Introduction to Computer Vision

> *"A picture is worth a thousand words — but in computer vision, it's also millions of pixels that need to be understood."*

An image is an important source of information because it contains rich visual details that words or numbers alone cannot capture.

Computer vision is the field of AI that enables machines to process and understand these images, transforming raw pixel data into meaningful features such as edges, shapes, textures, or entire objects. Traditionally, models like Convolutional Neural Networks (CNNs) dominated this field by exploiting spatial structure, but the introduction of **Transformers** has shifted the paradigm.

But

> *"The eye should learn to listen before it looks."* — Robert Frank

This quote works beautifully in context: **vision** is not just about seeing pixels, but about **understanding meaning and relationships** — exactly what Transformers bring to computer vision.

Major Computer Vision Tasks

These are the **core building blocks** in computer vision.

Task	What it does	Example
Image Classification	Assigns a single/multiple label to the whole image	Cat vs Dog vs Horse
Object Detection	Identifies & localizes objects with bounding boxes	Cars, pedestrians, traffic lights
Image Segmentation	Labels each pixel with a class (one mask per class)	Road, sidewalk, building pixels
Instance Segmentation	Labels each object separately, even within the same class	Two overlapping dogs → two masks
Image-to-Image Translation	Transforms an image from one style/domain to another	Day → Night, photo → Van Gogh
Image Generation	Creates entirely new images (often text-guided)	"Gothic castle under a full moon"

Task	What it does	Example
Image Captioning	Generates a natural-language description	“Woman in black dress dancing under a spotlight”
Visual Question Answering (VQA)	Answers questions about an image	Q: Instrument? → A: “Cello”
Image Retrieval / Similarity Search	Finds visually or semantically similar images	Retrieve similar handbag designs

Key Applications of Computer Vision

These tasks combine to power **real-world systems** across industries:

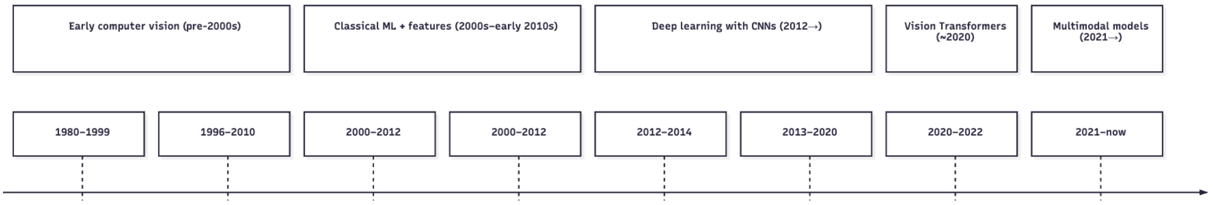
1. **Image Recognition & Classification** – facial recognition, disease detection, product recognition. *Examples: Google Vision AI, Zebra Medical*.
2. **Image Generation & Synthesis** – generative art, deepfakes, enhancement. *Examples: DALL·E, Adobe Firefly*.
3. **Video Analysis** – action recognition, summarization, AR/VR streaming. *Examples: TikTok effects, YouTube video summaries*.
4. **Autonomous Systems & Robotics** – self-driving cars, drones, warehouse robots. *Examples: Waymo, Boston Dynamics*.
5. **AR/VR** – filters, VR world building, gesture recognition. *Examples: Meta AR glasses, Apple Vision Pro*.
6. **Retail & E-Commerce** – visual search, stock tracking, personalized recommendations. *Examples: Google Lens, Walmart AI inventory*.
7. **Security & Surveillance** – anomaly detection, license plate recognition, crowd analysis. *Examples: Hikvision smart cameras*.
8. **Agriculture & Environment** – crop monitoring, wildlife tracking, deforestation analysis. *Examples: John Deere precision farming, WWF monitoring*.
9. **Creative Media** – style transfer, auto-editing, motion capture. *Examples: Runway ML, Pixar tools*.

Examples of Applications in Social Sciences

- **Facial Expression & Emotion Analysis** → studying political speeches, protest dynamics, or classroom engagement.
- **Crowd & Mobility Analysis** → quantifying migration flows, protest sizes, or urban foot traffic patterns.

- **Cultural Analytics** → analyzing visual memes, fashion trends, or art/historical imagery at scale.
- **Media & Misinformation Studies** → detecting deepfakes, manipulated images, or visual framing in news/social media.
- **Socioeconomic Indicators from Satellite/Street Images** → estimating poverty, infrastructure quality, or neighborhood safety.
- **Human–Environment Interaction** → monitoring public space use, green area access, or disaster response behaviors.

Evolution of CV approaches


Early computer vision (pre-2000s)
Classical ML + features (2000s–early 2010s)
Deep learning with CNNs (2012→)
Vision Transformers (~2020)
Multimodal models (2021→)
1980–1999
1996–2010
2000–2012
2000–2012
2012–2014
2013–2020
2020–2022
2021–now

- **Early computer vision (pre-2000s)**
  - Mostly rule-based: edge detectors, filters, geometric operations.
  - Features like **SIFT** (Scale-Invariant Feature Transform) or **HOG** (Histogram of Oriented Gradients) were hand-designed by researchers.
  - These worked for specific problems (e.g., matching points between two images), but they struggled when conditions changed (lighting, pose, noise).
- **Classical ML + features (2000s → early 2010s)**
  - The pipeline was: extract features → feed them into a classifier (like SVM, Random Forest, or logistic regression).
  - This separated the "what features to look at" step from the "how to classify them" step.
  - More flexible than rule-based, but still heavily dependent on good feature design by humans.
- **Deep learning with CNNs (2012 onward)**
  - The big leap came with **AlexNet** on ImageNet in 2012.
  - CNNs learn features automatically, starting with edges, then textures, then entire objects.
  - No longer needed manual feature engineering — the network learned the hierarchy from raw pixels.
  - CNNs became the default for almost every vision task: classification, detection, segmentation.
- **Vision Transformers (ViTs, ~2020)**

    - Borrowed from NLP Transformers: instead of words, an image is split into patches.
    - Uses **self-attention** to capture relationships across the whole image, not just locally like CNNs.
    - Very powerful at scale, but less data-efficient — they need huge datasets (e.g., JFT-300M) or pretraining.
    - Opened the door to treating images more like sequences, making vision and language models more similar.

- **Multimodal models (2021 →)**

    - Next step was combining vision and language in one system (e.g., **CLIP**, **BLIP**, **Flamingo**, **Gemini**).
    - These models can align text with images, do captioning, visual question answering, and power text-to-image systems (Stable Diffusion, DALL·E).
    - They reflect a shift: not just "see" or "read," but **understand both together**.
