AI in Vision and Perception

June 2025

Topics to be discussed

Part I. Introduction

1. Introduction to Artificial intelligence

Part II. Search and optimisation.

2. Search - basic approaches
3. Search - optimisation
4. Two-player deterministic games
5. Evolutionary and genetic algorithms

Part III. Machine learning and data analysis.

6. Regression, classification and clustering (Part I & II)
8. Artificial neural networks
9. Bayesian models
10. Reinforcement Learning

Part IV. Logic, Inference, Knowledge Representation

11. Propositional logic and predicate logic
12. Knowledge Representation

Part V. AI in Action: Language, Vision

13. AI in Natural language processing
14. AI in Vision and Perception

Part VI. Summary

15. AI engineering, Explainable AI, Ethics

Outline

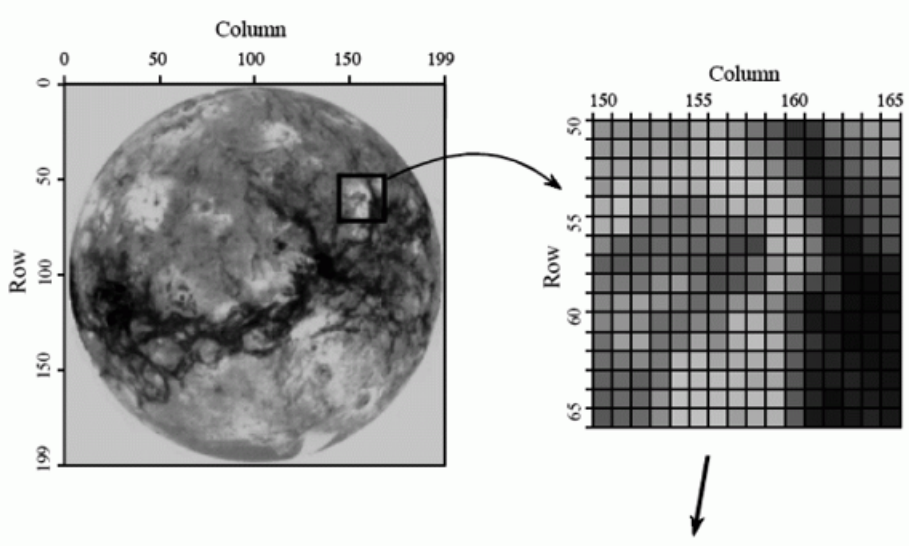
Vision: The ability to interpret and understand visual information from the surrounding environment.
Perception: The process of acquiring, organizing, and interpreting sensory information to understand the world.

AI in Vision and Perception:

- The field of AI aims to develop algorithms and systems that enable computers to perceive and understand visual information, similar to how humans do.
- AI in Vision and Perception involves the use of advanced techniques and algorithms to analyze and interpret visual data, such as images and videos, and extract meaningful information from them.

Introduction to AI in Vision and Perception

Digital Image Structure



Column

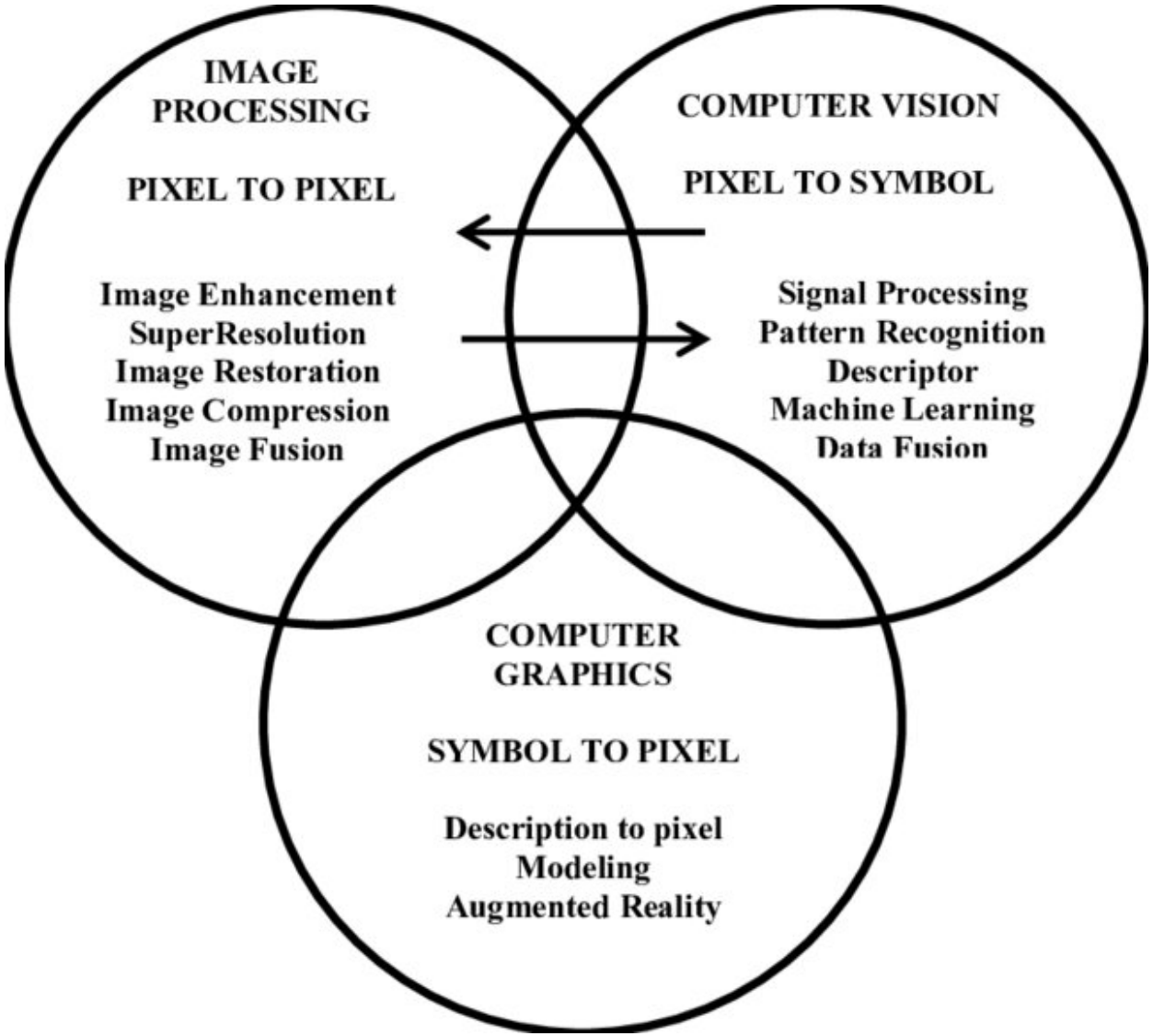
Row	150	155	160	165
50	183 183 181 184 177 200 200 189	159 135 94 105	160 174 191 196	
	186 195 190 195 191 205 216 206	174 153 112 80	134 157 174 196	
	194 196 198 201 206 209 215 216	199 175 140 77	106 142 170 186	
	184 212 200 204 201 202 214 214	214 205 173 102	84 120 134 159	
	202 215 203 179 165 165 199 207	202 208 197 129	73 112 131 146	
55	203 208 166 159 160 168 166 157	174 211 204 158	69 79 127 143	
	174 149 143 151 156 148 146 123	118 203 208 162	81 58 101 125	
	143 137 147 153 150 140 121 133	157 184 203 164	94 56 66 80	
	164 165 159 179 188 159 126 134	150 199 174 119	100 41 41 58	
	173 187 193 181 167 151 162 182	192 175 129 60	88 47 37 50	
60	172 184 179 153 158 172 163 207	205 188 127 63	56 43 42 55	
	156 191 196 159 167 195 178 203	214 201 143 101	69 38 44 52	
	154 163 175 165 207 211 197 201	201 199 138 79	76 67 51 53	
	144 150 143 162 215 212 211 209	197 198 133 71	69 77 63 53	
	140 151 150 185 215 214 210 210	211 209 135 80	45 69 66 60	
65	135 143 151 179 213 216 214 191	201 205 138 61	59 61 77 63	

FIGURE 23-1
Digital image structure. This example image is the planet Venus, as viewed in reflected microwaves. Digital images are represented by a two-dimensional array of numbers, each called a *pixel*. In this image, the array is 200 rows by 200 columns, with each pixel a number between 0 to 255. When this image was acquired, the value of each pixel corresponded to the level of reflected microwave energy. A *grayscale* image is formed by assigning each of the 0 to 255 values to varying shades of gray.

https://www.dspguide.com/ch23/1.html

Introduction to AI in Vision and Perception

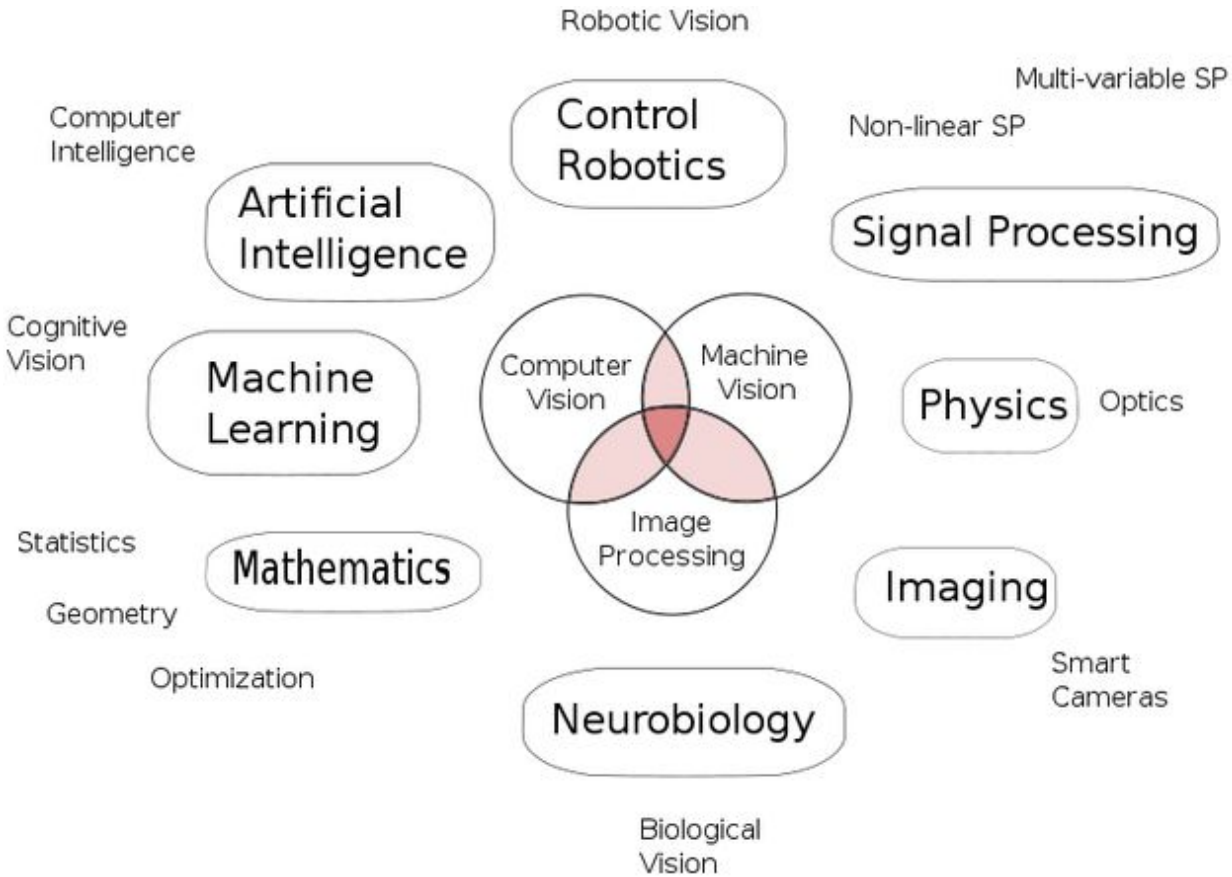
Classical Description of Computer Vision, Image Processing and Computer Graphics



https://www.researchgate.net/publication/235339437

Introduction to AI in Vision and Perception

Computer Vision and Corresponding Fields



https://www.c-sharpcorner.com/article/a-quick-introduction-to-computer-vision-using-c-sharp/

Introduction to AI in Vision and Perception

History of Computer Vision and Integration with AI Models

1960s: Early Years

- Initial research focused on simple image processing tasks, such as edge detection and pattern recognition.
- Development of early computer vision systems for industrial applications like quality control in manufacturing.

