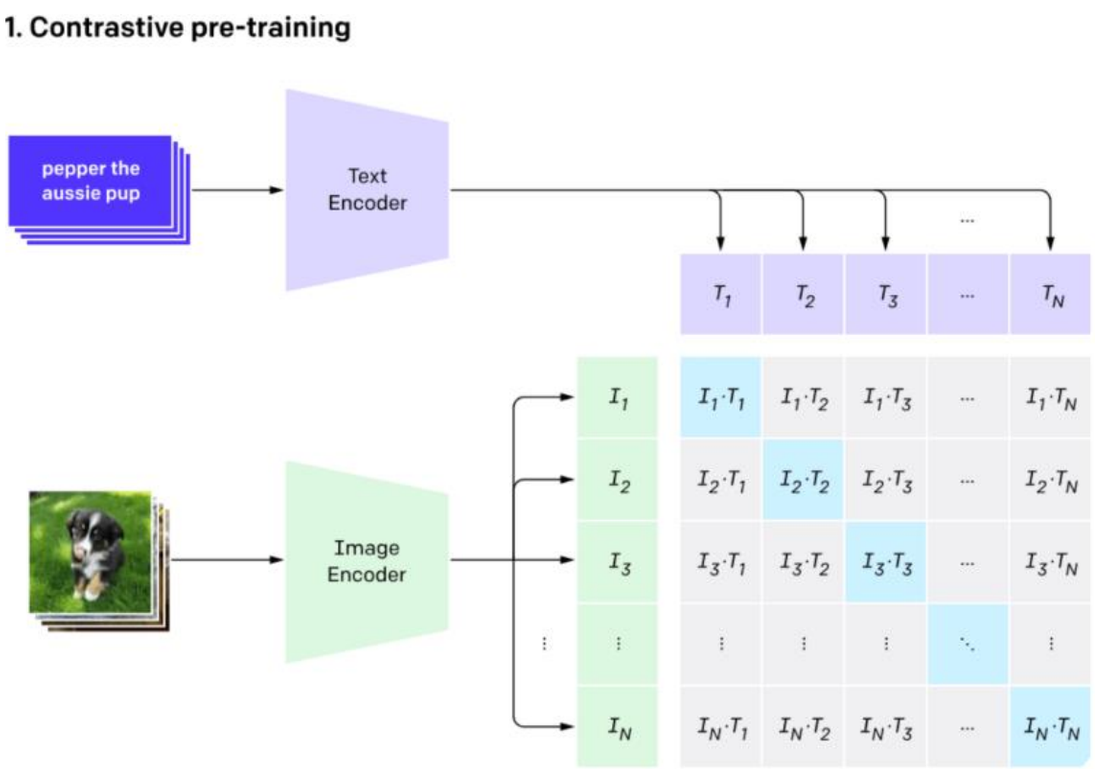
- **Output Generation**
  - Based on the task, the model generates appropriate outputs
  - Could be text descriptions, answers, or even new images

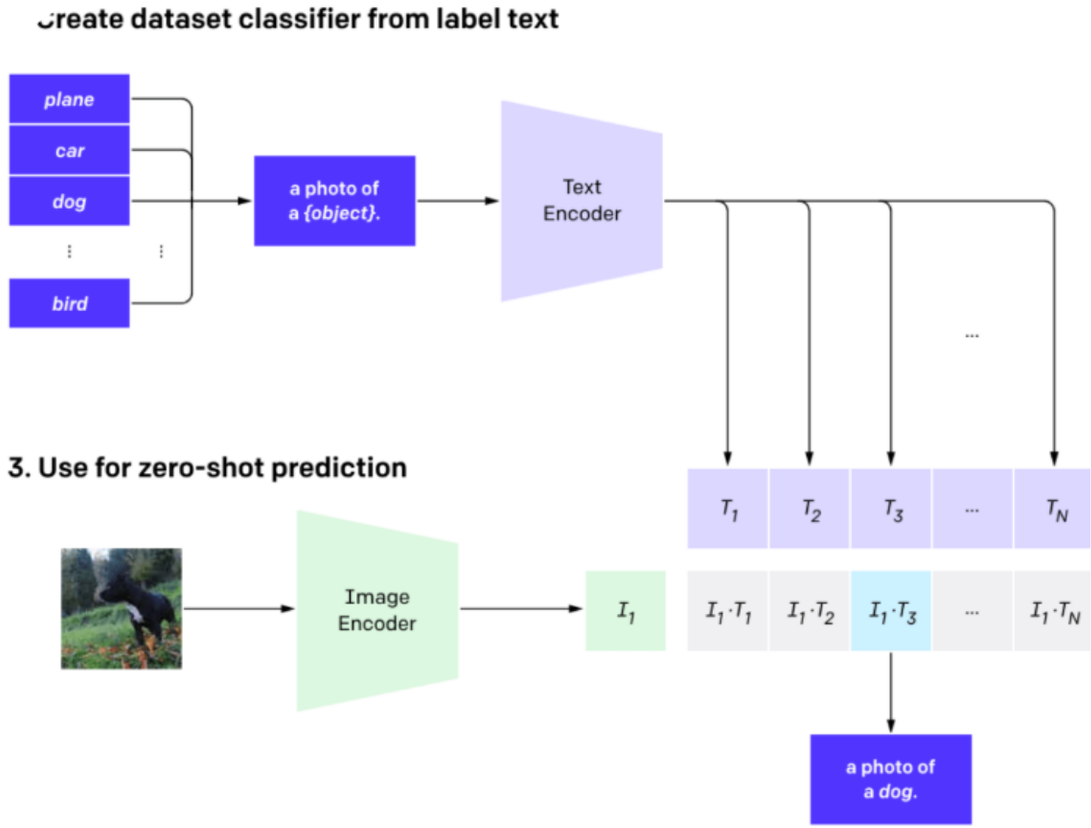
Output Head

- **Purpose**: Generate task-specific outputs
- **Variations**:
  - Classification heads for visual question answering
  - Generation heads for image captioning
  - Regression heads for similarity scoring

LVMs

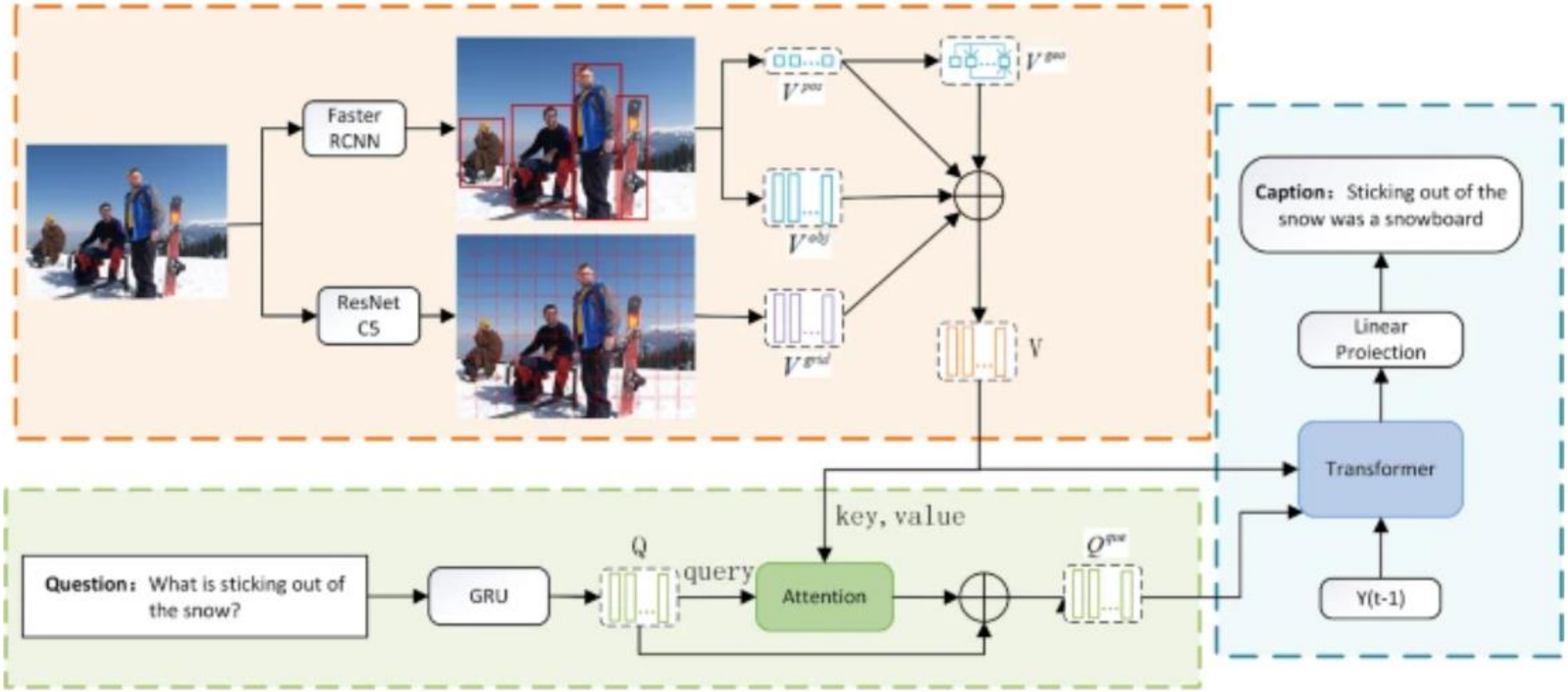
- Large Vision Models (LVMs)





LVMs

- Large Vision Models (LVMs)
- Visual Question Answering Models (VQA)



Uses of Vision Models

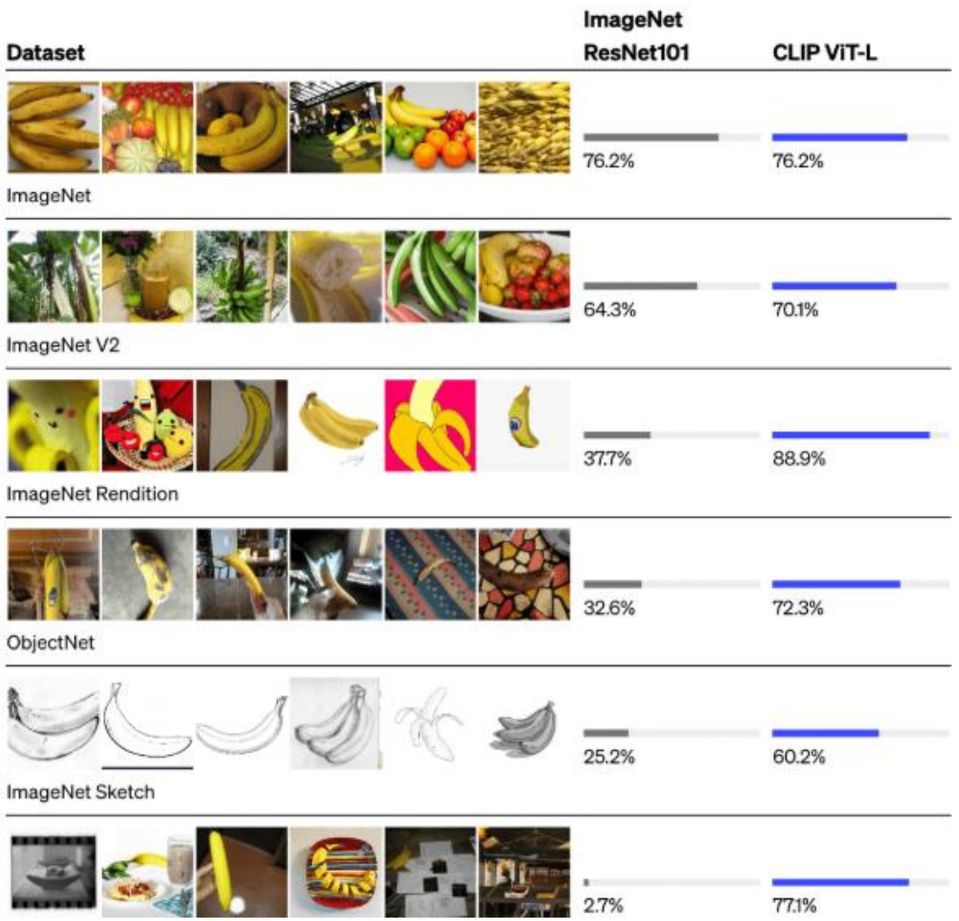
OpenAI's CLIP

- **CLIP or Contrastive Language-Image Pretraining**
- It is a neural network, that undergoes training using diverse sets of images and corresponding text captions.
- Through this process, it acquires the ability to comprehend and articulate the content depicted in images in a manner consistent with natural language descriptions.

