International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 6, Nov-Dec-2026, ISSN (Online): 2395-566X

Computer Vision: Techniques, Applications and Challenges

Rutuja Subhash Jadhav[1] Prof.Divya Rahane[2]
Department Of Artificial Intelligence And Data Science (Ai&Ds) Sahyadri Valley College Of Engineering And Technology, Rajuri, Pune

Abstract—Computer vision is the field of artificial intelligence that enables machines to extract meaningful information from images and video. Advances in deep learning, large annotated datasets, and powerful hardware have moved computer vision from hand-crafted feature engineering to end-to-end learned systems used in healthcare, transportation, manufacturing, agriculture, and security. This paper presents a structured study of computer vision. It reviews classical image-processing methods, Convolutional Neural Networks (CNNs), object detection architectures, image segmentation models, and Vision Transformers (ViTs). A general framework covering image acquisition, preprocessing, feature extraction, task-specific modelling, evaluation, and deployment is described. The paper discusses accuracy, localization quality, robustness, and efficiency as important evaluation dimensions, along with applications in medical imaging, autonomous driving, surveillance, and precision agriculture. Challenges such as limited labelled data, dataset bias, domain shift, adversarial vulnerability, privacy, and computational cost are also examined. The proposed framework is intended to support responsible development of vision systems; models should be validated on representative real-world data before deployment.

Keywords—Computer Vision, Deep Learning, Convolutional Neural Network, Object Detection, Image Segmentation, Vision Transformer, Image Classification.

I. INTRODUCTION

Computer vision aims to give machines the ability to see, interpret, and act on visual information. Applications such as medical image diagnosis, autonomous vehicles, face and document recognition, industrial inspection, crop monitoring, and video surveillance all depend on automated analysis of images and video. For many years, vision systems relied on hand-designed features such as edges, corners, and texture descriptors combined with shallow classifiers. These pipelines worked in controlled settings but struggled with variation in lighting, viewpoint, scale, and background.

The introduction of deep convolutional networks changed the field. A large CNN trained on the ImageNet dataset reduced classification error substantially and demonstrated that features could be learned directly from data [1], [8]. Deeper residual architectures, region-based and single-shot detectors, and encoder–decoder segmentation networks followed [2]–[5], and more recently Transformer-based models have been applied to images [6].

The objective of this paper is to study the main techniques of computer vision and describe a general framework for building and evaluating vision systems. Thestudy emphasizes that a high benchmark score does not automatically imply reliable real-world performance; robustness, fairness, and efficiency must be assessed for each application.

1. Objectives And Scope

The main objectives are: (1) to understand the need for computer vision in modern applications; (2) to study common vision techniques; (3) to describe a general workflow for developing vision systems; (4) to identify suitable evaluation measures for classification, detection, and segmentation; and (5) to discuss applications, limitations, and future improvements.

The scope includes still images and video, with emphasis on supervised learning tasks such as image classification, object detection, and semantic segmentation. The framework can be adapted to other tasks such as pose estimation and image captioning. Computer vision models may assist human experts or automate routine inspection, but they should not be assumed to generalize to conditions that were absent from the training data.

II. LITERATURE REVIEW

Krizhevsky et al. trained a deep convolutional network on ImageNet using GPUs, ReLU activations, and dropout, and achieved a large improvement over earlier methods in the ILSVRC competition [1]. Russakovsky et al. describe this large-scale benchmark and the progress it enabled in recognition research [8].

He et al. introduced residual learning, in which layers learn corrections to an identity mapping through skip connections. This made it practical to train networks with very many layers and improved accuracy on image recognition tasks [2].

© 2026 IJSRET

For object detection, Ren et al. proposed Faster R-CNN, which shares convolutional features between a region proposal network and a detection head [3]. Redmon et al. proposed YOLO, which treats detection as a single regression problem and runs fast enough for real-time use [4]. Ronneberger et al. presented U-Net, an encoder–decoder network with skip connections that works well for biomedical segmentation even with limited training images [5].

Dosovitskiy et al. showed that a standard Transformer applied to sequences of image patches can match or exceed CNNs on image classification when pre-trained on large datasets [6]. Lin et al. introduced the COCO dataset, which supports detection, segmentation, and captioning in everyday scenes [7]. These studies indicate that no single architecture is best for every task; selection depends on data size, latency limits, and accuracy requirements.