1970s: Image Understanding

- Research expanded to higher-level image understanding tasks, such as scene analysis and object recognition.
- Introduction of rule-based systems and symbolic representations for image interpretation.

1980s: Knowledge-Based Vision

- Emphasis on knowledge-based approaches that incorporated domain-specific knowledge and reasoning.
- Integration of geometric and physical constraints in computer vision algorithms.

1990s: Emergence of Machine Learning

- Machine learning techniques, such as neural networks and statistical models, gained prominence in computer vision.
- Development of robust algorithms for tasks like face detection, object recognition, and image segmentation.

2000s: Large-Scale Datasets and Deep Learning

- Introduction of large-scale annotated datasets, such as ImageNet, driving advancements in object recognition and image classification.
- Rise of deep learning with convolutional neural networks (CNNs) achieving state-of-the-art results in various computer vision tasks.

2010s: Advancements in Deep Learning and Real-Time Vision

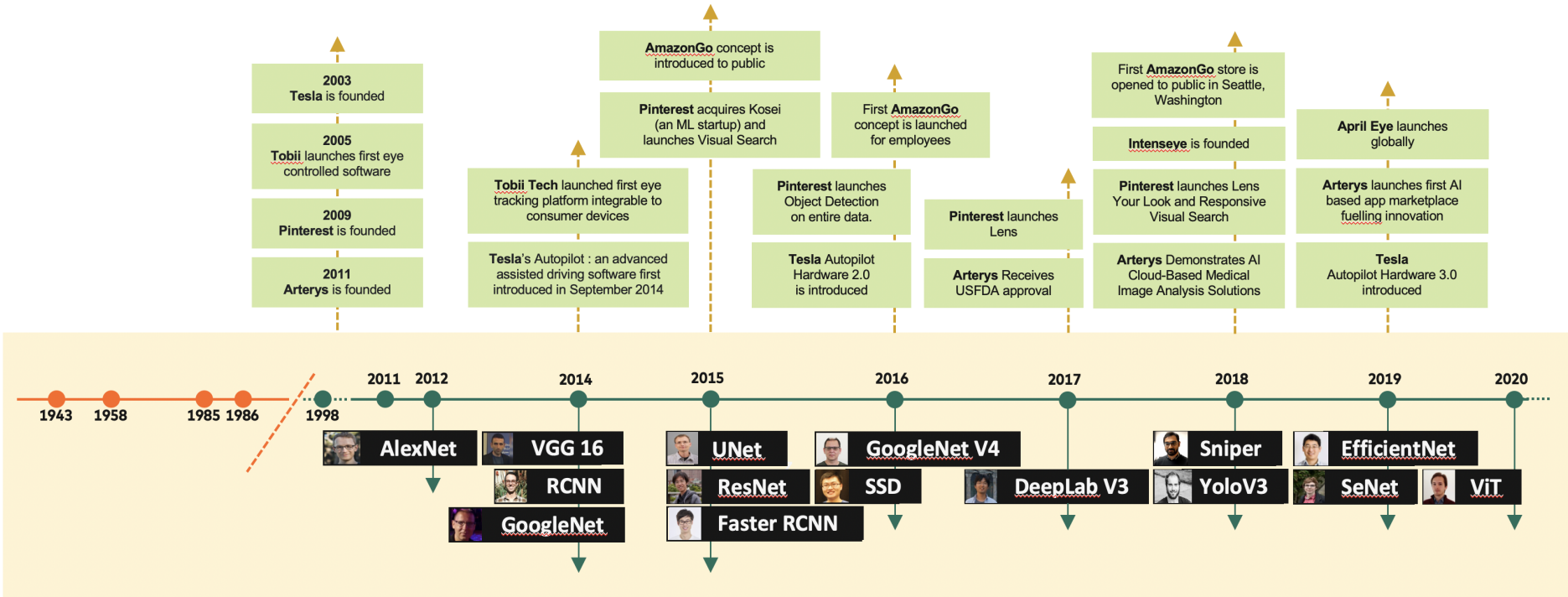
- Development of more sophisticated deep learning architectures, such as recurrent neural networks (RNNs) and generative adversarial networks (GANs).
- Real-time computer vision applications became more prevalent, enabled by faster hardware and optimized algorithms.
- Integration of computer vision with natural language processing (NLP) models like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers).

2020s: Transformer-based architectures, Integration with AI and Robotics

- AI models like GPT, BERT, and other Transformer-based architectures playing a crucial role in Computer Vision for tasks like image captioning, visual question answering, and multimodal learning.
- Computer vision also extensively used in robotics applications, including perception, object manipulation, and autonomous navigation.

Introduction to AI in Vision and Perception

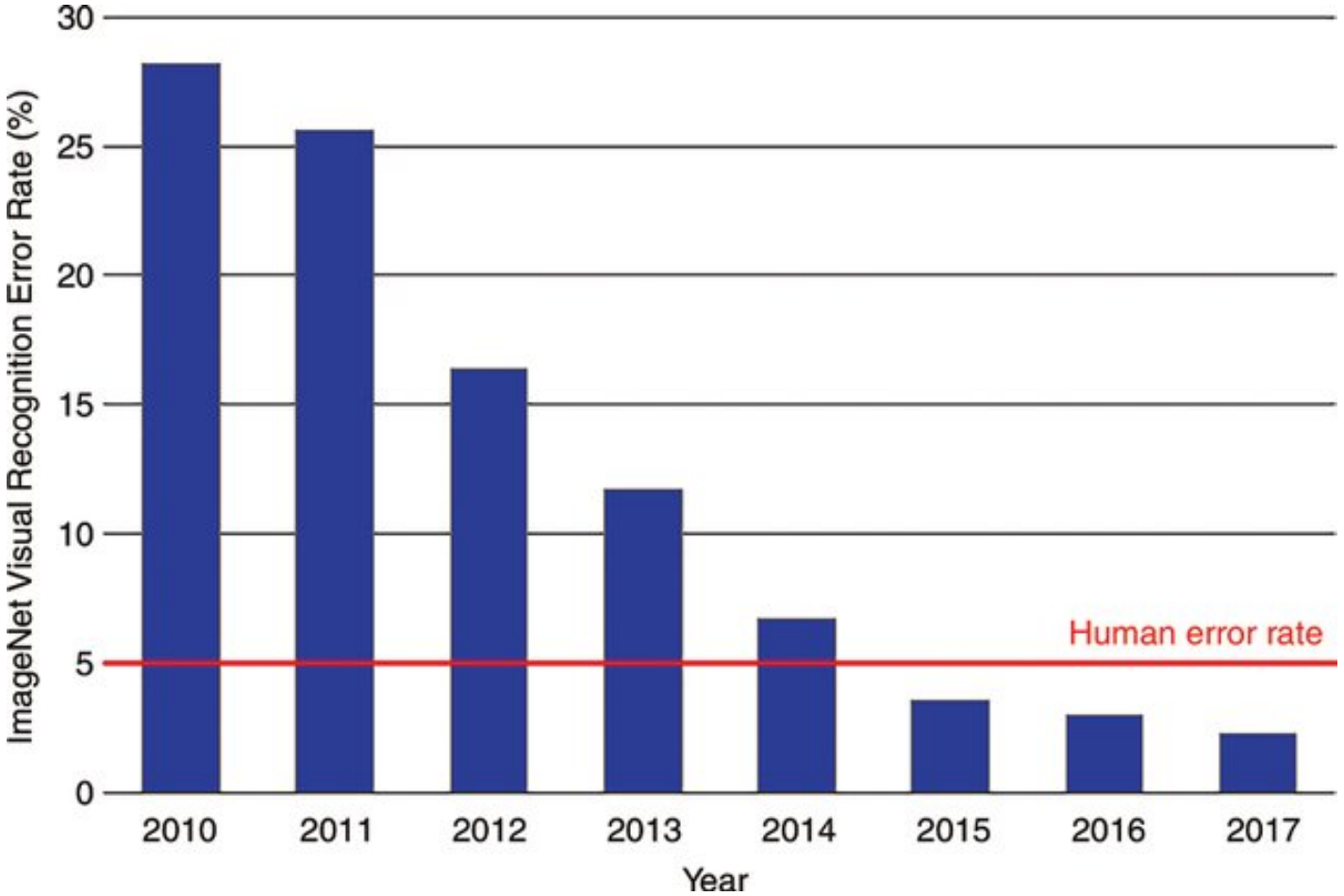
History of Computer Vision and Integration with AI Deep Learning Models (cont.)



https://www.analyticsvidhya.com/blog/2021/08/computer-vision-and-how-it-is-shaping-the-world-around-us/

Introduction to AI in Vision and Perception

ImageNet Error Rates



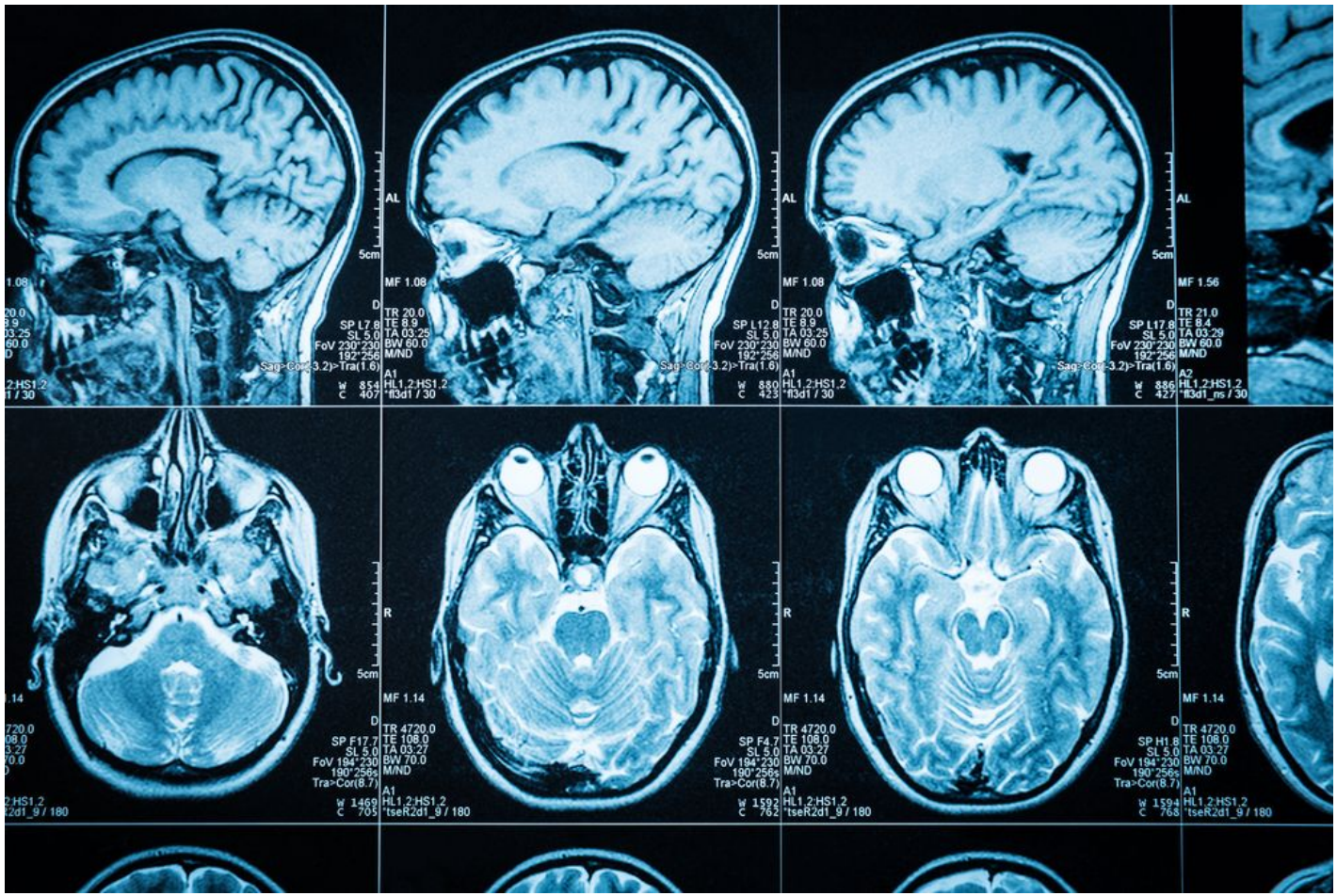
https://www.researchgate.net/publication/332452649