Text Encoder (Transformer) ——┐

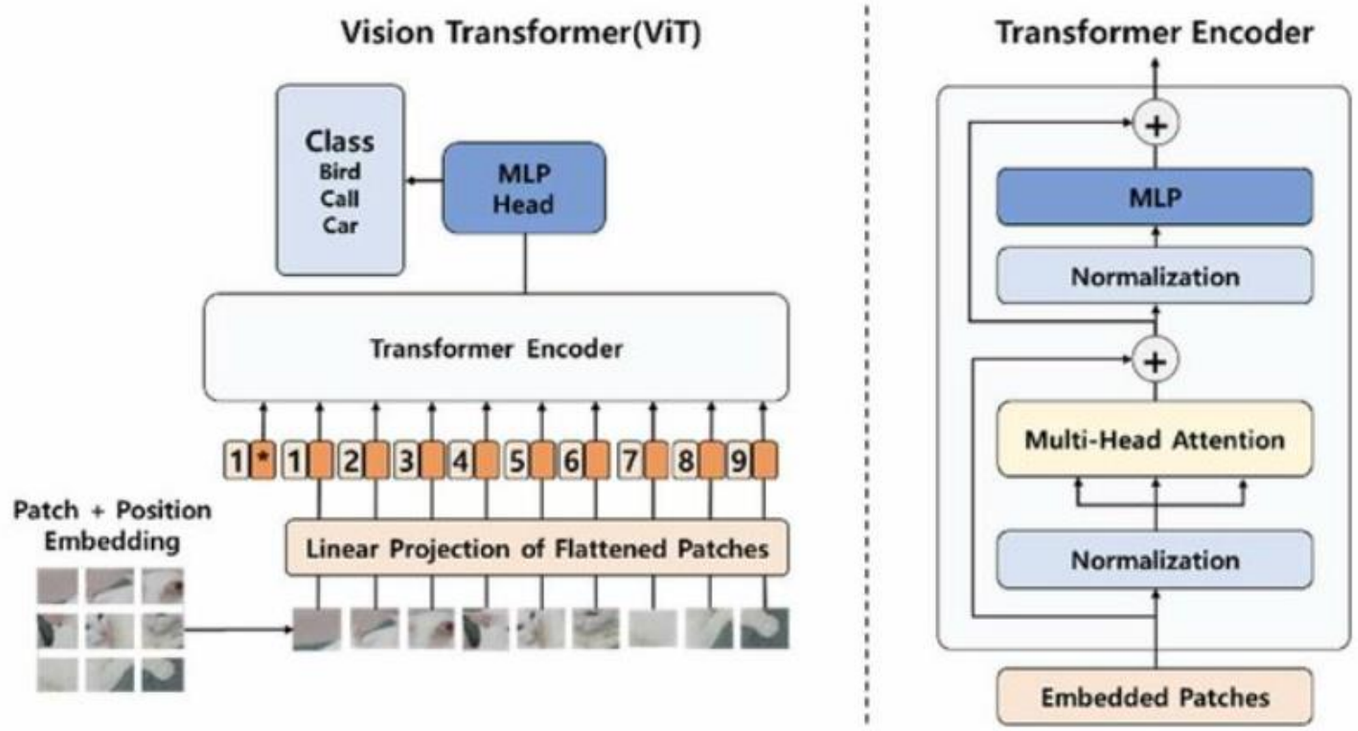
├— Contrastive Loss

Image Encoder (ViT/ResNet) ——┘



LVMs

- **Vision Transformers**
- **Self-attention** replaces convolutions.
- Process images as patches (like NLP tokens).
- Achieved state-of-the-art results with large data + compute.



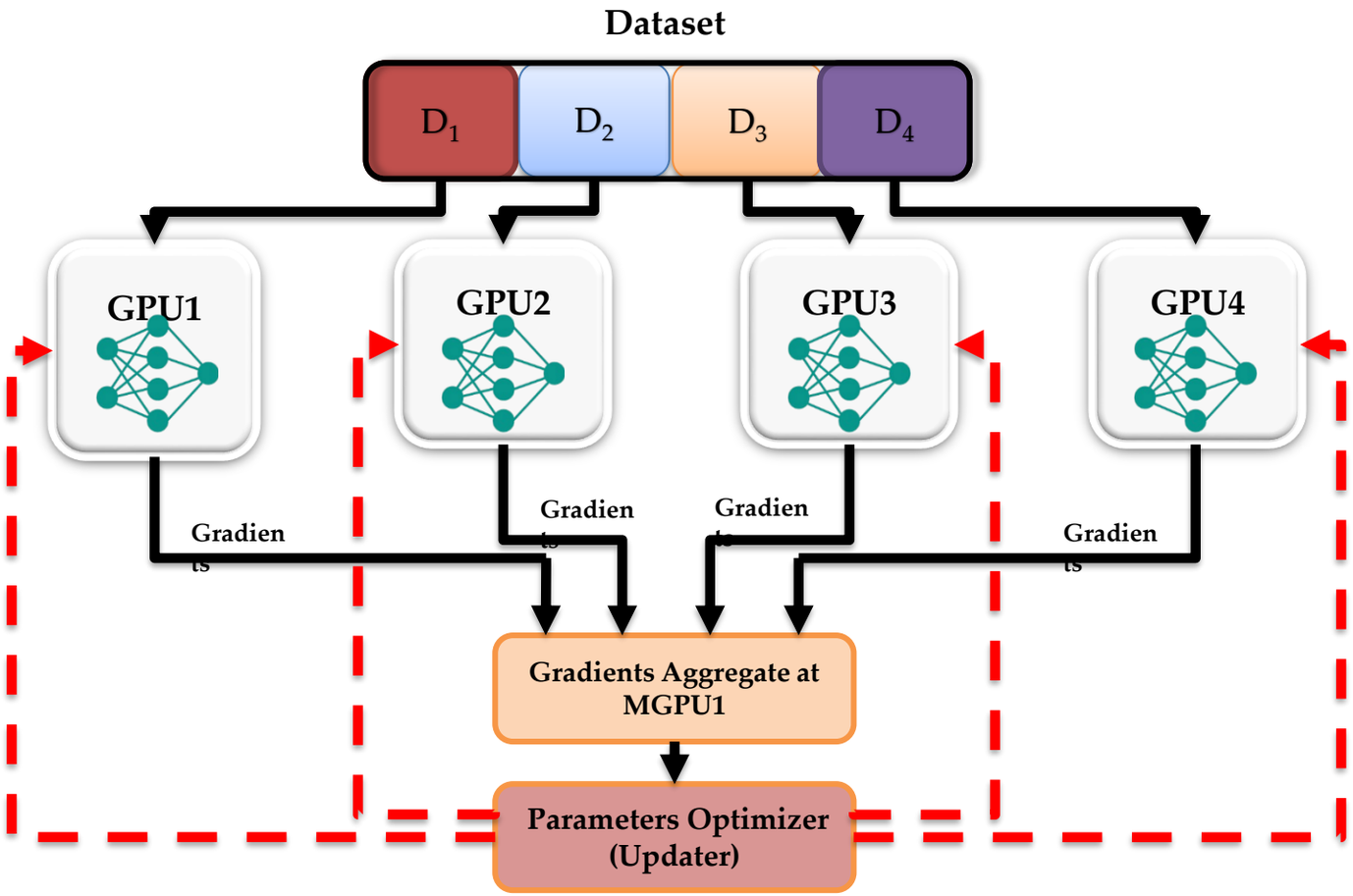
Distributed DL

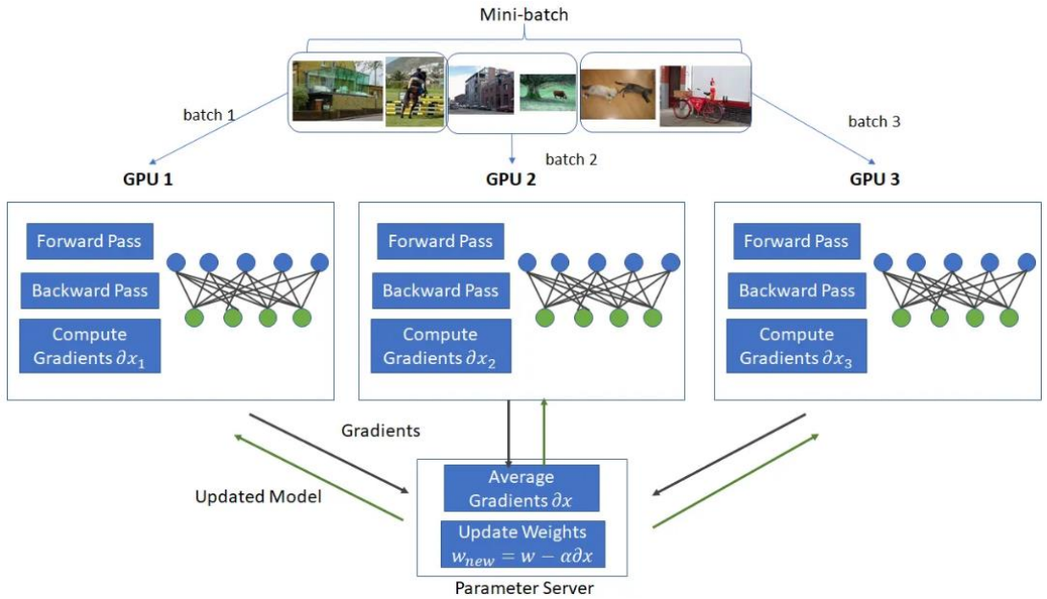
- In distributed deep learning, we trained the deep learning models in distributed manner.
- It is used to reduce the training time of DL model.
- It is of two types:

