Big Picture:

Machine learning (ML) in Computer Vision can be **Categorized** in several ways depending on the **Task, Approach**, and **Learning Paradigm**.

Here's a breakdown to help make sense of it all:

🧠 Categorizing by Learning Paradigm (Abstract)

- **Supervised Learning**
  - Models learn from labeled datasets
  - Common for tasks like image classification and object detection
- **Unsupervised Learning**
  - Models explore patterns without labels
  - Used for clustering, dimensionality reduction (e.g., autoencoders)
- **Semi-Supervised Learning**
  - Combines a small amount of labeled data with a large pool of unlabeled data
  - Helps when labeling is expensive or slow
- **Self-Supervised Learning**
  - Learns representations from the data itself (e.g., contrastive learning)
  - Powerful for pretraining models on large unlabeled datasets
- **Reinforcement Learning**

- Agents interact with environments and learn via rewards
- Used in robotics and vision-based navigation

🛠️ Categorizing by Technique or Architecture (Comprehensive)

Taxonomy of Machine-Learning Techniques & Architectures in Computer Vision

Below is a comprehensive, two-axis categorization. First by **Architectural Family**, then by **Learning Paradigm**.

These categories often overlap in modern research:

1. **Learning Technique** >> how the model learns
2. **Model Architecture** >> structural family of the network

1. Categorizing by Learning Technique

1.1 Supervised Learning

- Learns from labeled images or videos
- Common tasks: image classification, object detection, semantic/instance segmentation, pose estimation
- Typical architectures:
  - CNN backbones (ResNet, DenseNet, EfficientNet, U-Net)
  - Vision Transformers (ViT, Swin Transformer)
  - Hybrid CNN+Transformer (CvT, ConViT)
  - Metric networks for face recognition (Siamese, triplet)

1.2 Unsupervised Learning

- Extracts structure without labels
- Tasks: clustering, density estimation, feature encoding
- Key methods:
  - Autoencoders (vanilla AE, VAE)
  - Generative models (GANs, Normalizing Flows, Autoregressive PixelCNN/RNN)
  - Deep clustering (DeepCluster, IIC)

1.3 Self-Supervised Learning

- Creates proxy tasks from raw pixels
- Paradigms:
  - Contrastive learning (SimCLR, MoCo, BYOL)
  - Masked image modeling (MAE, BEiT)
  - Pretext tasks (rotation prediction, jigsaw puzzles)
- Backbones: CNNs or Vision Transformers

1.4 Semi-Supervised & Weakly-Supervised Learning

- Mixes labeled + unlabeled (or weak labels)
- Techniques: pseudo-labeling (FixMatch), consistency regularization (UDA), multi-instance learning (CAMs for localization)
- Architectures: primarily CNNs, with emerging Transformer-based variants

1.5 Transfer Learning

- Pretrain on large corpus, fine-tune on target domain
- Common in medical, satellite imagery
- Models: pretrained CNNs (ImageNet backbones), ViTs, CLIP

1.6 Meta-Learning & Few-Shot Learning

- Learns to adapt from very few examples
- Approaches:
  - Optimization-based (MAML, Reptile)
  - Metric-based (Prototypical Networks, Matching Nets)
- Backbones: lightweight CNNs, occasionally GNNs

1.7 Zero-Shot & Open-Set Learning

- Recognize unseen classes by attribute or language embedding
- Methods: semantic prototypes, CLIP-style contrastive image-text models

1.8 Domain Adaptation & Generalization

- Align distributions across source/target
- Techniques: adversarial alignment, cycle-consistency, feature normalization

1.9 Active Learning

- Queries most informative unlabeled samples
- Strategies: uncertainty sampling, diversity sampling
- Plugs into any backbone

1.10 Reinforcement Learning in Vision

- Vision-based control, navigation, robotics
- Architectures: CNN+RNN/Transformer for temporal state encoding

1.11 Federated & Distributed Learning

- Privacy-preserving cross-device training
- Challenges: non-IID data, communication efficiency

1.12 Continual & Lifelong Learning

- Incremental updates without catastrophic forgetting
- Methods: regularization (EWC), replay buffers, dynamic architectures

2. Categorizing by Model Architecture

2.1 Convolutional Neural Networks (CNNs)

- LeNet, AlexNet, VGG → ResNet, DenseNet, EfficientNet
- Specialized: U-Net, FCN, DeepLab, PSPNet
- Uses: all supervised, many self-supervised, generative (DCGAN)

2.2 Vision Transformers

- ViT, DeiT, T2T-ViT → Swin, PVT, CoaT
- Applied in supervised, masked-model pretraining, few-shot, detection (DETR)

2.3 MLP-Based Vision Models

- MLP-Mixer, ResMLP, gMLP
- Simpler alternatives to convolutions and transformers

2.4 Recurrent & Spatiotemporal Models

- ConvLSTM, PredNet for video prediction
- 3D CNNs (C3D, I3D, SlowFast) for action recognition