AI in Vision and Perception has numerous applications and benefits, including:

- **Automation:** AI systems can automate visual tasks, improving efficiency and accuracy in various industries.
- **Enhanced Decision Making:** AI algorithms can assist in decision-making processes by providing insights and analysis based on visual data.
- **Safety and Security:** AI-powered vision systems can enhance safety and security measures, such as surveillance and threat detection.
- **Medical Diagnosis:** AI in Vision and Perception can aid in medical diagnosis, helping detect diseases and anomalies in medical imaging.

Introduction to AI in Vision and Perception

Applications of Computer Vision in Medical Diagnosis



https://www.researchgate.net/publication/332452649

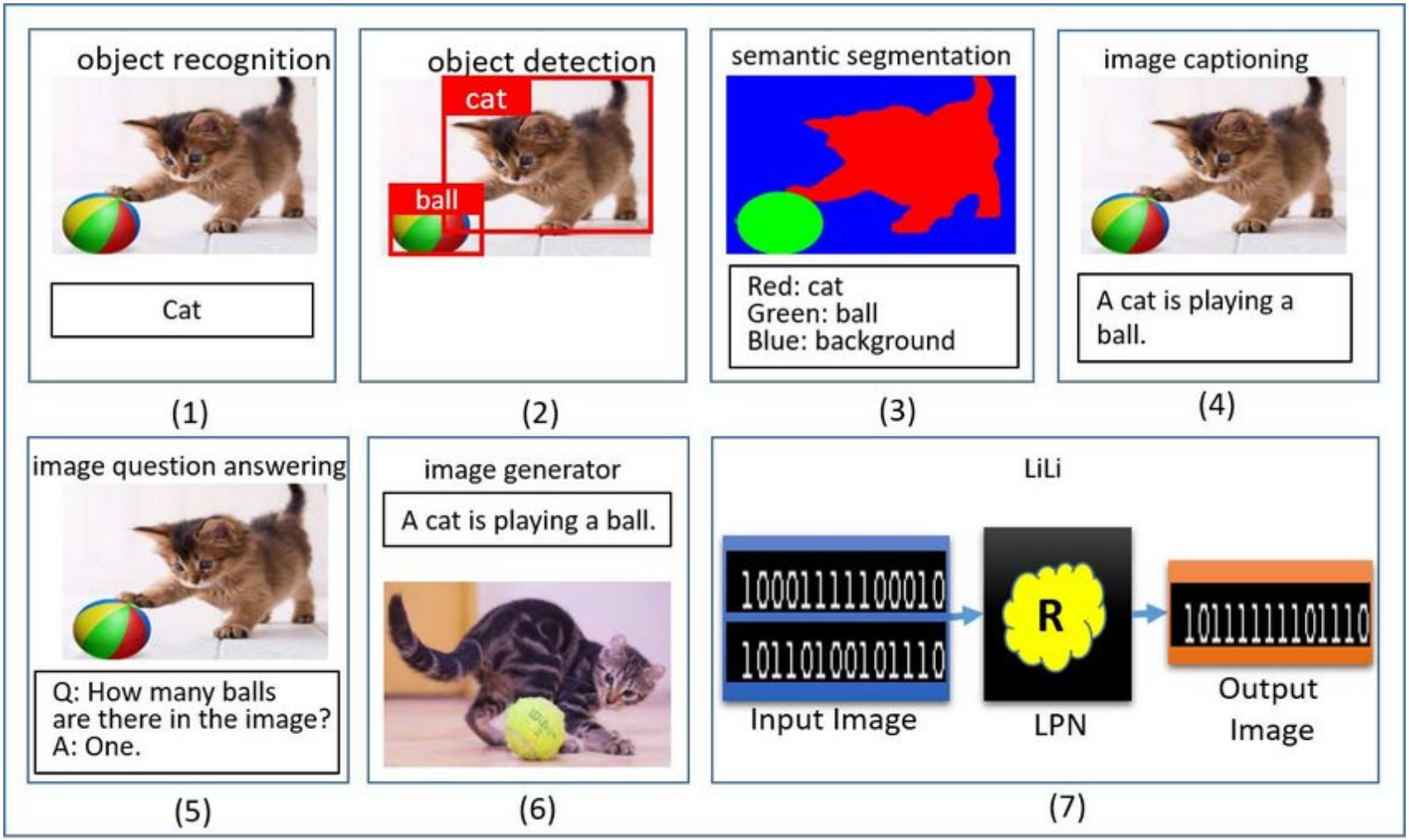
1. **Object Recognition and Detection:**
   - **Face Recognition:** Identify and verify individuals from images or video streams.
     - Example: Facial recognition systems used for biometric authentication in smartphones or surveillance systems.
   - **Object Detection:** Locate and classify objects of interest within images or video frames.
     - Example: Autonomous vehicles using object detection to identify pedestrians, vehicles, and traffic signs.

2. **Image Segmentation and Scene Understanding:**
   - **Semantic Segmentation:** Assign semantic labels to individual pixels or regions in an image.
     - Example: Medical imaging systems segmenting organs or tumors in CT scans for diagnosis and treatment planning.
   - **Scene Recognition:** Analyze complex scenes and infer their semantic meaning.
     - Example: Smart surveillance systems that automatically detect unusual activities or identify specific objects in a crowded scene.

3. **Image Captioning and Visual Question Answering:**
   - **Image Captioning:** Generate natural language descriptions for images.
     - Example: AI systems that provide captions for visually impaired users based on the content of images.
   - **Visual Question Answering (VQA):** Answer questions about images using both visual and textual information.
     - Example: AI assistants capable of understanding and answering questions based on images, such as "What breed of dog is in this picture?"

Introduction to AI in Vision and Perception

Six Popular Computer Vision Tasks



https://hackmd.io/@arkel23/computervision2021_starter

Introduction to AI in Vision and Perception

Types of AI in Vision and Perception

1. **Traditional AI:**
   - **Rule-based systems:** Use explicit rules and logical reasoning to process visual information.
     - Example: An expert system in medical imaging that applies predefined rules to analyze X-ray images for detecting abnormalities.
   - **Expert systems:** Employ domain-specific knowledge to analyze and interpret visual data.
     - Example: A computer-aided diagnostic system that uses expert knowledge in dermatology to identify skin lesions from images.
   - **Knowledge-based systems:** Represent and reason with knowledge to understand visual content.
     - Example: A system that utilizes a knowledge base of traffic rules and signs to recognize and interpret traffic signs in real-time video feeds.

2. **Machine Learning-based AI:**
   - **Supervised Learning:** Trained on labeled data to recognize and classify visual patterns.
     - Example: Image classification models trained on labeled images to classify objects, such as cats and dogs, in new images.
   - **Unsupervised Learning:** Discover hidden patterns and structures in unlabeled visual data.
     - Example: Clustering algorithms that group similar images together based on their visual features without any prior labels.
   - **Deep Learning:** Utilize deep neural networks to extract complex features and perform high-level visual tasks.
     - Example: Convolutional Neural Networks (CNNs) used for object detection and localization in images or videos.

Image perception refers to the cognitive process through which humans or computer systems interpret and understand visual information contained in images. It involves the ability to:

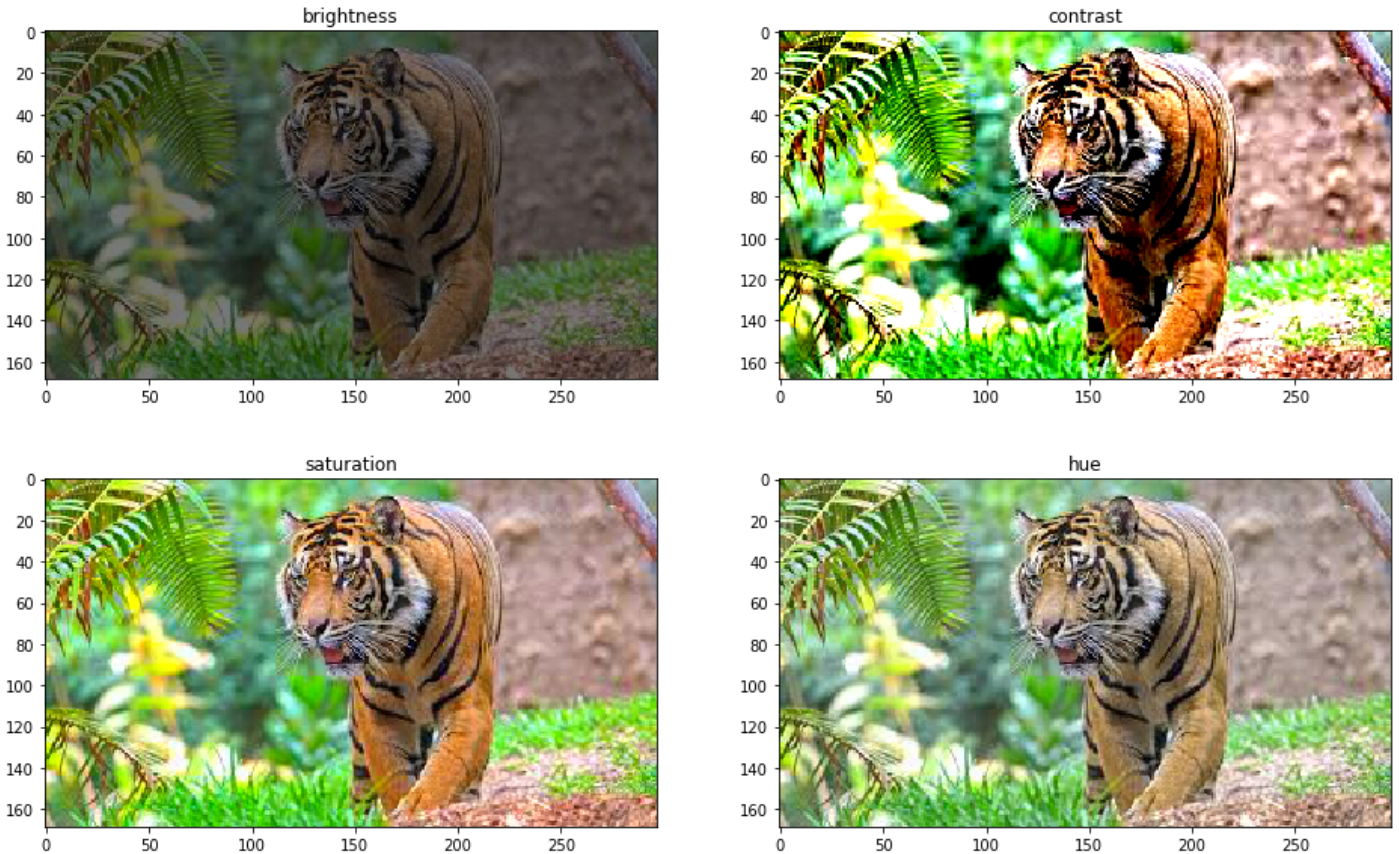
- Extract meaningful and relevant features, patterns, and structures from visual stimuli.
- Enable recognition, interpretation, and understanding of the content and context of images.
- Perform tasks such as object recognition, scene understanding, image segmentation, and visual reasoning.
- Utilize both symbolic and subsymbolic methods:
  - Symbolic methods: Representing visual knowledge using explicit rules, logical reasoning, and symbolic representations.
  - Subsymbolic methods: Employing machine learning techniques, such as neural networks, to learn representations directly from data without explicit rules.
- Extract low-level visual features (e.g., edges, colors, textures) and higher-level semantic information (e.g., objects, relationships, and context) from images.
- Integrate both bottom-up and top-down processing:
  - Bottom-up processing: Initial analysis of visual elements and features.
  - Top-down processing: Incorporation of prior knowledge, expectations, and contextual cues to interpret the visual information.
- Enable machines or computer systems to understand and interpret the visual world.