i) ***Data Parallelism***

ii) ***Model Parallelism***

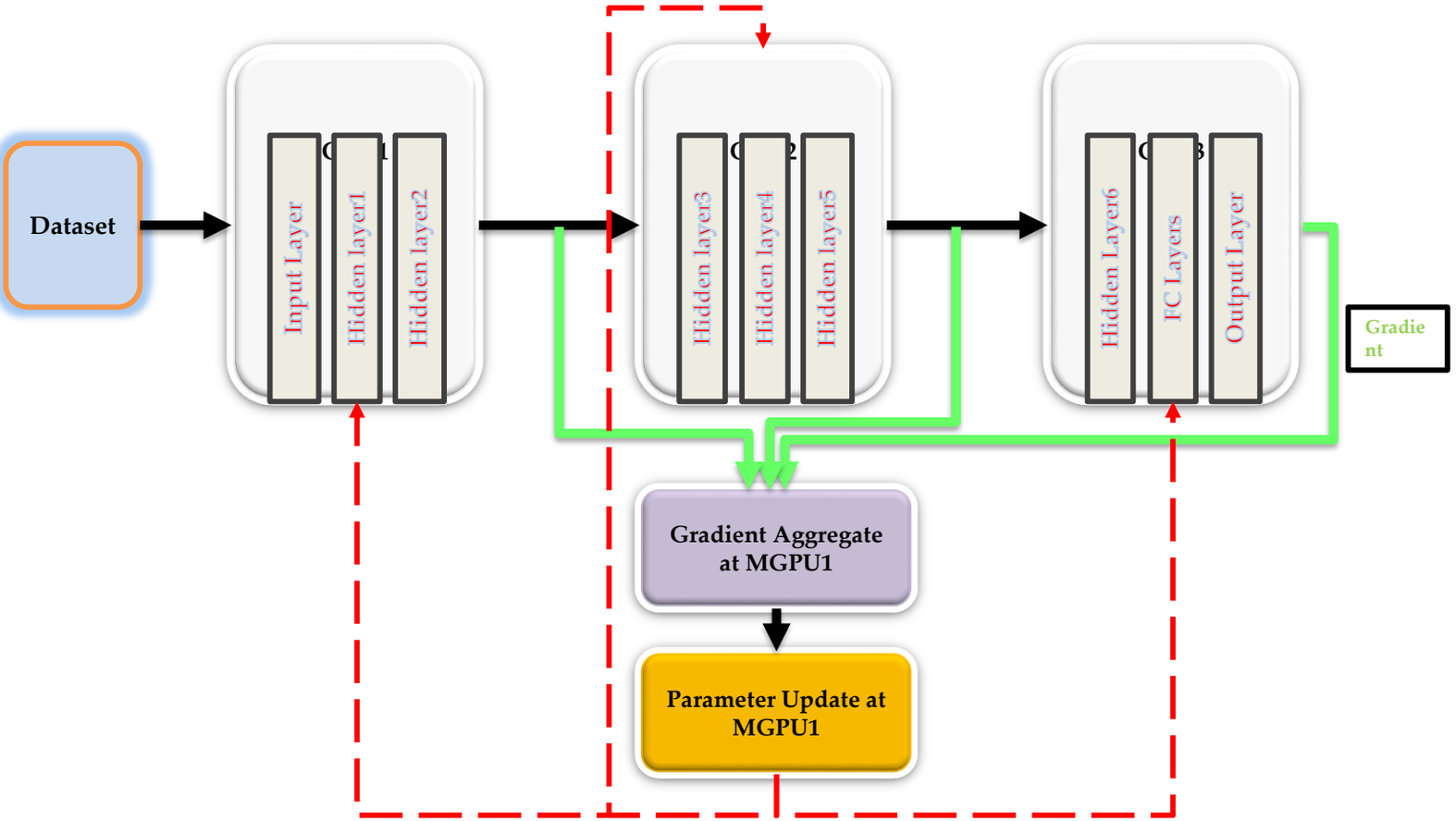
Data Parallelism

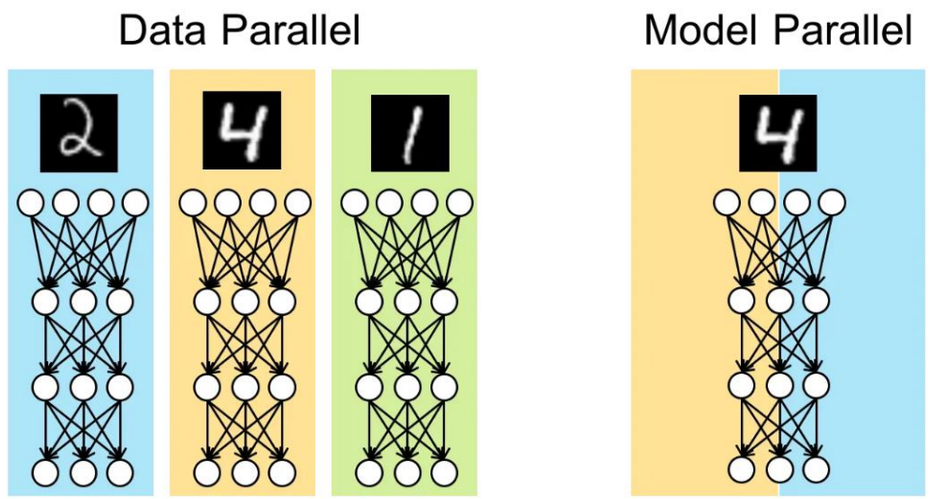




Model Parallelism

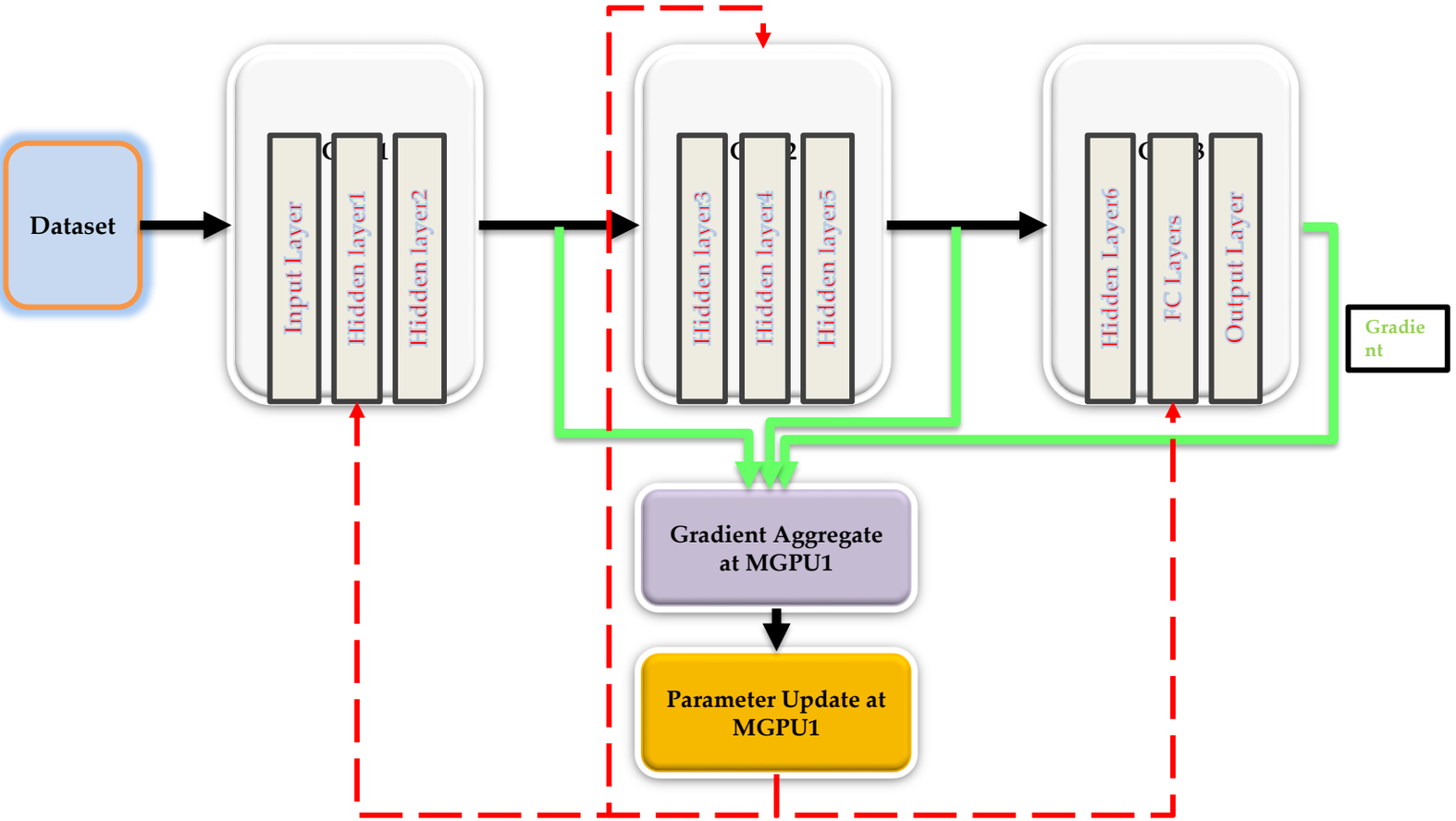
or “operator-level parallelism”