2.5 Graph Neural Networks (GNNs)

- Scene graphs, relational reasoning (Graph R-CNN, GATs)

2.6 Generative Architectures

- GAN variants: DCGAN, StyleGAN, CycleGAN, BigGAN
- VAEs, β-VAE, VQ-VAE
- Normalizing flows: RealNVP, Glow
- Autoregressive: PixelRNN, PixelCNN

2.7 Hybrid & Multi-Modal Models

- CNN+Transformer (CvT, ConViT)
- Vision+Language: CLIP, ALIGN, VisualBERT, VilBERT
- GNN+CNN for relational tasks

2.8 Metric & Siamese Networks

- Siamese nets, Triplet nets for face verification, one-shot recognition

2.9 Capsule Networks

- Dynamic routing for viewpoint equivariance (CapsNet)

2.10 Spiking & Event-Based Networks

- Neuromorphic sensors, spiking neuron models for ultra-low-latency vision

2.11 Implicit & 3D Representations

- Point clouds: PointNet, PointNet++, DGCNN
- Neural fields: NeRF, DeepSDF for volumetric rendering

2.12 Efficient, Pruned, Quantized Models

- MobileNet, ShuffleNet, GhostNet for edge devices
- Lottery ticket, structured pruning, post-training quantization

2.13 Explainable & Robust Models

- Interpretability: Grad-CAM, LIME, saliency maps
- Adversarial defense: robust training, certification methods

📘 Categorizing by Emerging & Orthogonal Themes

- Efficient/Pruned/Quantized Models (MobileNet, PruneMix)
- Federated & Distributed Learning for privacy
- Explainable Vision (Grad-CAM, LIME)
- Adversarial Robustness & Security

📸 Categorizing by Vision Task

Task	Description	Example Models/Techniques
Image Classification	Assign one or more labels to an entire image	CNNs, ResNet, EfficientNet, MobileNet, Vision Transformer (ViT)
Object Detection	Identify and localize objects with bounding boxes	YOLO (v1–v8), SSD, Faster R-CNN, RetinaNet, DETR
Semantic Segmentation	Classify each pixel into a semantic category	U-Net, DeepLab, FCN, SegNet
Instance Segmentation	Segment and distinguish individual object instances	Mask R-CNN, SOLOv2, PointRend
Panoptic Segmentation	Combine semantic and instance segmentation	Panoptic FPN, UPSNet, Detectron2
Keypoint Detection	Detect anatomical/structural points (e.g., joints on people)	OpenPose, HRNet, DeepCut
Pose Estimation	Predict human or object poses via keypoints	MediaPipe, AlphaPose, PoseNet
Facial Recognition	Identify or verify individual faces	FaceNet, Dlib, ArcFace, DeepFace
Image Generation (Synthesis)	Generate realistic or stylized images from input	GANs (StyleGAN, Pix2Pix), VAEs, Diffusion Models (Stable Diffusion)
Image-to-Image Translation	Convert one image domain to another	CycleGAN, Pix2Pix, SPADE
Super Resolution	Enhance image resolution	ESRGAN, SRGAN, Real-ESRGAN
Image Denoising	Remove noise from images	DnCNN, N2N (Noise2Noise), BM3D
Image Inpainting	Fill in missing regions	Contextual Attention GAN, DeepFill
Image Captioning	Generate natural language descriptions of images	CNN+LSTM, Transformer-based (BLIP, ViLBERT)

Task	Description	Example Models/Techniques
Visual Question Answering (VQA)	Answer questions based on image content	ViLT, VisualBERT, M4C Transformer
Depth Estimation	Infer depth from single or stereo images	Monodepth2, MiDaS, DPT
3D Reconstruction	Rebuild 3D models from 2D images or video	NeRF, COLMAP, Meshroom
Optical Flow Estimation	Track motion between video frames	FlowNet, PWC-Net, RAFT
Video Classification	Assign labels to video clips	I3D, C3D, SlowFast, TimeSformer
Action Recognition	Identify human actions from video	TSN (Temporal Segment Network), SlowFast, PoseC3D
Object Tracking	Follow object location over time in video	Deep SORT, ByteTrack, Tracktor
Multimodal Learning	Combine vision with language or audio	CLIP, BLIP, Flamingo, GIT
Scene Text Recognition (OCR)	Detect and read text from images	EAST, CRAFT (detection); CRNN, Tesseract (recognition)
Image Retrieval	Find visually similar images	Deep Metric Learning, CLIP, VGG embeddings
Domain Adaptation	Apply vision models to new but similar domains	DANN, ADDA, MMD
Zero-shot Learning (ZSL)	Recognize novel classes not seen during training	CLIP, ZSL with attributes
Few-shot Learning	Learn from very few labeled examples	Matching Networks, Prototypical Networks, Meta-learning models