Image Processing: The application of algorithms and techniques to modify or enhance digital images.
Image Representation: Digital images are represented as a matrix of pixels, where each pixel contains color or intensity information.
Image Processing Operations: Common image processing operations include:

- **Image Filtering:** Modifying pixel values based on neighboring pixels using filters like blurring, sharpening, and noise reduction.
- **Image Enhancement:** Improving the visual quality of an image by adjusting contrast, brightness, and color balance.
- **Image Transformation:** Applying geometric transformations like rotation, scaling, and cropping.

Image Processing and Feature Extraction

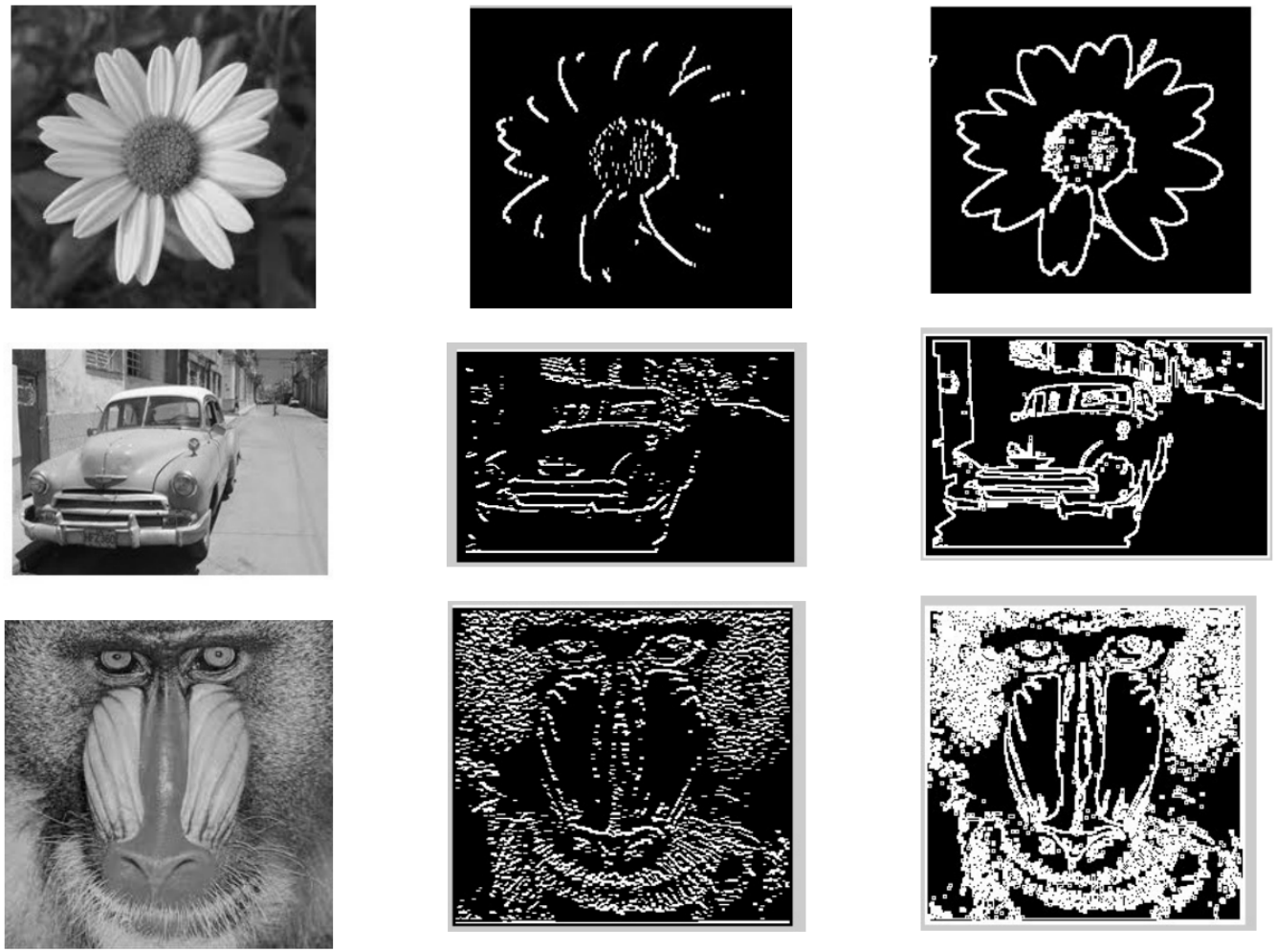
Image Preprocessing - An Example of Image Augmentation



https://www.v7labs.com/blog/data-augmentation-guide

Image Processing and Feature Extraction

Image Preprocessing - An Example of Edge Detection



https://aryamansharda.medium.com/how-image-edge-detection-works-b759baac01e2

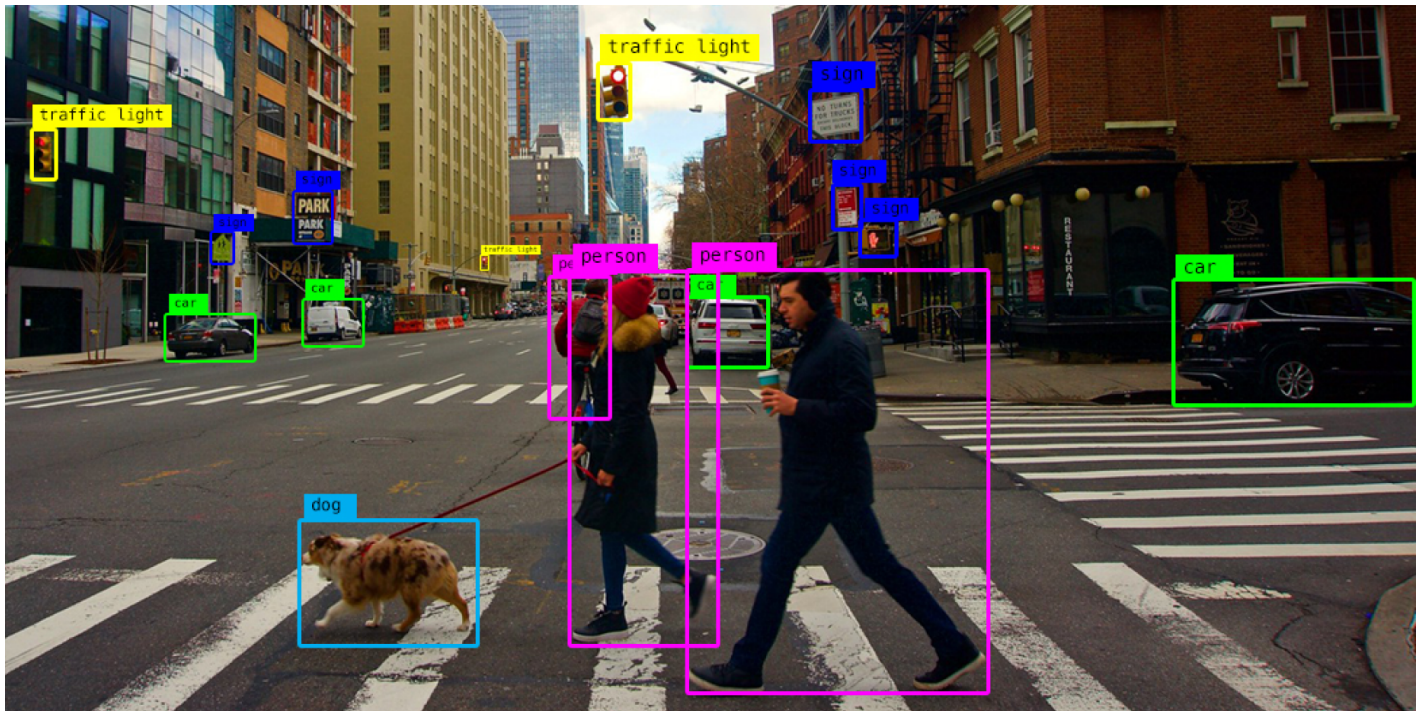
Image Processing and Feature Extraction

Image Preprocessing (cont.)

- **Image resizing:**
  - Adjust the size of images to a desired resolution or aspect ratio.
  - Helps in standardizing the input images and reducing computational complexity.
  - Example: Resizing high-resolution images to a smaller size for efficient processing in deep learning models.
- **Image augmentation:**
  - Generate new variations of images by applying transformations.
  - Helps in increasing the size of the training dataset and improving the model's robustness to variations in lighting, orientation, and scale.
  - Example: Randomly flipping, rotating, or cropping images to increase the diversity of the training dataset and improve model generalization.
- **Normalization:**
  - Standardize the pixel values of images to a common range.
  - Common techniques include mean normalization and min-max scaling.
  - Helps in reducing the influence of intensity variations and improving the convergence of learning algorithms.
  - Example: Scaling pixel values between 0 and 1 to ensure consistent input ranges for neural networks, enhancing convergence and performance.