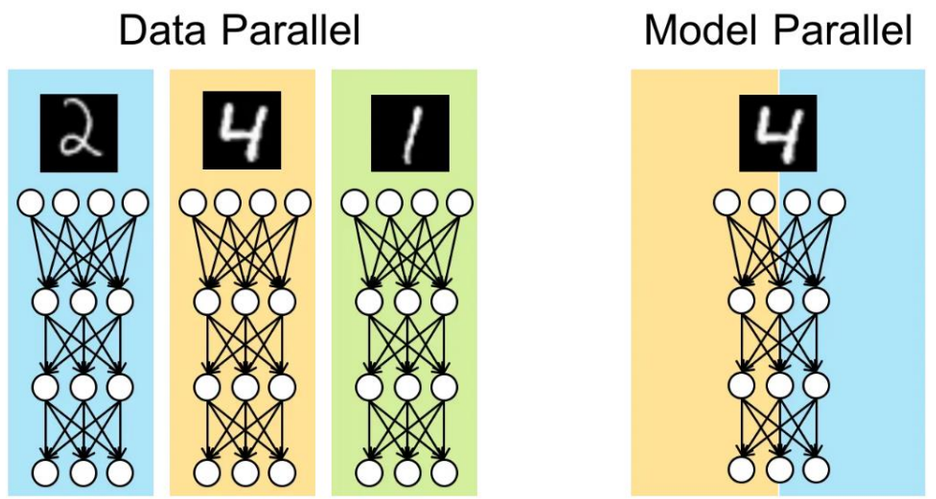




Model Parallelism

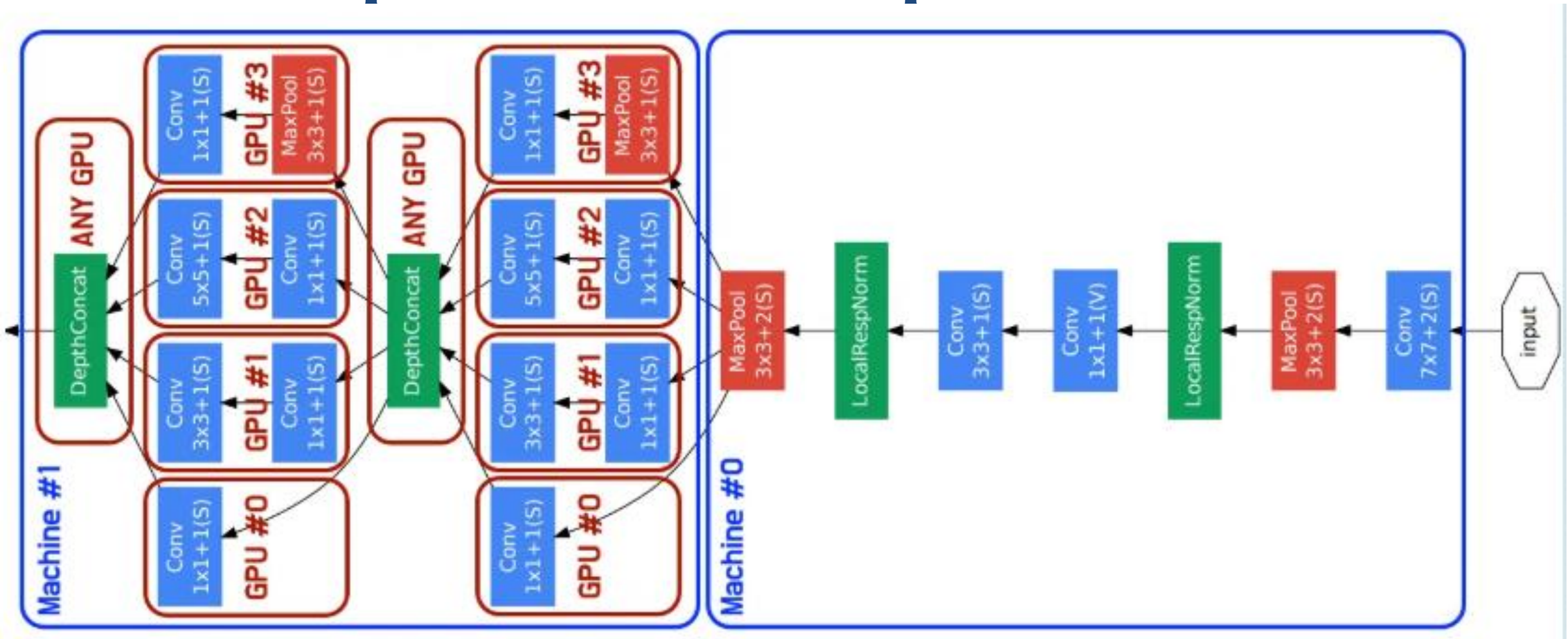
or “operator-level parallelism”





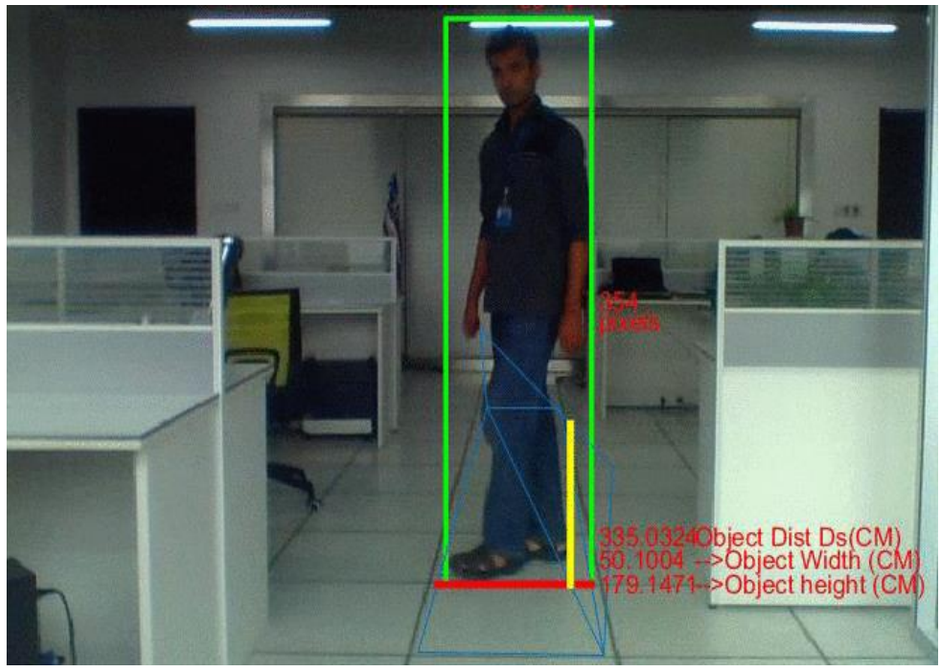
Model Parallelism

or “operator-level parallelism”

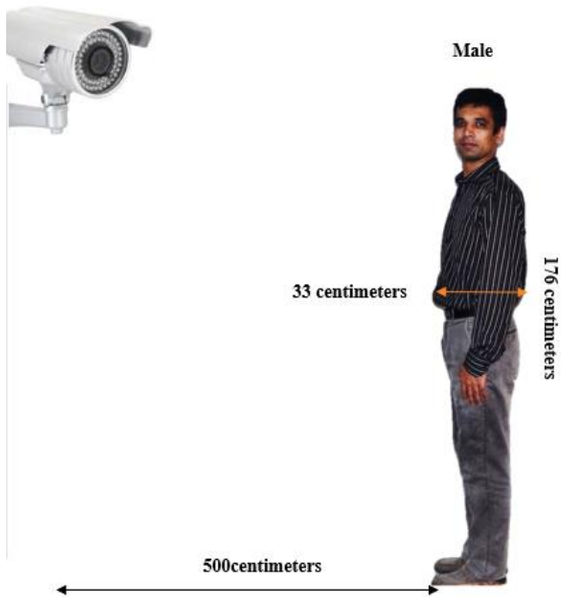


Real-world Exmples

- **Topic:** Automated Identification of Pedestrian' Attributes for Behavior Analysis in Surveillance Systems by Employing Deep Learning Techniques
- **Supervisor Name:** Chen Zonghai, Professor, University of Science and Technology of China (USTC), Anhui province, Hefei China

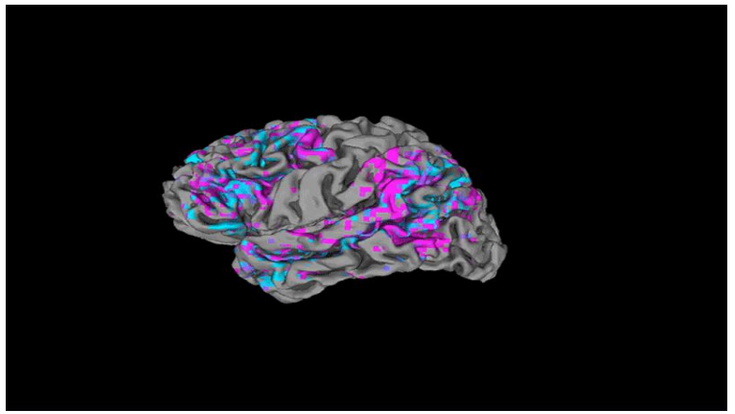


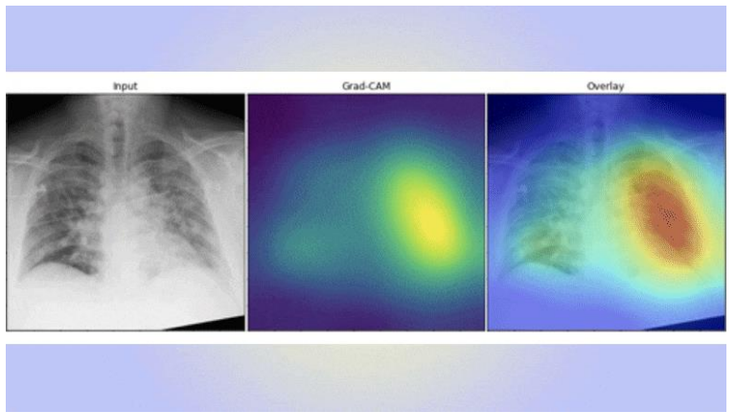
"A pedestrian detected is male and moving at an angle of 270^0 from the camera (He is directing towards a no-pass-through area) He is facing at an angle of 0^0 from the camera (Looking suspicious). Currently, his distance from the camera is 500 centimeters (CM). His height is 176 CM and is 33 CM fat."

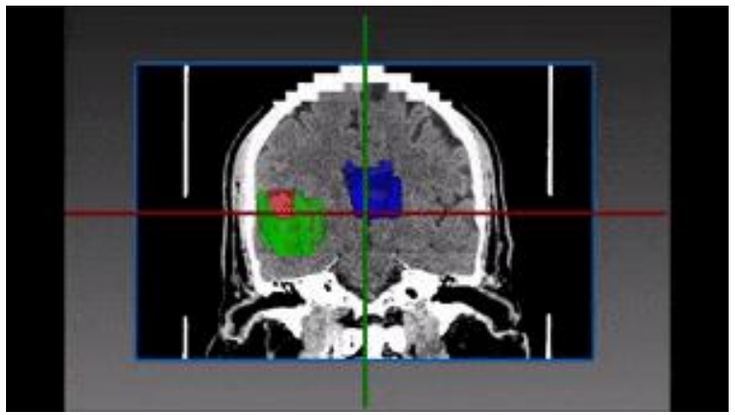


Medical Science

Detection and Diagnosis of Diseases (e.g., cancer, neurological disorders)







Artificial Intelligence

Medical Science

Surgical Planning and Patient Care.

