International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

Computer Vision: A Review of Foundational Methods, Deep Learning Architectures, Applications and Emerging Trends

Bagal Prem Sandip[1]; Prof. S. K. Said[2]; Prof. M. C. Kolse[3]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—Computer vision enables machines to extract and interpret meaningful information from images and video. It has evolved from hand-crafted image processing and feature descriptors to deep learning systems that now perform classification, detection, segmentation and recognition at or near human level on many benchmarks. This paper reviews the field in a structured way. It begins with classical foundations including edge detection, scale-invariant keypoints and histogram-of-gradient features, then traces the deep learning revolution enabled by large labelled datasets and convolutional neural networks, covering residual networks, region-based and single-stage object detectors, fully convolutional and instance segmentation, and vision transformers. Recent foundation models that align images with language or segment arbitrary objects from prompts are also discussed. Applications in autonomous driving, medical imaging, robotics, surveillance and augmented reality are summarized. A generic five-layer computer vision pipeline from image acquisition to decision output is proposed, along with research gaps, objectives, a comparison of architectures and an analysis of challenges such as data dependence, robustness, bias, privacy and computational cost. The review concludes that computer vision is moving toward general-purpose, multimodal models that must also become efficient, robust and trustworthy.

Index Terms—Computer vision, image processing, convolutional neural networks, object detection, semantic segmentation, vision transformers, foundation models

I. INTRODUCTION

1. Motivation

Humans recognize objects, read scenes and judge distances almost effortlessly, but for computers an image is only an array of numbers. Computer vision is the discipline that turns those numbers into understanding: what objects are present, where they are, how they move and what the scene means.For decades, progress relied on carefully designed algorithms for edges, corners and textures. The release of ImageNet [1] and the success of a deep convolutional network on it [2] changed the field within a few years. Learned features replaced hand-crafted ones, and accuracy on recognition tasks improved dramatically. Szeliski [3] gives a comprehensive account of both classical and modern methods.This paper reviews the evolution of computer vision, its main deep architectures and applications, and the challenges that remain.

2. Objectives Of The Review

- **O1:** To review the evolution of computer vision from classical methods to deep learning.
- **O2:** To examine major architectures for classification, detection and segmentation.
- **O3:** To propose a generic computer vision pipeline architecture.
- **O4:** To compare representative models in accuracy, speed and data needs.
- **O5:** To identify current challenges and future trends.

II. OVERVIEW OF KEY CONCEPTS

1. Core Tasks

Fig. 1 organizes the main tasks: image classification assigns a label to the whole image; object detection localizes objects with boxes; semantic segmentation labels every pixel by class; instance segmentation separates individual objects; and other tasks include pose estimation, tracking and 3-D reconstruction.

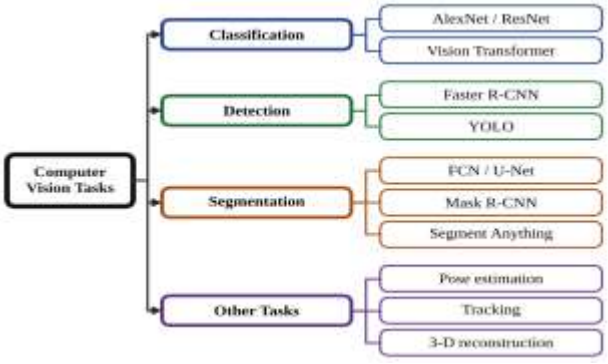


Fig. 1. Core computer vision tasks and representative methods.

© 2026 IJSRET

2. Classical Foundations

Canny's edge detector [4] defined optimal criteria for locating edges. Lowe's SIFT [5] produced keypoint descriptors invariant to scale and rotation, enabling robust matching, and Dalal and Triggs [6] showed that histograms of oriented gradients are effective features for detecting people. These ideas still underpin many lightweight systems.

III. REVIEW OF EXISTING WORK

Krizhevsky *et al.* [2] trained a deep CNN on ImageNet using GPUs, ReLU activations and dropout, cutting classification error sharply and starting the deep learning era. He *et al.* [7] introduced residual connections, which allowed networks with more than a hundred layers to train reliably.

For detection, Faster R-CNN [8] proposed regions with a learned network and classified them, achieving high accuracy, while YOLO [9] framed detection as a single regression problem and achieved real-time speed. The COCO dataset [10] became the standard benchmark for detection and segmentation in everyday scenes.

Long *et al.* [11] converted classification networks into fully convolutional networks for pixel-wise segmentation, and He *et al.* [12] extended Faster R-CNN with a mask branch for instance segmentation in Mask R-CNN.

Dosovitskiy *et al.* [13] showed that a pure transformer applied to image patches can match or exceed CNNs when pre-trained on enough data. CLIP [14] learned visual concepts from image–text pairs and enabled zero-shot recognition, and the Segment Anything model [15] segments objects from simple prompts such as points or boxes, illustrating the move toward foundation models. Table I summarizes these works.