Image Processing - Object Detection and Recognition

Object detection an Example



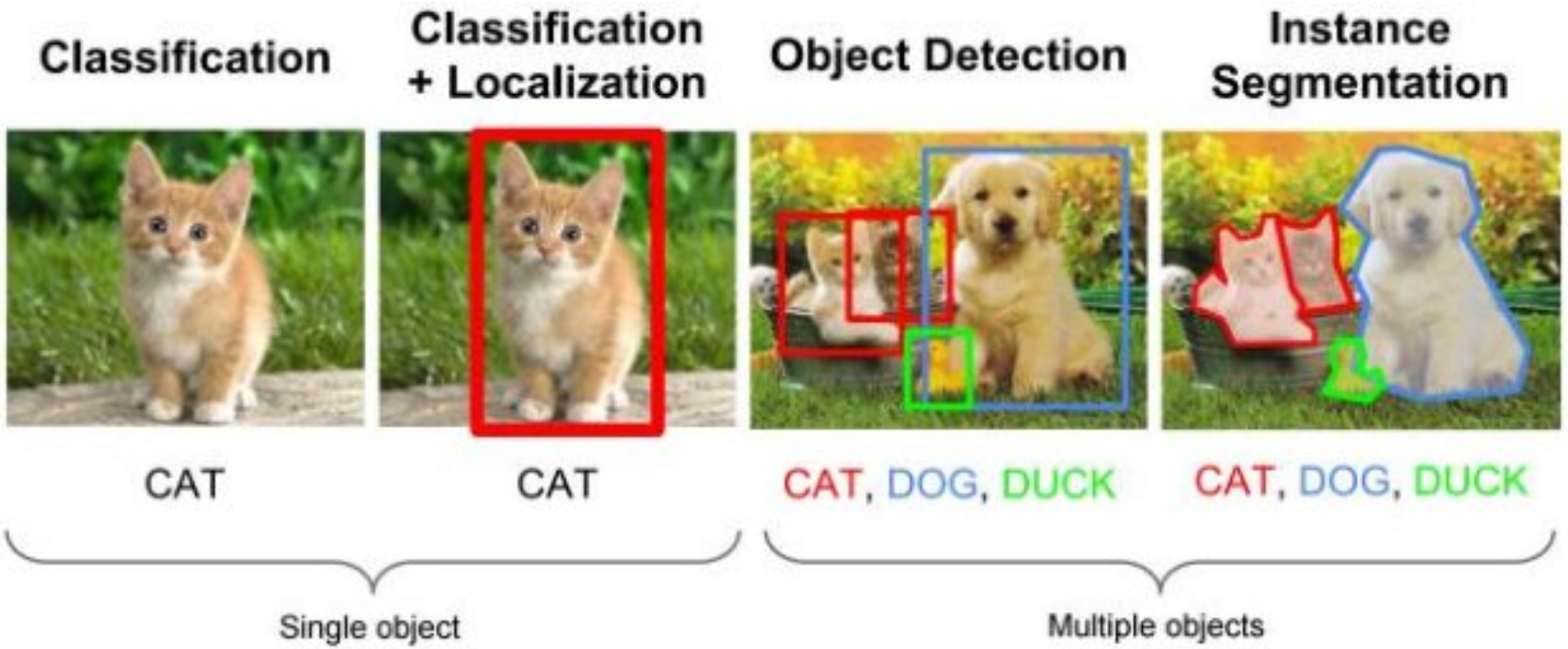
https://alwaysai.co/blog/object-detection-for-businesses

Object detection and recognition is a fundamental task in computer vision that involves identifying and localizing objects of interest within images or videos.

- Object Detection:
  - Object detection aims to locate and classify multiple objects within an image.
  - It provides bounding box coordinates and class labels for each detected object.
- Object Recognition:
  - Object recognition focuses on identifying and classifying objects without providing precise location information.
  - It assigns a class label to the entire image or individual objects within the image.

Image Processing - Segmentation

Difference Between Classification Localization Detection and Segmentation



https://www.researchgate.net/publication/355467756

- Definition:
  - Image segmentation aims to partition an image into multiple regions based on similarity criteria.
  - It assigns a unique label or identifier to each region, enabling the separation of objects and background or different objects from each other.

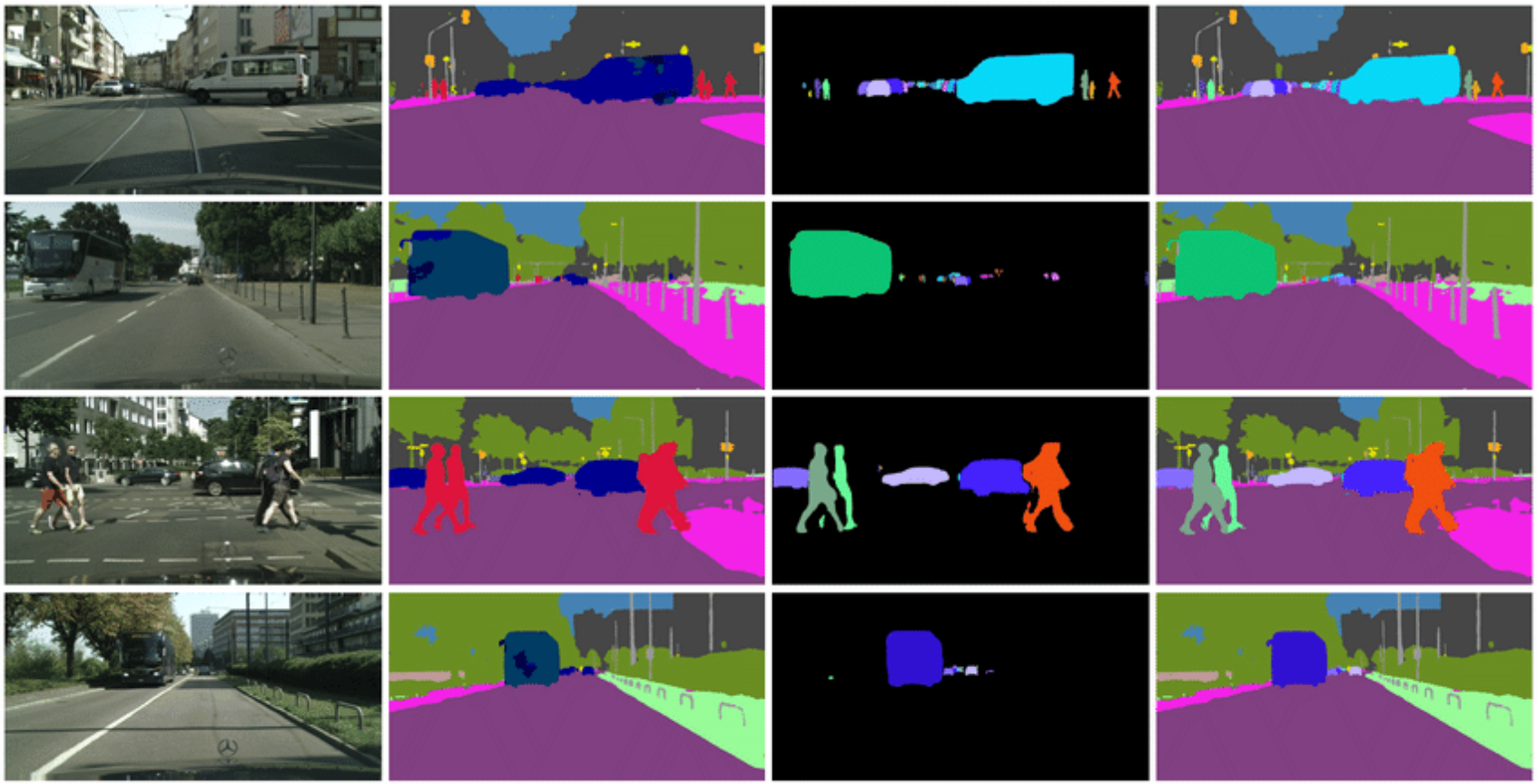
Image segmentation plays a crucial role in various applications, such as object recognition, scene understanding, medical imaging, and autonomous driving.

Image segmentation can be categorized into two main types: semantic segmentation and instance segmentation.

- Semantic Segmentation:
  - Semantic segmentation assigns a class label to each pixel or region in the image.
  - It aims to identify and label different objects or regions based on their semantic meaning.
  - For example, in a street scene, semantic segmentation can label pixels as road, pedestrians, cars, buildings, etc.
- Instance Segmentation:
  - Instance segmentation goes a step further and provides a unique label for each individual object instance within the image.
  - It not only differentiates between object classes but also separates different instances of the same class.
  - For example, in a crowded image, instance segmentation can distinguish between different people, cars, or other objects.

Image Processing - Segmentation

Panoptic Segmentation by Unifying Semantic and Instance Segmentation



https://www.researchgate.net/publication/329616112

Object tracking is a crucial task in computer vision that involves estimating the position and trajectory of an object of interest in a sequence of images or videos.

- Definition:
  - Object tracking aims to locate and follow a specific object over time, even as its appearance and position may change.
  - It involves associating the target object across consecutive frames, accounting for occlusions, scale changes, and other challenges.

Object tracking plays a vital role in various applications, such as surveillance, autonomous driving, video analysis, and augmented reality.

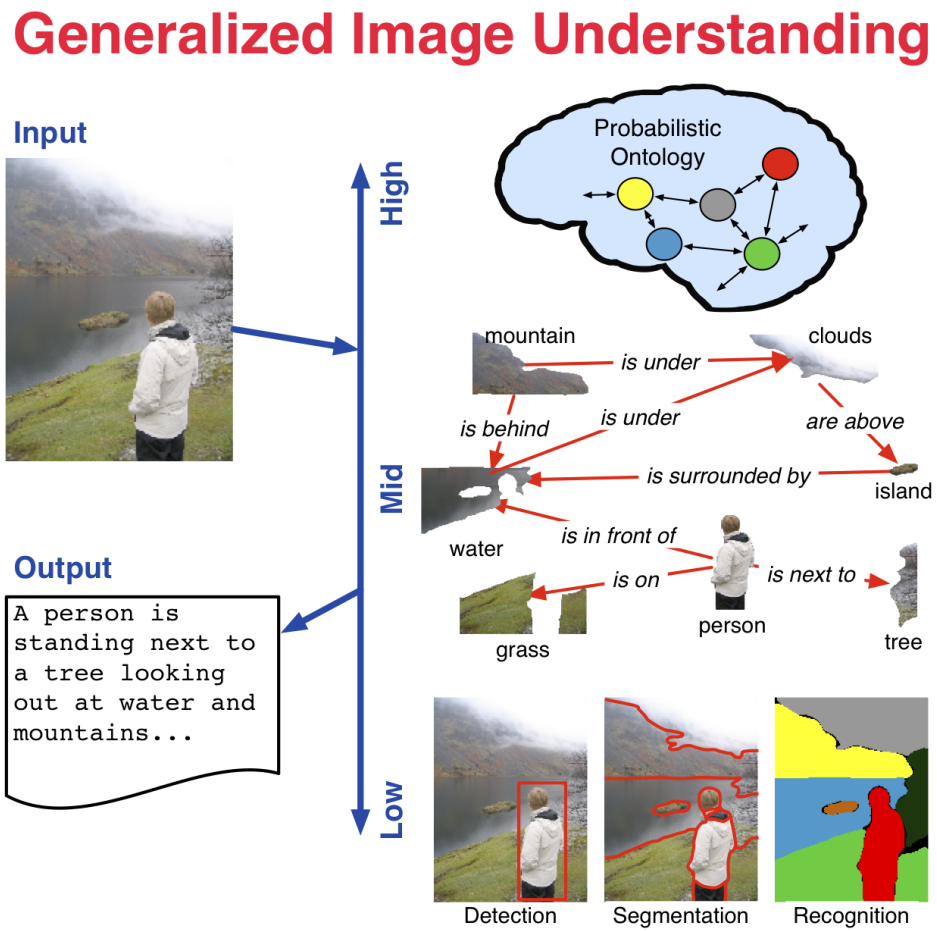
Image Processing - Object Tracking

Video Example of AI Methods in Object Tracking

Scene Understanding involves analyzing the context of a scene to infer higher-level information, such as object relationships, scene semantics, and contextual dependencies.

- AI methods, including deep learning and probabilistic models, enable scene understanding by capturing the complex relationships between objects and their context.
- Contextual analysis considers factors like object co-occurrence, spatial layout, and semantic relationships to derive meaningful insights from the scene.
- Scene understanding finds applications in image captioning, scene generation, and intelligent surveillance systems.

By comprehensively analyzing the scene context, AI algorithms can extract rich information and facilitate more advanced computer vision tasks.



https://www.hindawi.com/journals/complexity/2021/5541134/

Supervised Learning Algorithms:

- **Convolutional Neural Networks (CNNs):** Highly effective for image classification, object detection, and semantic segmentation tasks. They use convolutional layers to extract hierarchical features from images.
- **Support Vector Machines (SVMs):** Well-suited for classification tasks, SVMs aim to find an optimal hyperplane that separates different classes in feature space.
- **Random Forests (RFs):** Ensemble learning method that combines multiple decision trees to make predictions. RFs are useful for tasks like object recognition and feature selection.
- **Gradient Boosting Machines (GBMs):** Boosting algorithms that create an ensemble of weak learners, such as decision trees, to improve prediction accuracy. GBMs, like XGBoost and LightGBM, are commonly used for various computer vision tasks.