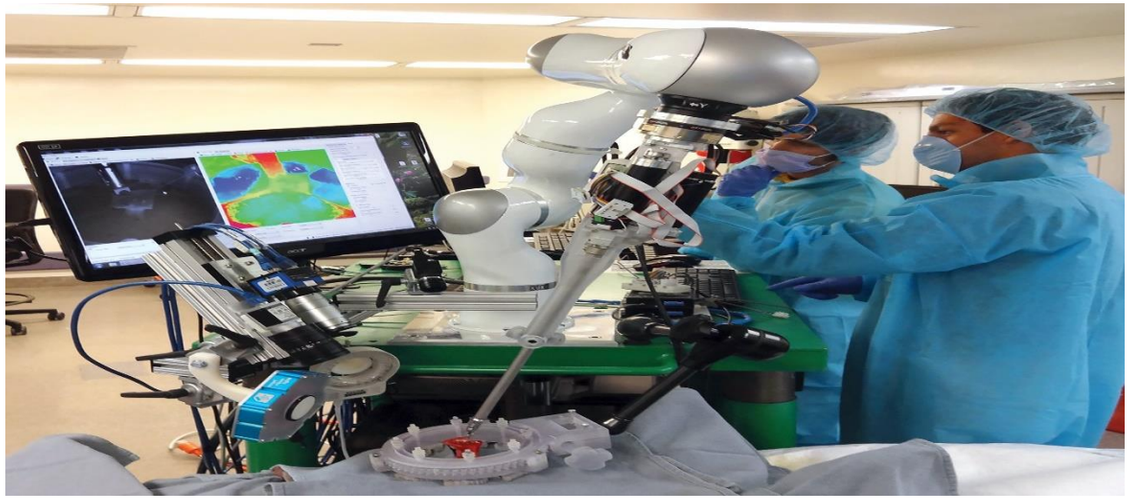
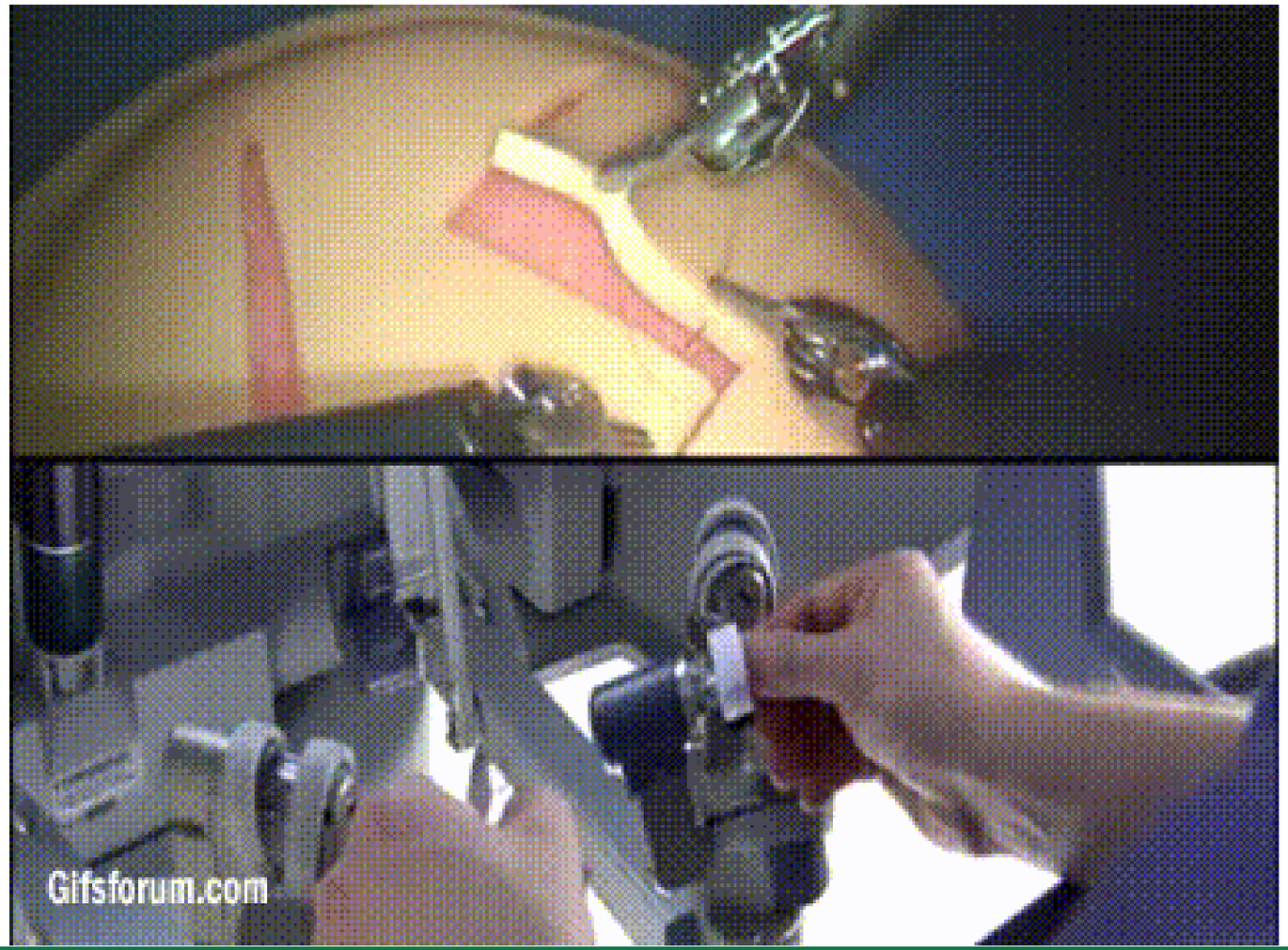
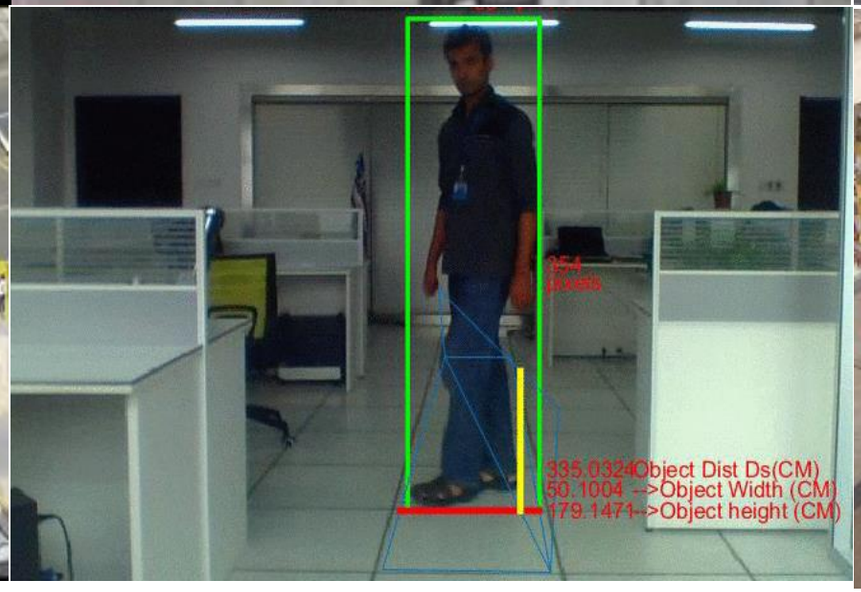
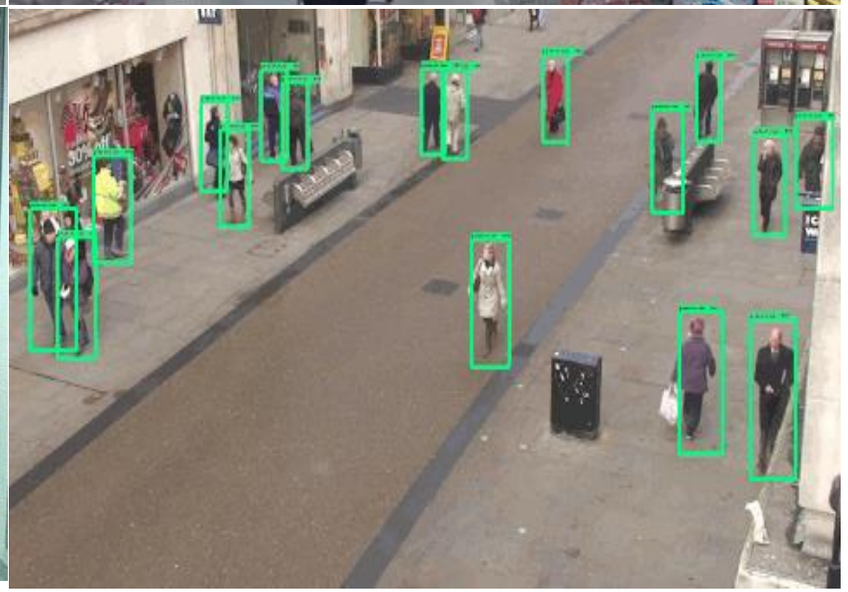


Image guided surgery



Surveillance: Public Safety and Crime Prevention





Street crimes monitoring

General proposed model for automated crime monitoring

Field of View

Camera with alarm

Surveillance setup at home connected to central network

On detection of crime like activity, alarm will be generated. Meanwhile, the location and the scene snapshots will be sent to control room for quick response.

Network

Control room

Exam Proctoring

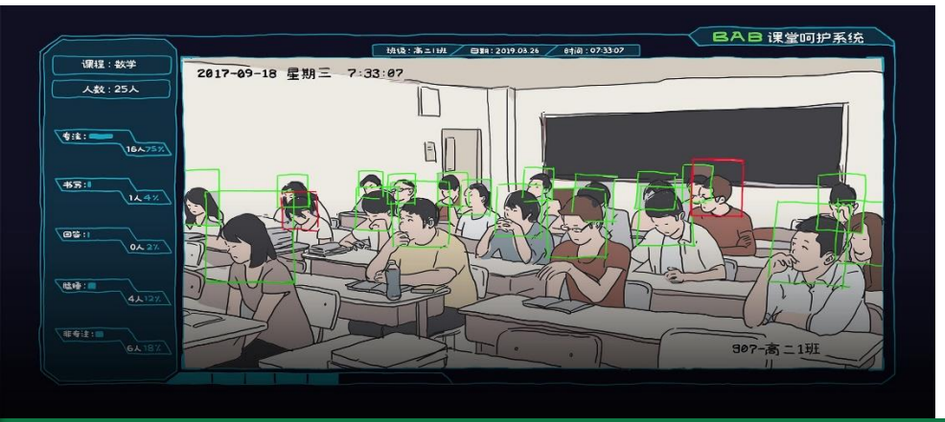
IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