III. PROPOSED SYSTEM

1. **System Overview**

The proposed framework accepts images or video as input. The data is inspected, cleaned, annotated, and divided into training, validation, and test subsets. A suitable model is selected according to the task and the available computing resources. The model learns visual features from the training data and produces predictions such as class labels, bounding boxes, or pixel masks. The results are evaluated using task-specific metrics, robustness tests, and speed measurements. Only models that satisfy the application requirements are deployed.

2. **System Architecture**

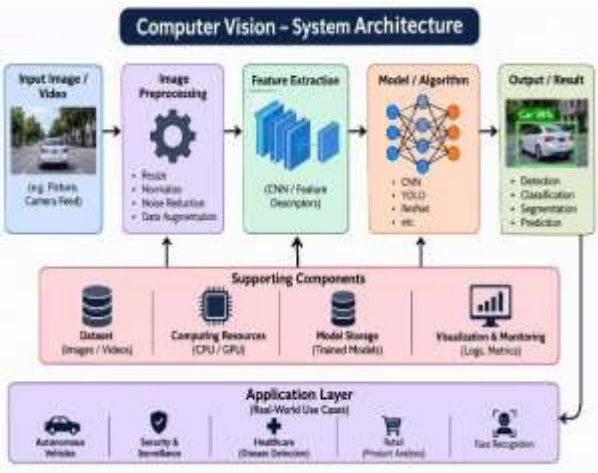

Fig. 1. General architecture of the proposed computer vision framework.

3. **Working Procedure**

First, images are collected from an authorized source such as cameras, public datasets, or institutional archives. Second, preprocessing handles corrupted files, duplicates, inconsistent labels, resizing, and normalization. Third, the dataset is split before training, and near-duplicate images are kept within the same split to prevent information leakage. Fourth, a backbone network is selected and trained, often starting from pre-trained weights. Fifth, the model is tuned using the validation set and tested on unseen images, including difficult cases. Finally, accuracy, speed, and robustness are reported before the model is deployed and monitored.

IV. COMPUTER VISION TECHNIQUES

1. Classical Image Processing

Classical methods include filtering, edge detection, thresholding, morphological operations, and feature descriptors such as SIFT and HOG. They are fast, interpretable, and need little data. Their limitation is that hand-crafted features are sensitive to changes in lighting, pose, and background, so they are now mostly used for preprocessing or on low-power devices.

2. Convolutional Neural Networks

A CNN applies learned convolution filters across an image to detect local patterns, with pooling layers reducing spatial size and deeper layers combining simple patterns into complex ones. Weight sharing makes CNNs efficient for images. Architectures such as AlexNet and ResNet have become standard backbones for many tasks [1], [2]. CNNs are strong for local texture and shape, although they may need large labelled datasets.

3.Object Detection

Object detection predicts both the class and the location of objects. Two-stage detectors such as Faster R-CNN first propose candidate regions and then classify and refine them, generally giving high accuracy [3]. One-stage detectors such as YOLO predict boxes and classes directly in a single pass, giving higher speed that suits real-time video [4]. Non-maximum suppression is commonly used to remove duplicate boxes.

4. Image Segmentation

Semantic segmentation assigns a class label to every pixel, while instance segmentation also separates individual objects of the same class. Encoder–decoder networks such as U-Net combine low-level detail from early layers with high-level context from deeper layers, which is valuable for medical scans and satellite imagery [5].

© 2026 IJSRET

5. Vision Transformers
A Vision Transformer divides an image into fixed-size patches, embeds each patch, and processes the sequence using self-attention, allowing the model to relate distant regions directly [6]. ViTs perform well when pre-trained on large datasets but may need more data and computation than CNNs when trained from scratch. Hybrid models that combine convolution and attention are also widely explored.

6. Dataset And Preprocessing
The choice of dataset depends on the intended task. Public benchmarks such as ImageNet and COCO are widely used for classification and detection research [7], [8], while domain-specific collections are needed for medical, industrial, or agricultural tasks. Medical, face, and surveillance images may contain sensitive personal information and should be used only with appropriate authorization and safeguards.

Preprocessing may include removing corrupted or duplicate files, resizing, colour normalization, noise reduction, and examining class imbalance. Data augmentation such as flipping, rotation, cropping, brightness change, and scaling increases training variety and reduces overfitting. The train–test split should be performed before augmentation so that augmented copies of the same image do not appear in both sets. For video, frames from the same sequence should stay in one split, and the test set must remain separate to provide an unbiased evaluation.