Unsupervised Learning Algorithms:

- **Clustering Algorithms:** Unsupervised algorithms like k-means, hierarchical clustering, and DBSCAN group similar data points together based on their features. These algorithms can be applied to image segmentation tasks.
- **Generative Adversarial Networks (GANs):** Consist of a generator and a discriminator network that compete against each other. GANs can generate realistic images, perform image-to-image translation, and learn meaningful representations.
- **Autoencoders:** Neural network architectures that aim to reconstruct the input data, forcing the model to learn compressed and meaningful representations. Autoencoders are useful for dimensionality reduction and anomaly detection.

Semi-Supervised and Transfer Learning:

- **Semi-Supervised Learning:** Combines labeled and unlabeled data to train models. Techniques like self-training, co-training, and multi-view learning can leverage unlabeled data to improve performance.
- **Transfer Learning:** Involves pretraining a model on a large dataset and then fine-tuning it on a smaller dataset for a specific task. Transfer learning enables models to leverage knowledge learned from one task to improve performance on another related task.

Conclusion:

- Machine learning algorithms, both supervised and unsupervised, play a crucial role in solving various computer vision tasks.
- Understanding the characteristics and capabilities of different algorithms is essential for selecting the most appropriate approach for specific computer vision problems.

Deep Learning for Computer Vision

Neural Networks Types and Their Applications (selected examples)

Neural Network Type	Application in Computer Vision
Convolutional Neural Network (CNN)	Image Classification, Object Detection Semantic Segmentation, Facial Expression Recognition Medical Image Analysis, Image Super-Resolution
Recurrent Neural Network (RNN)	Image Captioning, Video Analysis OCR, Handwriting Recognition Human Action Recognition
Generative Adversarial Network (GAN)	Image Generation Image-to-Image Translation Super-Resolution Imaging Style Transfer Data Augmentation Anomaly Detection
Siamese Neural Network	Face Recognition Person Re-Identification Signature Verification Similarity Matching Biometric Identification
Transformer Network	Object Detection Image Captioning Visual Question Answering Video Understanding Autonomous Driving
You Only Look Once (YOLO)	Real-time Object Detection Vehicle and Pedestrian Detection Object Tracking
Mask R-CNN	Instance Segmentation Object Detection with Masking Human Pose Estimation

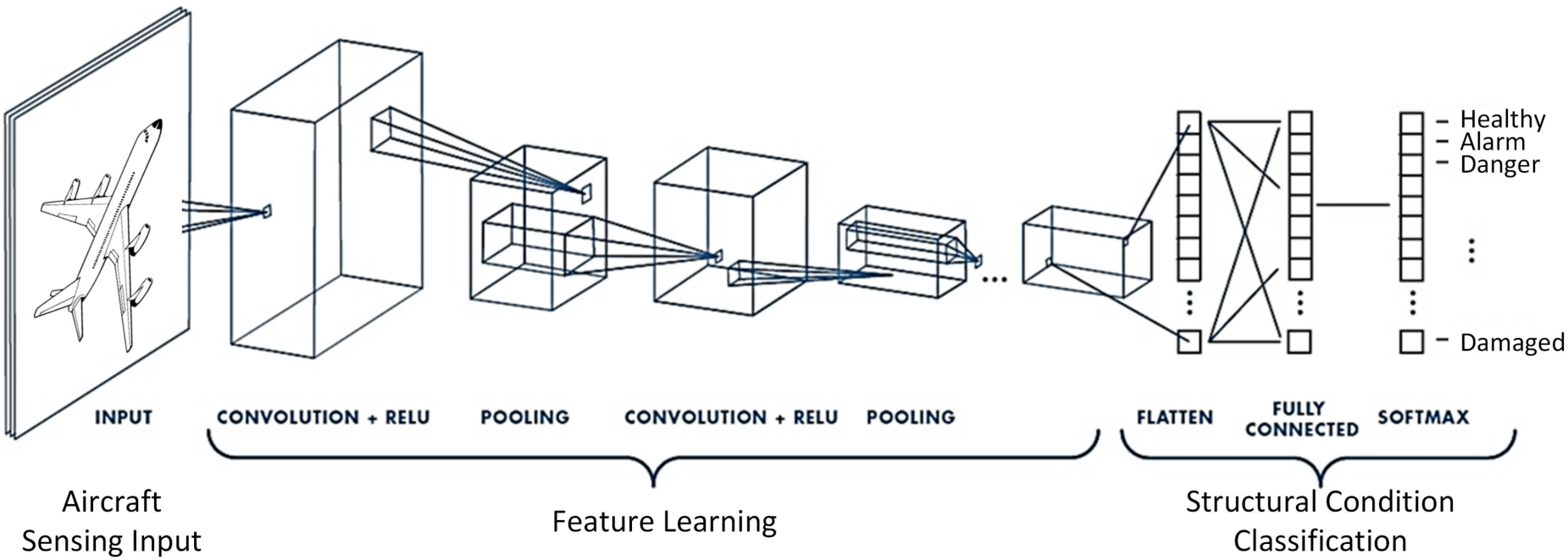
Convolutional Neural Networks (CNNs) are a type of deep learning model specifically designed for processing visual data.

- Definition of CNN:
  - CNNs are composed of multiple layers of interconnected neurons, including convolutional, pooling, and fully connected layers.
  - They employ convolutional operations to extract features from input images, enabling automatic feature learning.

CNNs have revolutionized various computer vision tasks, including image classification, object detection, and image segmentation.

Deep Learning for Computer Vision - CNN

Convolutional Neural Networks (CNNs) (cont.)



https://www.mdpi.com/1424-8220/19/22/4933

The architecture of a CNN typically consists of several key components:

- Convolutional Layers:
  - Convolutional layers perform the main feature extraction step by applying convolutional filters to input images.
  - These filters capture spatial patterns and local dependencies.
- Pooling Layers:
  - Pooling layers downsample feature maps, reducing the spatial dimensions and computational complexity.
  - Common pooling operations include max pooling and average pooling.
- Fully Connected Layers:
  - Fully connected layers take the extracted features and map them to the desired output classes or predictions.
  - They enable high-level reasoning and decision-making based on the learned features.

This architecture allows CNNs to capture hierarchical and translation-invariant features in visual data.

Several CNN architectures have been proposed, each with its unique characteristics and performance:

- VGG (Visual Geometry Group):
  - VGG is known for its simplicity and uniform architecture.
  - It consists of multiple stacked convolutional layers, followed by fully connected layers.
- ResNet (Residual Network):
  - ResNet introduced residual connections to address the degradation problem in deep networks.
  - It allows the gradient flow through the network, enabling the training of very deep models.
- Inception:
  - Inception modules incorporate multiple parallel convolutional operations at different scales.
  - They capture multi-scale features efficiently, leading to improved performance.

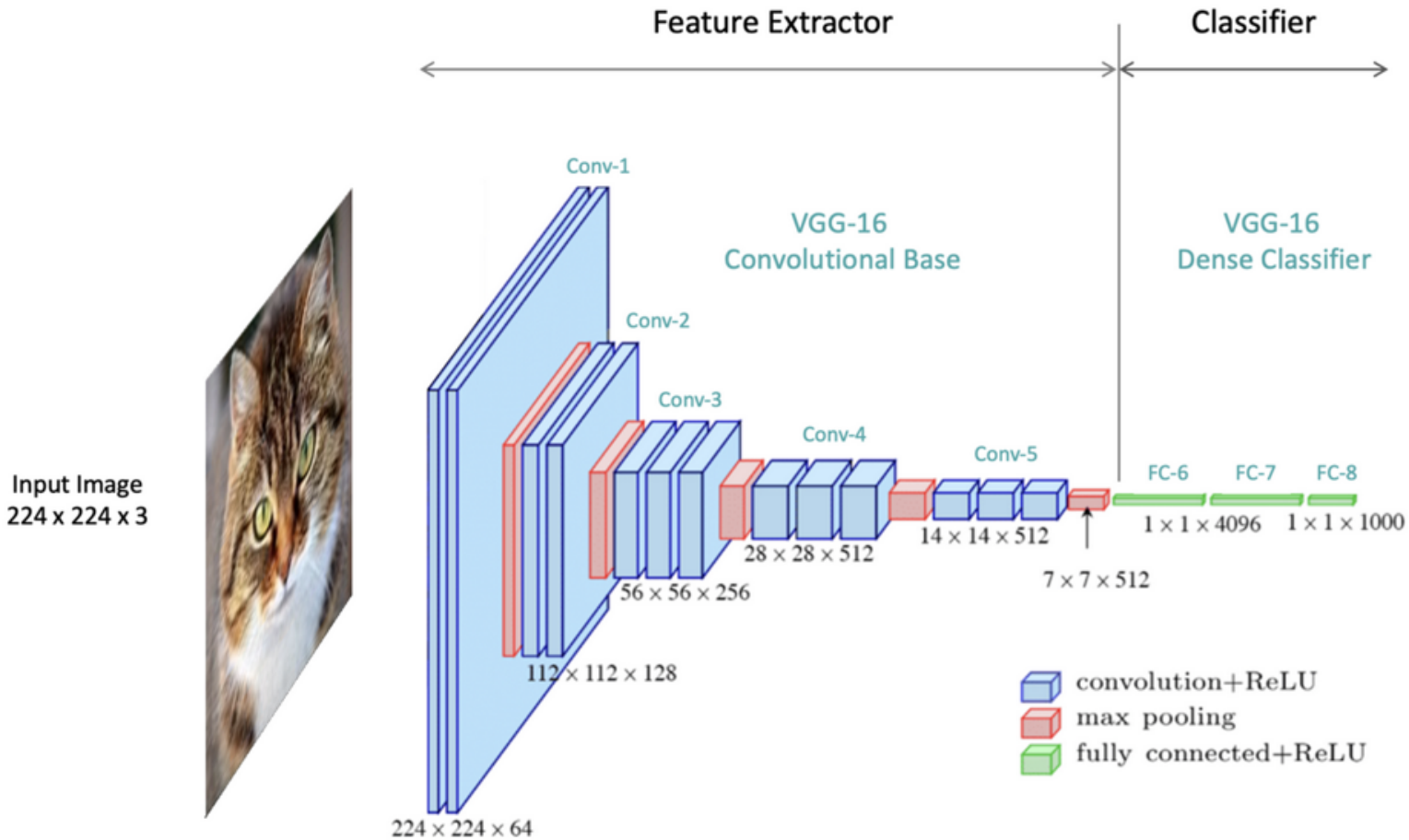
Deep Learning for Computer Vision

VGG (Visual Geometry Group) Architectures: Deep Learning Models in Computer Vision

- The VGG architectures are deep convolutional neural network models known for their simplicity and effectiveness in image classification tasks.
- Developed by the Visual Geometry Group at the University of Oxford.
- Key features of VGG architectures:
  - Sequential Structure: Consists of a series of convolutional layers, followed by fully connected layers for classification.
  - Small Convolutional Filters: Uses 3x3 convolutional filters throughout the network, which allows for deeper networks with fewer parameters.
  - Pooling Layers: Employs max pooling layers to reduce spatial dimensions and extract robust features.
  - Multiple Architectures: VGG16 (16 layers) and VGG19 (19 layers) are commonly used variants.
- VGG architectures prioritize depth over width, resulting in highly expressive and discriminative features.
- Transfer Learning: Pretrained VGG models on large datasets like ImageNet can be used as feature extractors or fine-tuned for specific tasks.
- Benefits of VGG architectures:
  - Strong performance in image classification tasks, especially for fine-grained recognition.
  - Simplicity and easy reproducibility.
  - Robust feature extraction capabilities.
- VGG architectures have been widely adopted as baselines and benchmarks in Computer Vision research.
- Applications of VGG architectures:
  - Image classification and recognition tasks.
  - Object detection and localization.
  - Image style transfer and generation.
  - Medical image analysis and diagnosis.