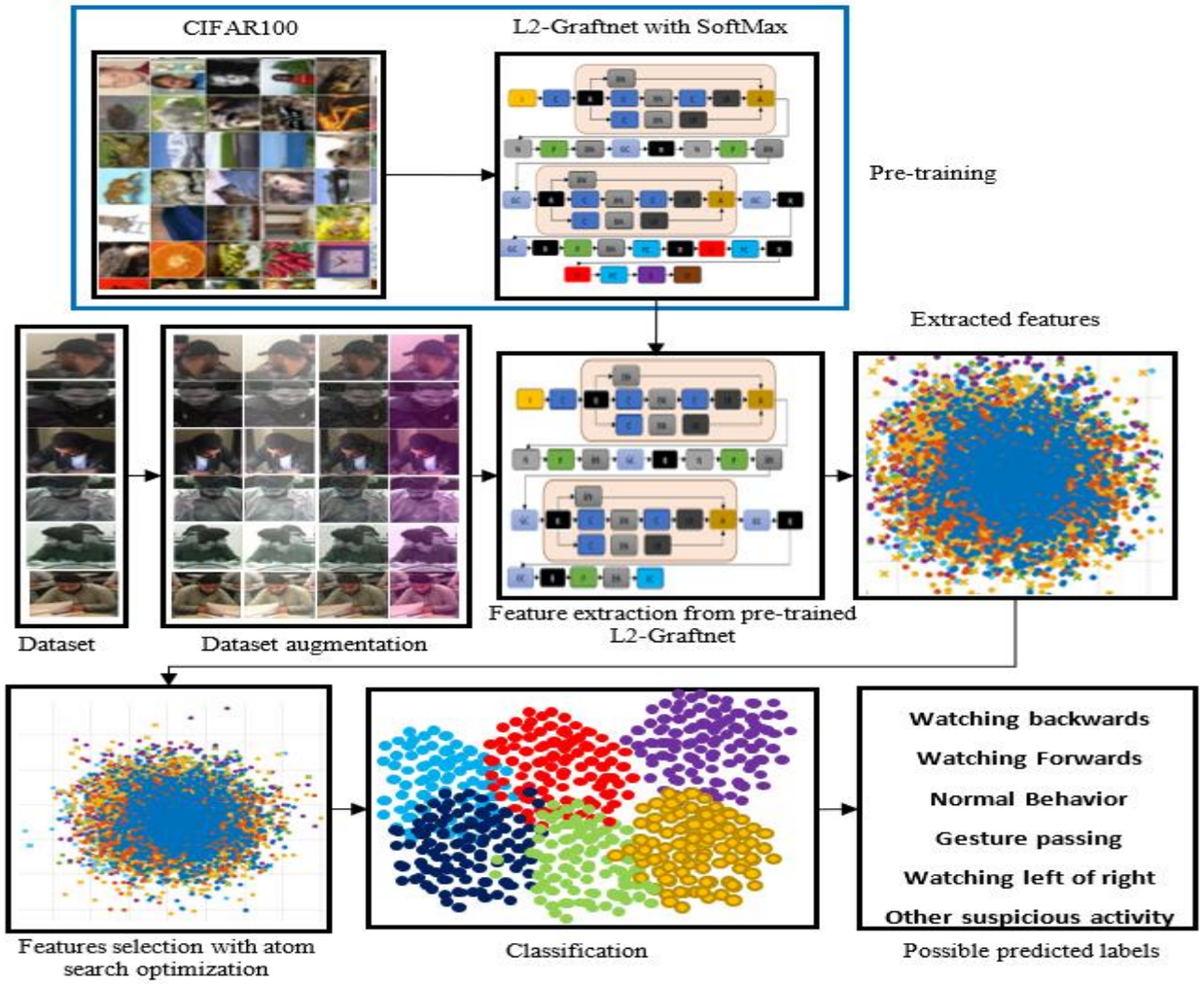
Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Categorizing the Students' Activities for Automated Exam Proctoring using Proposed Deep L2-GraftNet CNN Network and ASO Based Feature Selection Approach

Tanzila Saba[1], Amjad Rehman[1], Nor Shahida Jamail[1], SouadLarabi Marie-Sainte[1], Mudassar Raza[2], and Muhammad Sharif[2]

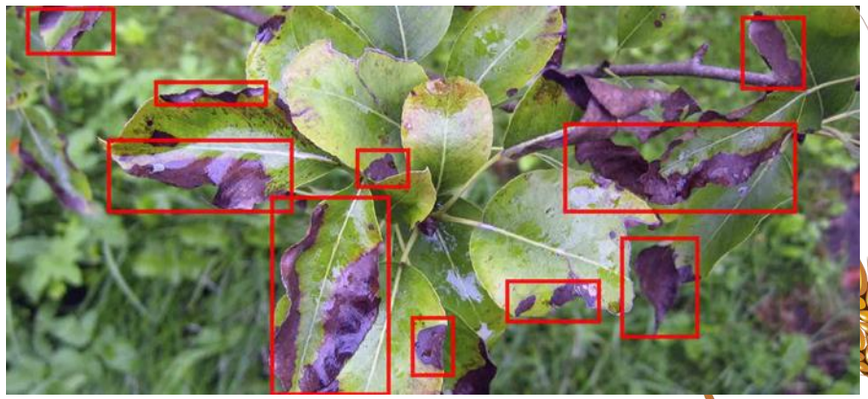




Agriculture

- Disease detection
- Crop readiness identification
- Field management
- Soil survey and mapping
- Monitoring Health of Crops
- Precision Farming

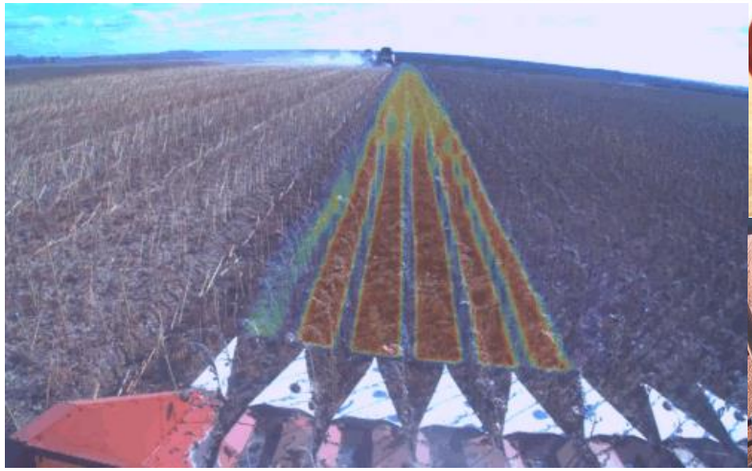
Photo-powered app could help farmers diagnose crop diseases themselves



The Rise of Agbots

ROBOTICS FEATURE AN ARMY OF GRAIN-HARVESTING ROBOTS MARCHES ACROSS RUSSIA

https://spectrum.ieee.org/robotic-farming-russia



Agriculture

Cont…

Automated harvesting systems;

Iron Ox Lettuce Robot

MIT Robot

It looks like a tractor. This machine uses sensors and robotic arms to detect ripe berries and pick these up from the ground.

Agrobot SW6010

A network of sensors attached to each plant monitors the soil humidity and call the robot for water.

Cucumber

WP5

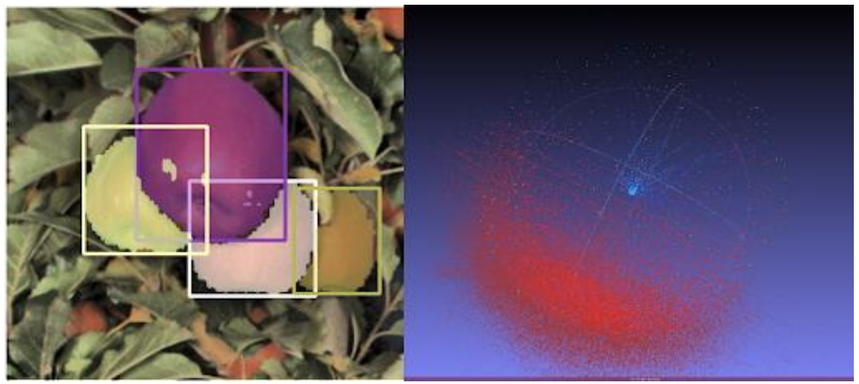
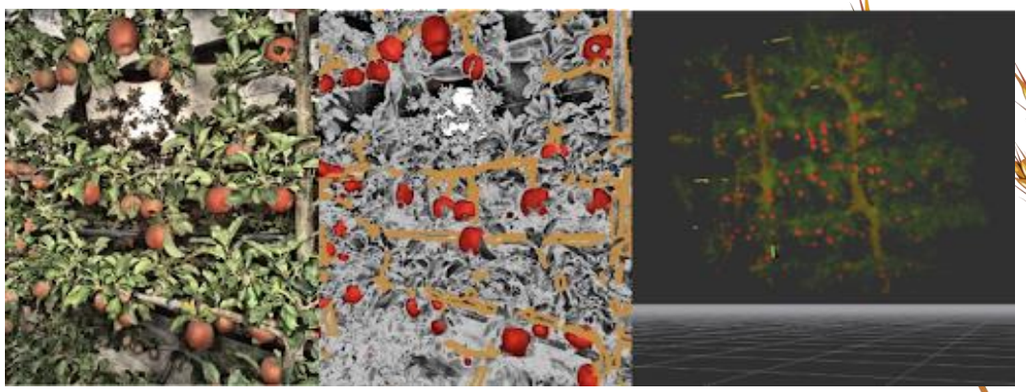
Uses a robotic arm with a gripper to cut the sweet pepper fruit. Two mini-cameras attached to the gripper help the robot to detect the fruits

Agriculture

Cont…

Apple harvesting, automated

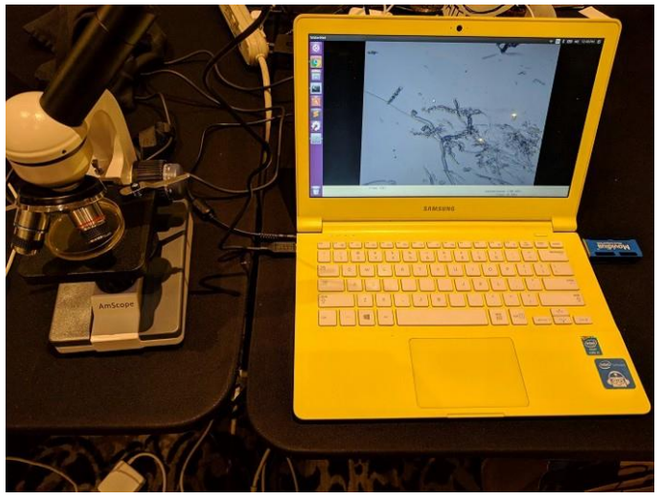


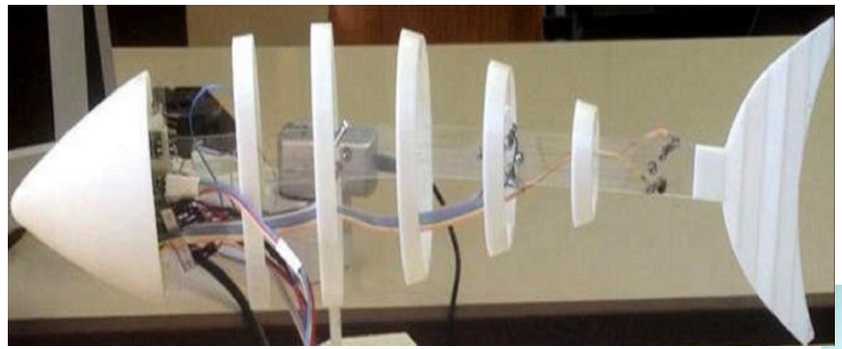




https://waxinvest.com/projects/abundant-robots/

Water Management

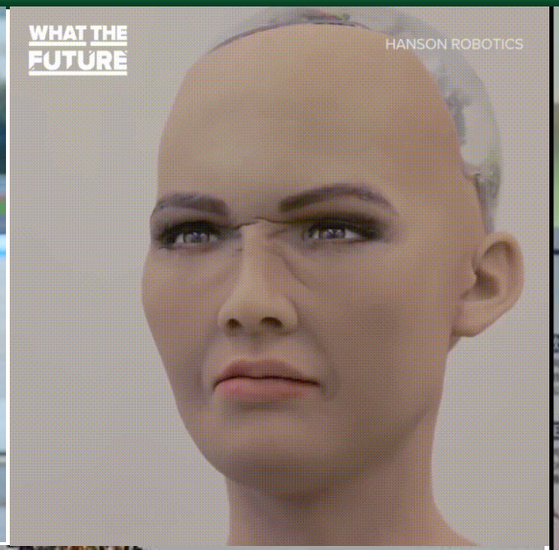
- AI detects bacteria in water
- Every minute a newborn dies from infection caused by lack of safe water and an unclean environment. (World Health Organization, 2017)
- Autonomous robotic fish designed to monitor water quality