VII. PERFORMANCE EVALUATION

A vision system should be assessed from multiple perspectives. A single accuracy figure is not sufficient because a model can score well on a benchmark while failing on rare classes, new environments, or limited hardware.

1. Classification Metrics
Classification is commonly evaluated using accuracy, precision, recall, F1-score, top-k accuracy, and the confusion matrix. For imbalanced data, per-class recall and F1-score are more informative than overall accuracy.

2. Detection Metrics
Detection quality depends on Intersection over Union (IoU), the overlap between a predicted box and the ground-truth box. A prediction is counted as correct when its IoU exceeds a threshold. Average precision (AP) is computed from the precision–recall curve, and mean average precision (mAP) averages it over classes and, in COCO evaluation, over several IoU thresholds [7].

3 Segmentation Metrics
Segmentation is commonly evaluated using mean IoU (mIoU), pixel accuracy, and the Dice coefficient. Dice and IoU are preferred when the target region is small compared to the background, as in lesion segmentation.

4. Robustness And Efficiency
Robustness testing examines performance under noise, blur, lighting change, occlusion, and unseen domains. Efficiency measures include inference latency, frames per second, model size, and memory use, which determine whether a model can run on edge devices. Fairness checks across demographic groups or conditions are also important for sensitive applications.

5. Applications
In healthcare, computer vision supports analysis of X-ray, CT, MRI, and pathology images, helping clinicians detect abnormalities. In autonomous driving, vision models detect vehicles, pedestrians, lanes, and traffic signs. In manufacturing, automated inspection finds surface defects and missing components. In agriculture, drone and field images are used to monitor crop health and detect disease and weeds. It is also used in surveillance, retail analytics, document recognition, and augmented reality. In every case, the system should be validated against domain requirements and, where decisions are high-impact, reviewed by a human expert.

6. Challenges And Limitations
The quality of a vision model depends on the representativeness and accuracy of its training data. Labelling images, particularly for detection and segmentation, is costly and requires expertise. If the dataset contains bias, missing groups, or incorrect labels, the model may reproduce or amplify these problems. Domain shift, where deployment images differ from training images in camera, lighting, or location, can cause a sharp drop in performance.

Deep models can also be vulnerable to adversarial perturbations, are often difficult to interpret, and may require substantial computation for training and inference. Facial and surveillance applications raise privacy and ethical concerns. Therefore, accuracy, robustness, fairness, privacy, efficiency, and explainability should be evaluated together.

7. Future Scope
Future work may compare CNN, Transformer, and hybrid architectures on common datasets using consistent evaluation protocols. Research can focus on learning from limited labels through self-supervised, semi-supervised, and few-shot learning, as well as on synthetic data generation for rare cases. Lightweight models for edge

© 2026 IJSRET

devices, robustness to domain shift, explainable vision systems, and privacy-preserving methods such as federated learning are further directions. Multimodal models that combine images with text and standardized benchmarks can improve trust and reproducibility. Human review will remain important before vision systems are used in high-impact decisions.

XI. CONCLUSION

Computer vision is a powerful approach for extracting useful information from images and video and has transformed healthcare, transportation, industry, and agriculture. This paper described a general framework involving image collection, preprocessing, model selection, training, evaluation, and deployment. Classical methods, CNNs, object detectors, segmentation networks, and Vision Transformers provide different strengths and limitations. Reliable use requires more than a high benchmark score: accuracy, robustness, fairness, privacy, and efficiency must all be considered. Vision systems should complement carefully governed data and expert validation rather than be treated as universally accurate.

REFERENCES

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, vol. 25, 2012.
2. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
3. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.
4. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Proc. IEEE CVPR, pp. 779–788, 2016.
5. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Proc. MICCAI, LNCS vol. 9351, pp. 234–241, 2015.
6. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Int. Conf. Learning Representations (ICLR), 2021.
7. T. -Y. Lin et al., "Microsoft COCO: Common Objects in Context," Proc. ECCV, LNCS vol. 8693, pp. 740–755, 2014.
8. O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," Int. J. Computer Vision, vol. 115, no. 3, pp. 211–252, 2015. AUTHOR PROFILE

© 2026 IJSRET