Deep Learning for Computer Vision - CNN

CNN VGG architecture

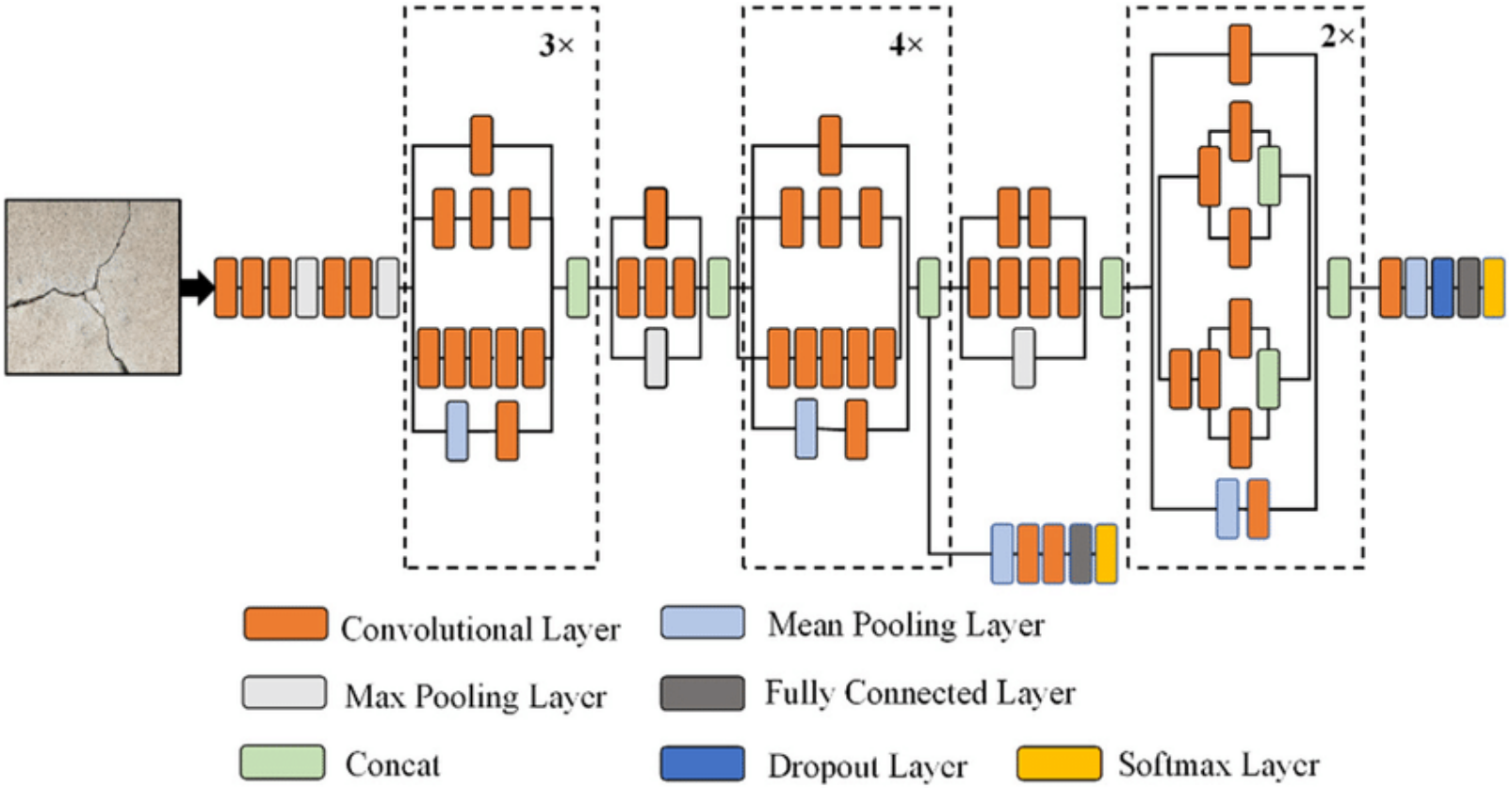


https://www.mdpi.com/1424-8220/19/22/4933

- The Inception architectures, also known as GoogLeNet, are deep convolutional neural network models designed to achieve high accuracy while maintaining efficiency.
- Key features of Inception architectures:
  - Inception Module: Introduced the idea of parallel convolutional filters of different sizes within a single layer.
  - Dimensionality Reduction: Utilizes 1x1 convolutions to reduce the number of input channels and improve computational efficiency.
  - Multi-Scale Feature Extraction: Extracts features at different spatial scales by incorporating parallel convolutions with different filter sizes.
  - Auxiliary Classifiers: Adds auxiliary classifiers during training to provide additional gradients and combat the vanishing gradient problem.
- Inception variants:
  - Inception-v1 (GoogLeNet): The original Inception architecture, with multiple inception modules and auxiliary classifiers.
  - Inception-v2 and Inception-v3: Improved versions with additional optimizations, such as factorized convolutions and batch normalization.
  - Inception-v4: Further improved version with increased depth and reduced computational cost.
- Benefits of Inception architectures:
  - Effective balance between accuracy and computational efficiency.
  - Improved gradient flow and reduced vanishing/exploding gradient problem.
  - Multi-scale feature extraction for capturing both fine-grained and high-level features.
- Applications of Inception architectures:
  - Large-scale image recognition competitions (e.g., ImageNet challenge).
  - Fine-grained object classification.
  - Visual question answering.
  - Medical image analysis and diagnosis.

Deep Learning for Computer Vision - CNN

CNN Inception architecture



https://www.researchgate.net/publication/349717475

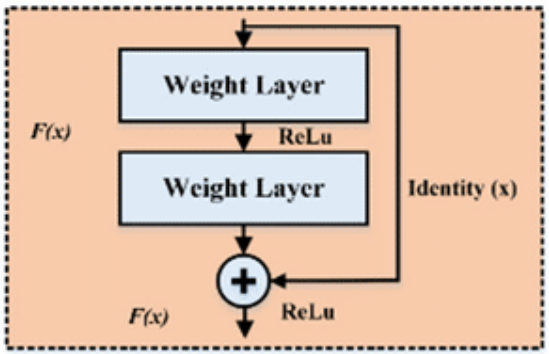
Deep Learning for Computer Vision

ResNet (Residual Neural Network): Deep Learning in Computer Vision

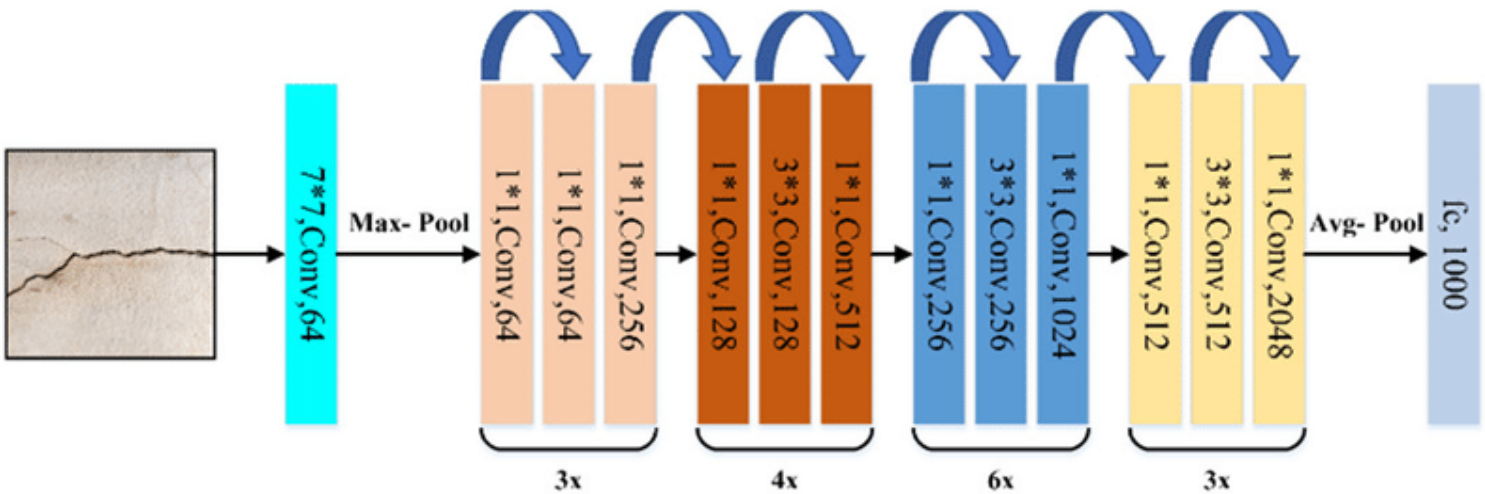
- ResNet is a deep neural network architecture that addresses the problem of vanishing gradients during training of very deep networks.
- It introduces residual connections, allowing the network to learn residual mappings rather than directly learning the desired underlying mapping.
- Key features of ResNet:
  - Residual Blocks: Consist of multiple convolutional layers with shortcut connections that bypass some of the layers.
  - Skip Connections: Enable the network to propagate gradients more effectively and alleviate the vanishing gradient problem.
  - Identity Mapping: Allows the network to learn the residual mapping by simply learning the difference between the input and output.
  - Deep Stacking: ResNet architectures can be significantly deeper (e.g., 50, 101, 152 layers) than traditional networks without compromising performance.
- Pretrained Models: Pretrained ResNet models on large-scale datasets (e.g., ImageNet) are widely available, enabling transfer learning for various computer vision tasks.
- ResNet Variants: Different variants of ResNet have been proposed, including ResNet-18, ResNet-50, ResNet-101, and ResNet-152, with varying depths and complexities.
- Applications of ResNet:
  - Image classification and recognition tasks.
  - Object detection and localization.
  - Semantic segmentation.
  - Image super-resolution.
  - Visual question answering.

Deep Learning for Computer Vision - CNN

CNN Resnet architecture



Residual Learning Block



https://www.researchgate.net/publication/349717475

Deep Learning for Computer Vision - RNN

Application of RNN Architecture in Computer Vision: Video Captioning

- Video captioning is an application of RNN architecture in Computer Vision that aims to generate textual descriptions of videos.
- Problem Statement:
  - Given a video sequence, the task is to automatically generate a descriptive caption that accurately represents the content and context of the video.
  - This involves understanding and summarizing the visual information over time and generating coherent and meaningful sentences.
- Architecture Overview:
  - RNN-based architectures, such as LSTM (Long Short-Term Memory), are commonly used for video captioning tasks.
  - The video frames are first fed into a pre-trained convolutional neural network (CNN) to extract visual features from each frame.
  - The extracted features are then input to the RNN, which sequentially processes them along with the previously generated words.
  - At each time step, the RNN generates a word based on the input features and the hidden state, which represents the understanding of the video content so far.
  - The process continues until an end-of-sequence token is generated or a maximum caption length is reached.
- Training Process:
  - Training data consists of video-caption pairs, where human-generated captions are used as ground truth.
  - During training, the RNN is trained to minimize the discrepancy between the generated caption and the ground truth caption.
  - This is typically done using techniques like teacher forcing, where the ground truth caption is used as input during training.
- Evaluation and Metrics:
  - The quality of generated captions is evaluated using metrics like BLEU (Bilingual Evaluation Understudy), METEOR, and CIDEr, which measure the similarity between generated and reference captions.
  - Human evaluation, involving subjective assessment by human judges, is also performed to gauge the quality and coherence of the generated captions.