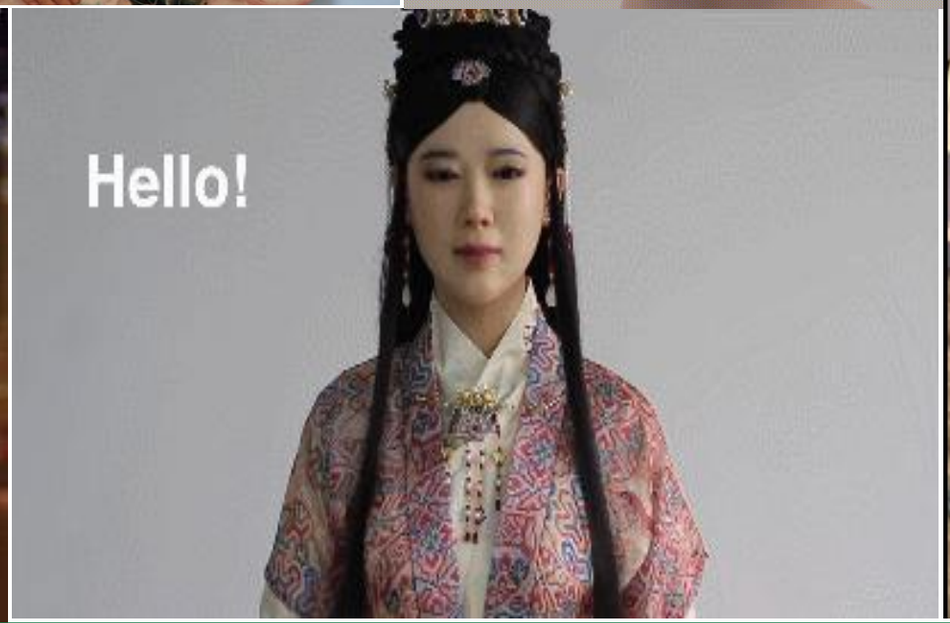
https://becominghuman.ai/how-is-ai-saving-the-planet-92473b41cfa0?gi=ef90e594eb5a

Robotics


WHAT THE FUTURE
HANSON ROBOTICS

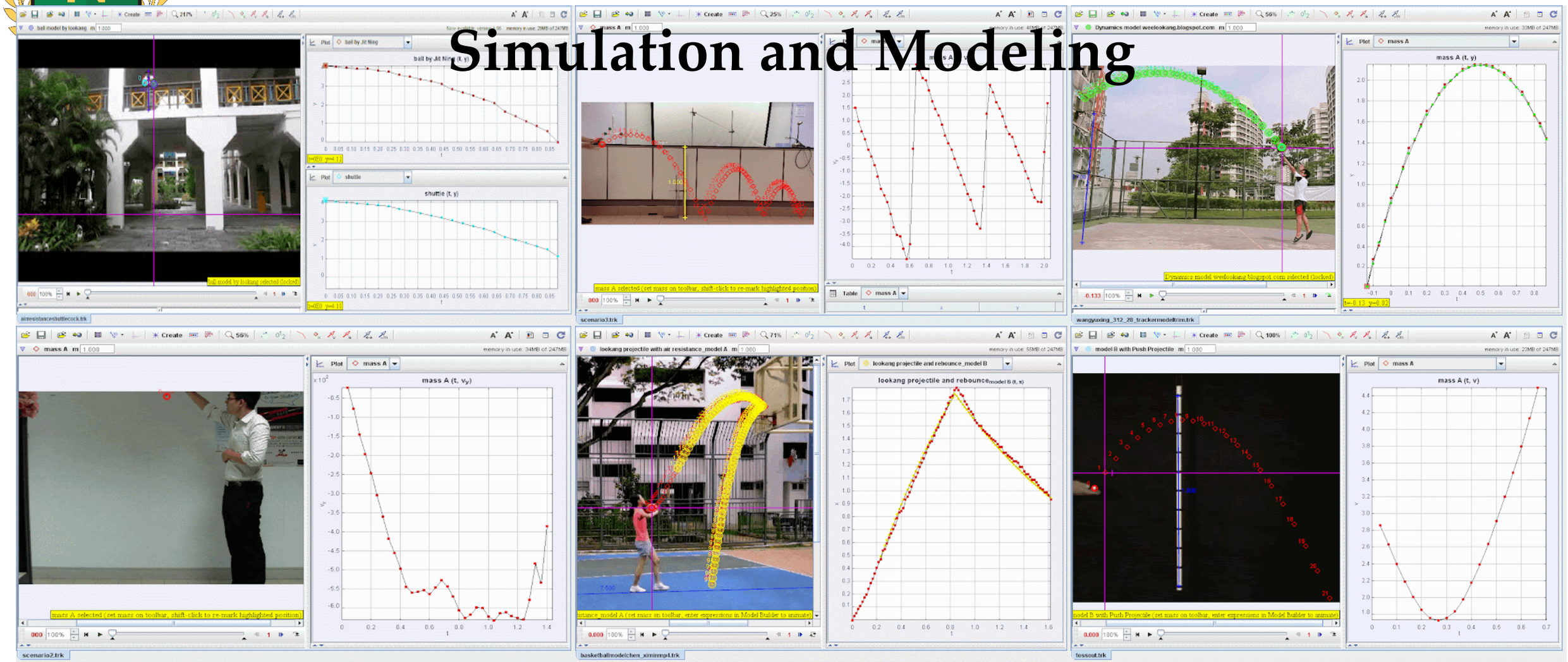

展区
1958
University of Science and Technology of China


#FALLONTONIGHT


Hello!

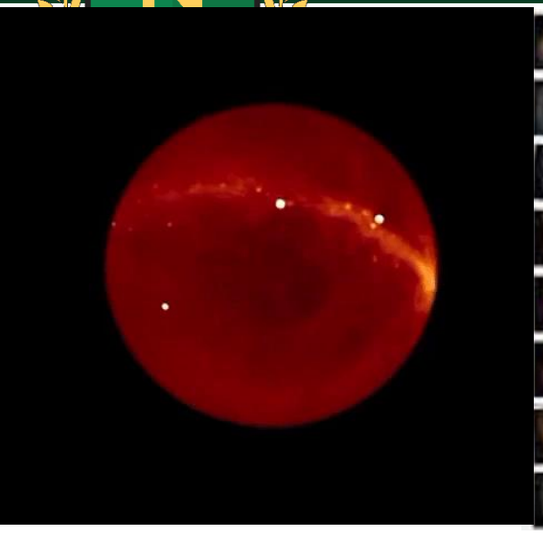
Autonomous Vehicles

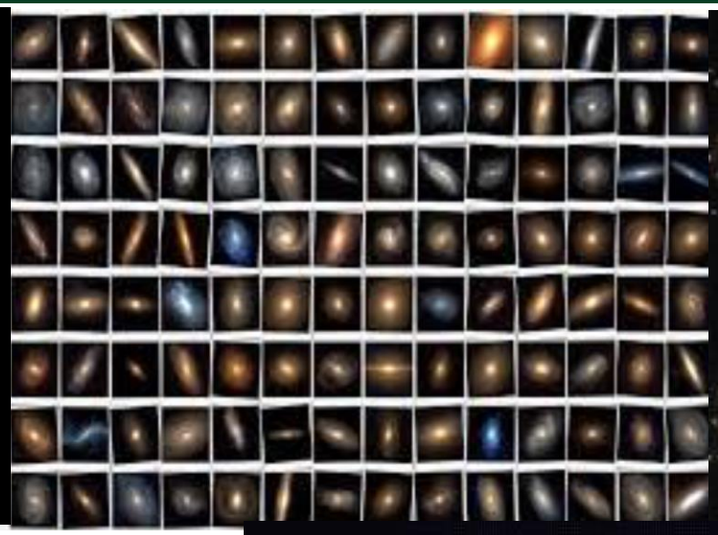
Simulation and Modeling



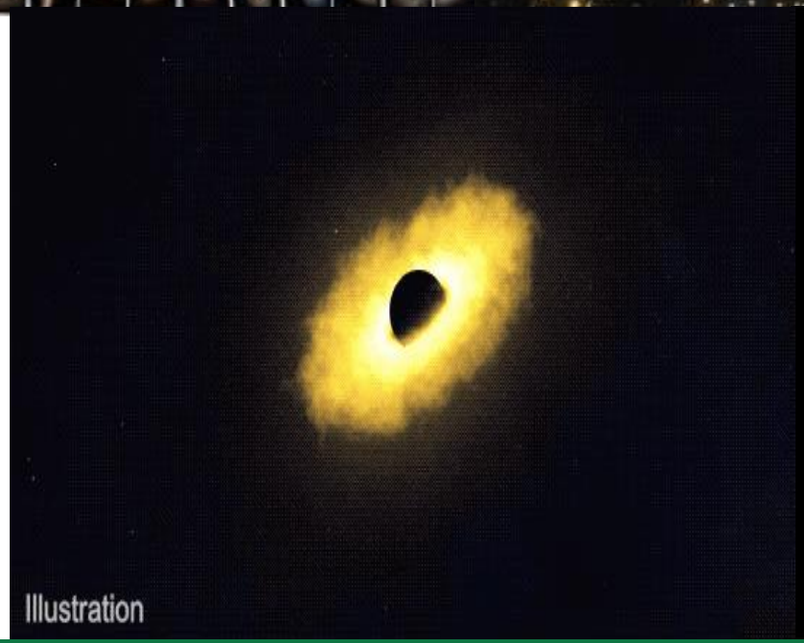
https://weelookang.blogspot.com/2017/03/tracker-animated-gifs.html

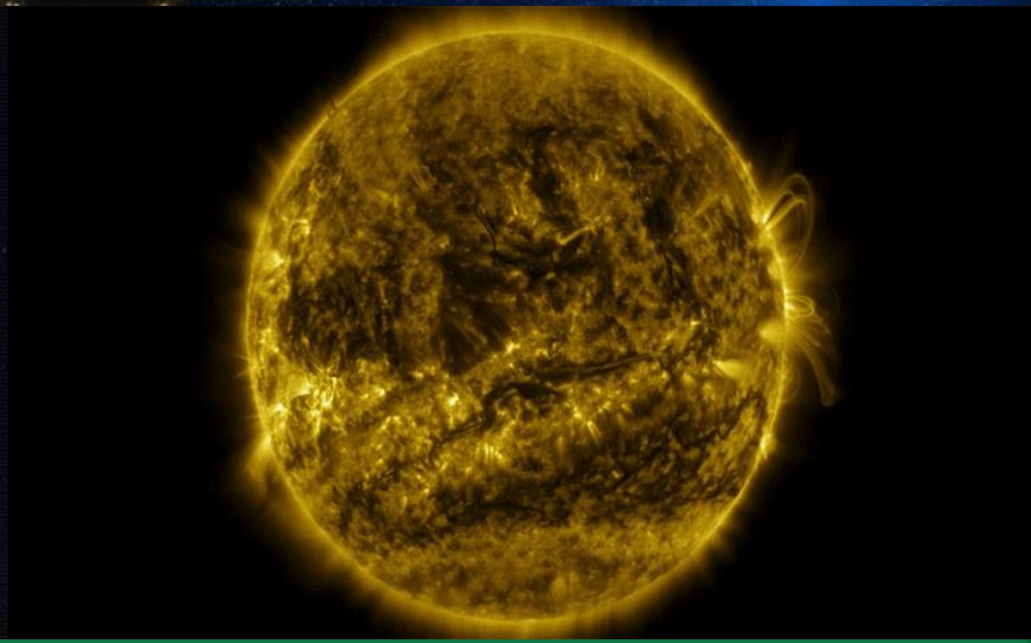
Cosmology and Astrophysics





AI can process and analyze vast amounts of astronomical data, helping astronomers discover new **celestial objects, identify gravitational wave signals, and study the cosmic microwave background radiation**.