TABLE I Summary of Literature Survey

Ref.	Author(s) / Year	Technique / Approach	Key Contribution	Limitation / Gap
[2]	Krizhevsky et al., 2012	Deep CNN (AlexNet) on ImageNet	Large drop in classification error	Needs large labelled data
[7]	He et al., 2016	Residual networks	Very deep networks train reliably	Compute-intensive
[8]	Ren et al., 2017	Faster R-CNN two-stage detector	High detection accuracy	Slower than single-stage
[9]	Redmon et al., 2016	YOLO single-stage detector	Real-time detection	Weaker on small objects
[11]	Long et al., 2015	Fully convolutional networks	End-to-end pixel labelling	Coarse boundaries
[12]	He et al., 2017	Mask R-CNN	Instance segmentation	Heavy for edge devices
[13]	Dosovitskiy et al., 2021	Vision Transformer	Transformer competitive with CNNs	Data hungry
[14]	Radford et al., 2021	CLIP image–text pre-training	Zero-shot recognition	Inherits web-data biases
[15]	Kirillov et al., 2023	Segment Anything	Promptable general segmentation	No semantic labels by itself

IV. IDENTIFIED RESEARCH GAPS

- Deep models require large labelled datasets that are costly to create for specialized domains.
- Performance can drop sharply under distribution shift, poor lighting, weather or adversarial perturbations.
- Datasets under-represent many regions and populations, causing biased results.
- Large models are expensive to train and run, limiting use on edge devices.
- Explaining why a vision model made a decision remains difficult.

1. System Architecture

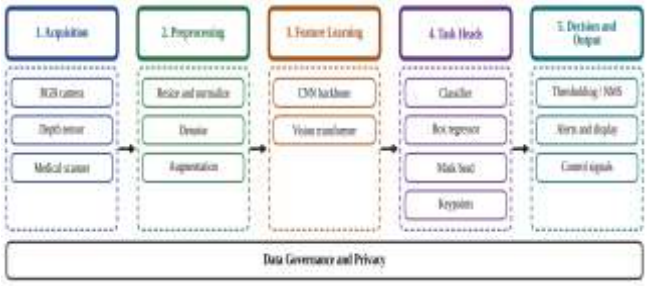


Fig. 2 shows a generic five-layer computer vision pipeline applicable to most applications.

Fig. 2. Generic five-layer computer vision pipeline.

© 2026 IJSRET

2. Acquisition Layer

Cameras, depth sensors or medical scanners capture images or video, together with metadata such as time and location.

3. Preprocessing Layer

Images are resized, normalized and denoised; augmentation creates variations during training.

C. Feature Learning Layer

A backbone network such as ResNet or a vision transformer converts the image into feature maps or tokens.

4. Task Head Layer

Task-specific heads produce class labels, bounding boxes, masks or keypoints.

5. Decision And Output Layer

Predictions are filtered with confidence thresholds and non-maximum suppression and passed to the application for display, alerts or control.

V. IMPLEMENTATION METHODOLOGY

Fig. 3 describes how a vision model is developed. A convolution layer computes each output feature as

$$y(i,j) = \Sigma_m \Sigma_n x(i+m, j+n) \cdot w(m,n) + b \qquad (1)$$

Detection quality is measured by mean average precision, which relies on Intersection over Union between predicted and true boxes:

$$IoU = |B_p \cap B_g| / |B_p \cup B_g| \quad (2)$$



Fig. 3. Development workflow for a computer vision model.

VI. PERFORMANCE COMPARISON

Table II compares representative architectures. CNNs remain efficient and strong on moderate data, transformers scale better with large data, and single-stage detectors trade a little accuracy for speed.

TABLE II Comparison of Vision Architectures

Model	Task	Speed	Data Need
ResNet-50	Classification	Fast	Medium
ViT-B/16	Classification	Medium	High
Faster R-CNN	Detection	Medium	Medium
YOLO	Detection	Very fast	Medium
Mask R-CNN	Instance segmentation	Slow	High

VII. OPEN ISSUES AND FUTURE DIRECTIONS

Table III summarizes challenges. Future trends include multimodal vision–language models, self-supervised learning that reduces labelling, efficient models for phones and embedded devices, 3-D scene understanding and stronger safeguards for privacy in surveillance applications.

TABLE III Challenges and Mitigation

Challenge	Mitigation
Labelled data cost	Self-supervised and transfer learning
Distribution shift	Augmentation; domain adaptation
Bias	Diverse datasets; subgroup evaluation
Compute cost	Pruning, quantization, efficient backbones
Privacy	On-device processing; face blurring

VIII. CONCLUSION

Computer vision has progressed from hand-crafted features to deep networks and foundation models that perform a wide range of visual tasks with high accuracy. The reviewed literature shows how convolutional networks, residual learning, modern detectors, segmentation networks and transformers each expanded what machines can see. The next stage requires models that are not only accurate but also efficient, robust, fair and respectful of privacy.

1. Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

© 2026 IJSRET

REFERENCES

1. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE CVPR, 2009, pp. 248–255.
2. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NeurIPS, 2012, pp. 1097–1105.
3. R. Szeliski, Computer Vision: Algorithms and Applications, 2nd ed. Cham, Switzerland: Springer, 2022.
4. J. Canny, “A computational approach to edge detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-8, no. 6, pp. 679–698, 1986.
5. D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.
6. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE CVPR, 2005, pp. 886–893.
7. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.
8. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017.
9. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE CVPR, 2016, pp. 779–788.
10. T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. ECCV, 2014, pp. 740–755.
11. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE CVPR, 2015, pp. 3431–3440.
12. K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE ICCV, 2017, pp. 2961–2969.
13. A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.
14. A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. ICML, 2021, pp. 8748–8763.
15. A. Kirillov et al., “Segment anything,” in Proc. IEEE/CVF ICCV, 2023, pp. 4015–4026.

© 2026 IJSRET