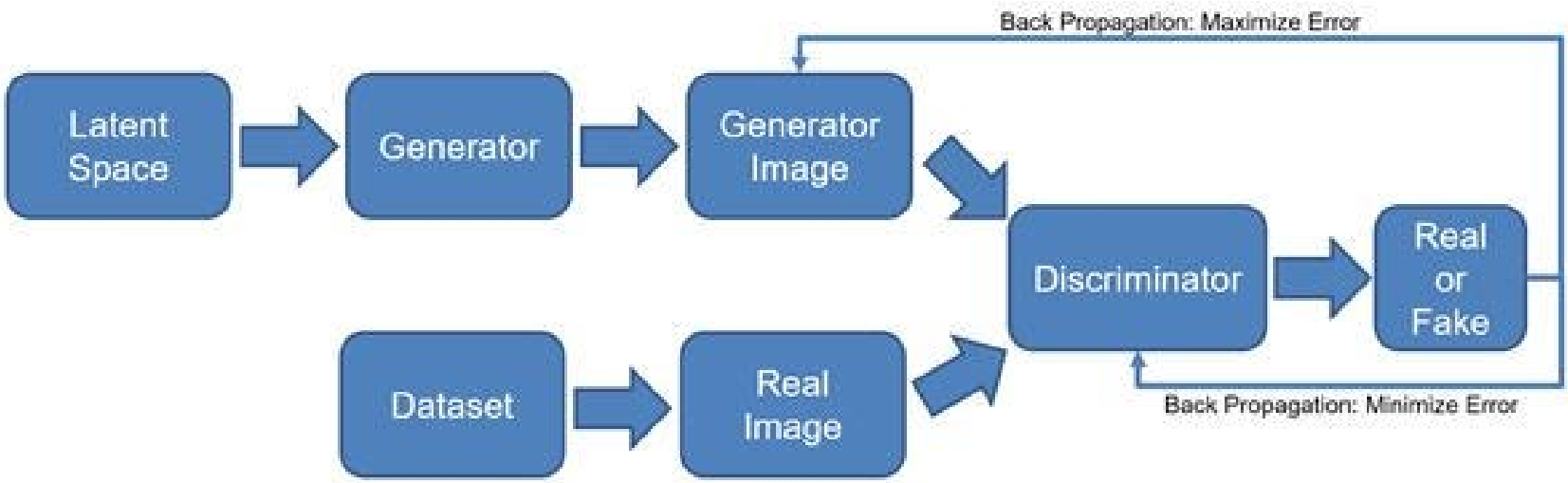
Deep Learning for Computer Vision - GAN

Image Synthesis with GANs

GANs have been successfully used for image synthesis tasks, where the generator network learns to generate realistic images from random noise as input.

- The generator network typically consists of convolutional layers that upsample the noise vector to generate an image.
- The discriminator network is trained to distinguish between real images from the training dataset and the fake images generated by the generator.
- Through the adversarial training process, the generator gradually improves its ability to generate high-quality images that resemble the training data.

Example: Generating realistic human faces using GANs, as seen in projects like StyleGAN and Deepfake technology.



https://medium.com/@Packt_Pub/inside_the_generative_adversarial_networks_gan_architecture_2435afbd6b3b

Visual Perception and Human-Computer Interaction (HCI) involve the interaction between humans and machines through visual information. AI methods play a crucial role in enhancing the capabilities and improving the user experience in these domains.

- Introduction to Visual Perception and HCI:
  - Visual Perception studies how humans interpret and understand visual information.
  - HCI focuses on designing systems that facilitate human-computer interaction, leveraging visual interfaces.

Gesture Recognition and Tracking involve understanding and interpreting human gestures, enabling natural and intuitive interaction with machines.

- Gesture Recognition:
  - AI methods, such as deep learning, enable accurate recognition and classification of hand gestures.
  - These techniques are used in applications like sign language translation, virtual reality control, and human-robot interaction.
- Gesture Tracking:
  - AI algorithms can track and follow the movement of human hands or body parts in real-time.
  - This enables applications like motion-based gaming, interactive displays, and augmented reality.

Facial Recognition and Emotion Detection technologies have gained significant attention in recent years, with numerous applications in security, entertainment, and user experience enhancement.

- Facial Recognition:
  - AI-based facial recognition systems can identify and verify individuals based on their facial features.
  - These systems have applications in access control, surveillance, and identity verification.
- Emotion Detection:
  - AI algorithms can analyze facial expressions to detect and recognize emotions.
  - This is used in applications like sentiment analysis, user experience evaluation, and affective computing.

Examples of AI methods used in Gesture Recognition:

- Convolutional Neural Networks (CNNs):
  - CNNs have shown excellent performance in recognizing hand gestures from images or video frames.
  - They can classify a wide range of gestures and enable real-time gesture recognition.
- Recurrent Neural Networks (RNNs):
  - RNNs are suitable for capturing the temporal dynamics of gesture sequences.
  - They can model the sequential nature of gestures, allowing for gesture prediction and tracking.
- Depth Sensing Technologies:
  - Depth sensors, such as Microsoft Kinect or LiDAR, provide 3D information for precise gesture recognition.
  - AI algorithms can analyze the depth data to recognize intricate hand movements and gestures.

Examples of AI methods used in Facial Recognition:

- Deep Face Recognition Models:
  - Deep learning models, such as FaceNet and VGGFace, have achieved remarkable accuracy in face recognition tasks.
  - They can learn discriminative features from facial images, enabling reliable face identification.
- Facial Expression Analysis:
  - AI algorithms can analyze facial expressions to detect emotions like happiness, sadness, or anger.
  - They use techniques like facial landmark detection, feature extraction, and machine learning to classify emotions.
- Generative Adversarial Networks (GANs):
  - GANs can generate realistic and high-resolution facial images, enabling synthetic data augmentation.
  - This helps in training robust facial recognition models and addressing data scarcity challenges.

- **Recurrent Neural Networks (RNNs)**
  - Used for processing and generating sequences of text and video frames.
  - Variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) improve long-term dependency learning.
- **Convolutional Neural Networks (CNNs)**
  - Effective for spatial feature extraction in video frames.
  - Used in combination with RNNs for spatio-temporal processing.
- **Generative Adversarial Networks (GANs)**
  - Consist of a generator and a discriminator network.
  - Useful for generating realistic video frames from text descriptions.
- **Transformer Networks**
  - Handle sequential data with self-attention mechanisms.
  - Effective for capturing long-range dependencies in text and video sequences.
- **Variational Autoencoders (VAEs)**
  - Used for encoding text into latent representations.
  - Generate diverse and coherent video frames from these representations.
- **Diffusion Models**
  - Model the data distribution by iteratively adding and removing noise.
  - Effective for generating high-quality video frames from text descriptions.

Computer graphics - GenAI applications

An example of GenAI - a rivalry between AI and humans in terms of creativity?

https://cyfrowa.rp.pl/technologie/art38343551-bezczelna-malpa-sztuczna-inteligencji-pokonala-zawodowych-fotografow

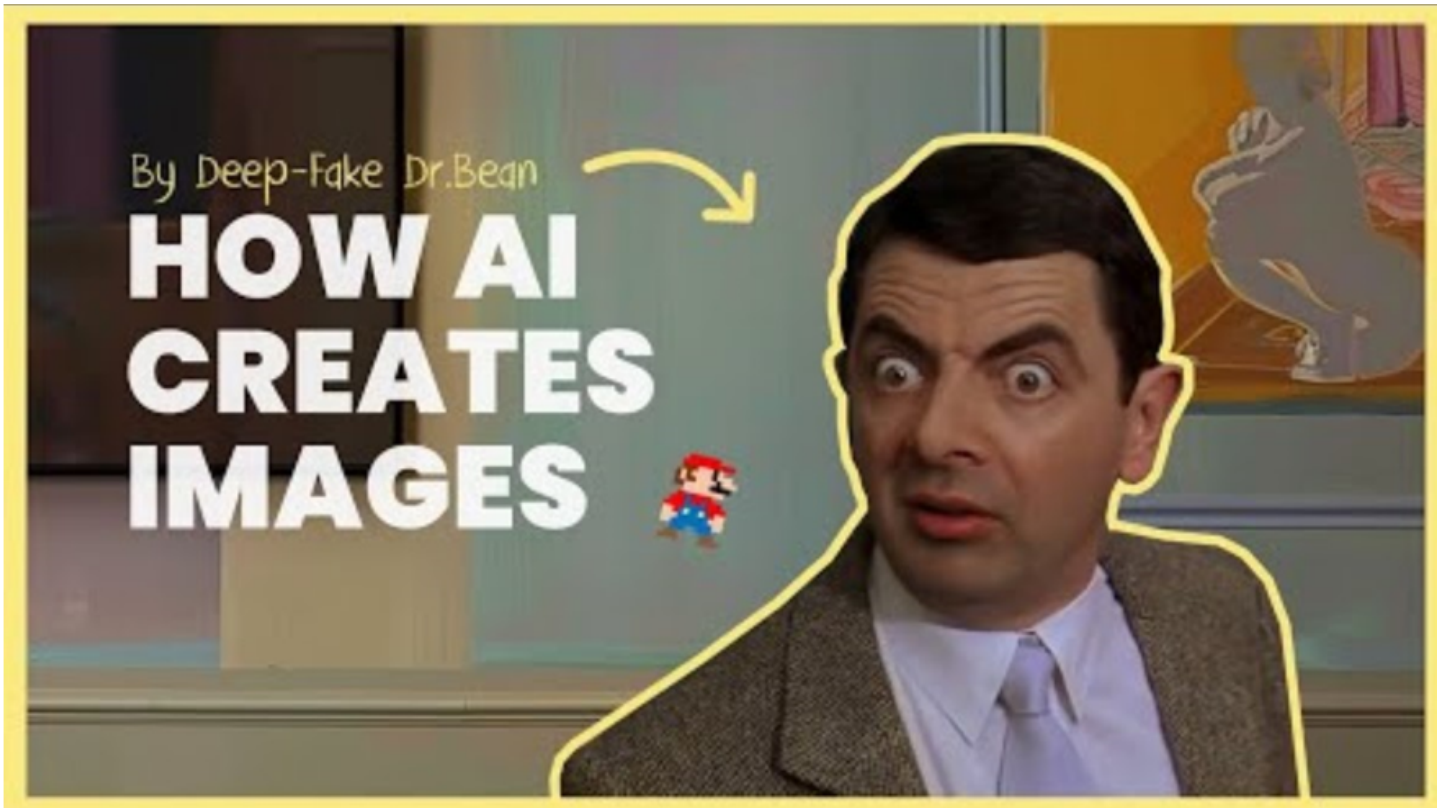
Computer graphics - GenAI applications

GenAI example - TexttoVideo creation

https://www.youtube.com/watch?v=vWZEb6DbSxs

Computer graphics - GenAI applications

Video Example of Deep Fake



https://www.youtube.com/watch?v=k357YPzJqgA

Computer graphics - GenAI applications

Video Example of Deep Fake

https://www.youtube.com/watch?v=vxro94cJJxM

Summary - AI in Computer Vision

1. Introduction to AI in Vision and Perception
   - Applications of AI in Vision and Perception
   - Types of AI in Vision and Perception
2. Image Processing and Feature Extraction
   - Basics of Image Processing and Feature Extraction
   - Object Detection and Recognition
   - Image Segmentation
   - Object Tracking
   - ...
3. Machine Learning for Computer Vision
   - Overview of Machine Learning Algorithms for Computer Vision
   - CNN
   - RNN
   - GAN
   - ...
4. Visual Perception and Human-Computer Interaction
5. Computer graphics
6. Challenges and Future Directions in AI for Vision and Perception
7. References

References

1. Richard Szeliski, "Computer Vision. Algorithms and Applications", Springer Int. Publ., 2022
2. S. J. Russell, P. Norvig, "Artificial Intelligence: A Modern Approach", Financial Times Prentice Hall, 2019.
3. M. Flasiński, "Introduction to Artificial Intelligence", Springer Verlang, 2016
4. M. Muraszkiewicz, R. Nowak (ed.), "Sztuczna Inteligencja dla inżynierów", Oficyna Wydawnicza PW, 2022
5. J. Prateek , "Artificial Intelligence with Python", Packt 2017