Prompt based AI: Towards the Modern Era

Copilot | Designer

Want to see how Image Creator works? Select Surprise Me, then Create

Explore ideas Creations



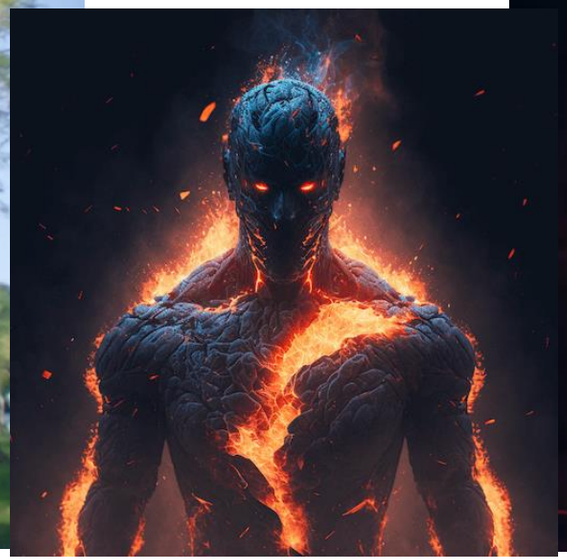





Generative AI: Image to Video

Generative AI: Image to Video

Generative AI: Video to Video



Generative AI: Text to Video

More LVM Applications

More LVM Applications

1. Healthcare

- **Medical Image Analysis**: Combine radiology images with clinical notes
- **Diagnostic Assistance**: Generate reports from medical scans
- **Patient Education**: Create visual explanations of medical conditions

2. Autonomous Systems

- **Self-Driving Cars**: Process visual input with navigation instructions
- **Robotics**: Understand visual scenes and verbal commands
- **Drones**: Navigate using visual input and textual waypoint descriptions

3. Content Creation and Media

- **Automatic Captioning**: Generate descriptions for images and videos
- **Content Moderation**: Identify inappropriate visual and textual content
- **Creative Tools**: Generate images from text descriptions (DALL-E, Midjourney)

4. E-commerce and Retail

- **Product Search**: Find products using natural language descriptions
- **Visual Shopping**: Search for products using images
- **Automated Cataloging**: Generate product descriptions from images

5. Education and Accessibility

- **Visual Learning Aids**: Generate explanations for educational images
- **Accessibility Tools**: Describe images for visually impaired users
- **Language Learning**: Connect visual concepts with vocabulary

6. Social Media and Communication

- **Content Understanding**: Analyze posts with images and text
- **Automatic Tagging**: Generate relevant hashtags and descriptions
- **Content Recommendation**: Suggest content based on visual and textual preferences

Challenges

Challenges

- **Data Challenges**
- Data quality: noisy, biased, unbalanced.
- Annotation costs very high for segmentation.
- Domain adaptation issues.

Challenges

- **Computational Costs**
- LVM training needs GPU/TPU clusters.
- Billions of parameters = expensive training.
- Raises a barrier for small labs/universities.

Challenges

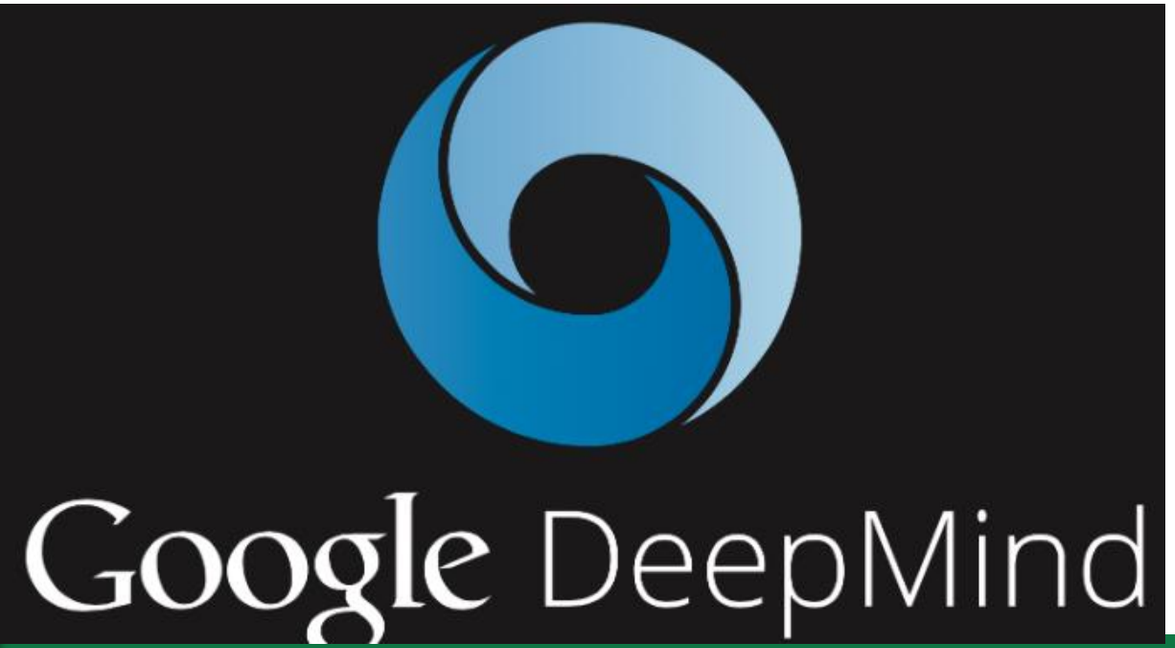
- **Energy & Sustainability**
- Training large models consumes megawatt-hours of energy.
- Carbon footprint of AI models is rising.
- Push towards green AI and efficiency.

Challenges

- **Energy & Sustainability**
- Training large models consumes megawatt-hours of energy.
- Carbon footprint of AI models is rising.
- Push towards green AI and efficiency.

Case Study: DeepMind & NHS Privacy Issue

- DeepMind collaborated with the UK NHS in 2016.
- Patient data was shared without adequate consent.
- This raised questions of medical data privacy.
- Transparency and consent are essential in healthcare AI.



1. **Lack of Patient Consent**
   - Patients were **not informed** that their medical data was being shared with DeepMind.
   - No option for patients to opt out.
2. **Excessive Data Sharing**
   - The volume of data shared went **beyond what was needed** to test and develop the AKI detection app.
   - Data on patients who **did not even have kidney problems** was included.
3. **Regulatory Ruling**
   - In **2017**, the UK's **Information Commissioner's Office (ICO)** ruled that the data sharing was **illegal**, because:
     - The NHS trust did not properly inform patients.
     - Data sharing violated the **UK Data Protection Act** (the law at that time).

Ethical Concerns

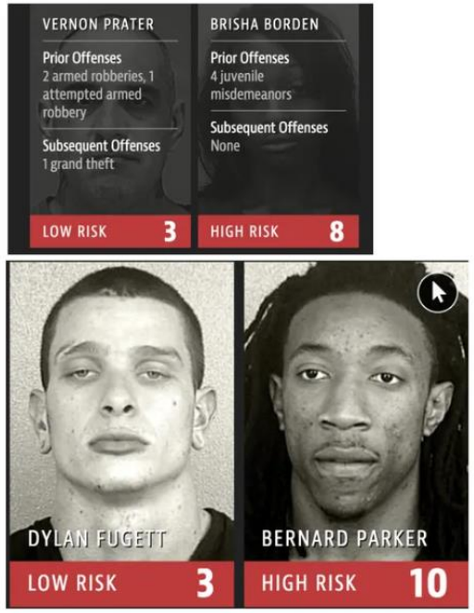
Case Study: Tesla Autopilot Safety Concerns

- Tesla's Autopilot has been linked to fatal crashes.
- Responsibility between driver and AI was unclear.
- Raised issues about testing and deploying AI safely.
- Highlighted gaps in transportation regulation.

Safety concerns for Tesla's Autopilot system, currently under scrutiny by the NHTSA, include its failure to adequately detect hazards, leading to crashes with stationary objects and fatalities.

Case Study: COMPAS Algorithm Bias

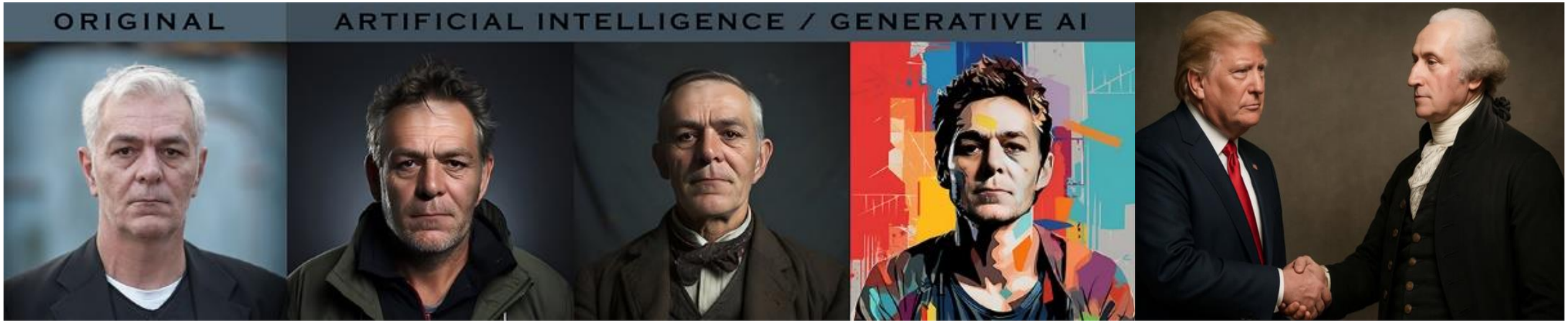
- COMPAS was used in the US justice system to predict recidivism.
- It showed higher false positive rates for African-Americans.
- This case highlighted AI's bias risks and lack of transparency.
- It sparked global discussions on fairness in AI.



The COMPAS software employs an algorithm to evaluate the likelihood of future criminal behaviour, which we terms commonly as ***recidivism***.

Case Study: Deepfakes & Generative AI

- Deepfakes generate realistic but fake videos and images.
- They threaten democracy, security, and personal reputation.
- Examples include fake political speeches and celebrity videos.
- Detection tools and regulation are urgently needed.



Responsible AI

- AI Governance **ensures the responsible use of artificial intelligence** across industries.
- Responsible AI means **AI that is fair, transparent, and accountable**.
- It **prioritizes human dignity** and safety while **protecting rights**.
- Responsible AI also **ensures sustainability** and inclusivity.

Future Directions

- Efficient LVMs for edge devices.
- Multimodal models: text, vision, audio, video.
- Human-in-the-loop AI for safer decisions.

Thank You

- mudassar.raza@namal.edu.pk