World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Advancements and challenges in computer vision: From pixels to perception

Abdulrahman Balogun *

Department of Computer Science, European Institute of Management and Technology, Switzerland; North Wales Management School, Wrexham University, United Kingdom.

World Journal of Advanced Research and Reviews, 2026, 30(03), 1406-1417

Publication history: Received on 10 May 2026; revised on 16 June 2026; accepted on 18 June 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.3.1714

Abstract

Computer vision has undergone remarkable transformation during the last decade, driven largely by advances in deep learning, foundation models, generative artificial intelligence, and three-dimensional vision technologies. These developments have improved performance in many areas, including healthcare, autonomous driving, robotics, manufacturing, and security. Despite these achievements, high accuracy alone does not guarantee reliable performance in real-world environments. This paper reviews the major developments and challenges in modern computer vision. The discussion focuses on deep learning architectures, transfer learning, three-dimensional vision, generative models, data quality, explainability, robustness, and ethical issues. It highlights the need for computer vision systems that are not only accurate but also transparent, reliable, and trustworthy. In addition, a conceptual framework is presented to illustrate how these challenges interact across different stages of the computer vision pipeline.

Keywords: Computer Vision; Deep Learning; Vision Transformers; Transfer Learning; Foundation Models; Explainable Artificial Intelligence; Generative Models; Adversarial Robustness; Ethical Artificial Intelligence; Three-Dimensional Vision.

1. Introduction

Computer vision is one of the most important branches of artificial intelligence. Its main objective is to enable machines to interpret and understand visual information obtained from images, videos, and sensor streams. Unlike traditional image processing, which focuses mainly on enhancing image quality, computer vision aims to extract meaningful information that can support intelligent decision-making.

Computer vision technologies are now used in many sectors. In healthcare, they assist with medical diagnosis and image analysis. In transportation, they support autonomous vehicles and traffic management systems. Other important applications include robotics, agriculture, manufacturing, surveillance, and retail.

Early computer vision systems relied heavily on handcrafted features. Techniques such as Scale Invariant Feature Transform (SIFT) (Lowe, 2004), Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005), and Speeded Up Robust Features (SURF) (Bay et al., 2008) were widely used to extract image features before classification. Although these methods achieved reasonable performance, they often struggled when images contained variations in scale, illumination, background complexity, or object orientation.

The emergence of deep learning transformed the field. Instead of relying on manually engineered features, deep neural networks learned useful representations directly from data. This shift marked an important transition from low-level image processing to high-level visual understanding. Models such as AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan & Zisserman, 2015), ResNet (He et al., 2016), EfficientNet (Tan & Le, 2019), and Vision Transformers (Dosovitskiy et al., 2021) have greatly improved performance across many visual tasks.

* Corresponding author: Abdulrahman Balogun

Despite these advances, several challenges remain. Deep learning models often require large datasets and considerable computational resources. Issues relating to explainability, robustness, bias, privacy, and ethics have also become increasingly important. These challenges demonstrate that improving benchmark accuracy alone is insufficient for building dependable real-world systems.

This review examines eight major areas that continue to shape the development of computer vision. These include deep learning, transfer learning, three-dimensional vision, generative models, data quality, explainability, robustness, and ethical considerations. Understanding these challenges is essential for developing computer vision systems that are accurate, reliable, and trustworthy.

2. The Deep Learning Revolution in Computer Vision

Deep learning has changed computer vision more than any other technological development in the last decade. The introduction of convolutional neural networks (CNNs) led to remarkable improvements in image classification and object recognition. A major breakthrough occurred in 2012 when AlexNet (Krizhevsky et al., 2012) achieved outstanding performance in the ImageNet challenge, demonstrating the effectiveness of deep learning for visual tasks.

Following the success of AlexNet, several architectures were developed to improve performance and efficiency. VGGNet (Simonyan & Zisserman, 2015) showed that deeper networks with smaller filters could achieve better results. Inception networks (Szegedy et al., 2015) introduced multi-scale processing to improve computational efficiency, while ResNet (He et al., 2016) addressed the vanishing gradient problem and enabled the training of very deep networks. EfficientNet (Tan & Le, 2019) later demonstrated that balancing depth, width, and image resolution could achieve high accuracy with fewer parameters.

More recently, Vision Transformers (ViTs) (Dosovitskiy et al., 2021) have introduced a new approach to image understanding. Unlike convolutional networks, Vision Transformers use attention mechanisms originally developed for natural language processing (Vaswani et al., 2017). These models can capture long-range relationships within images and have demonstrated excellent performance in large-scale applications. Models such as CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) have further expanded the capabilities of computer vision by combining visual and textual information.

Although these developments have improved performance significantly, they have also introduced new challenges.

First, modern models require large amounts of data and powerful computing resources. Training large networks can be expensive and difficult for smaller organisations with limited infrastructure.

Second, many deep learning models operate as black boxes. While they achieve impressive accuracy, understanding how predictions are generated remains difficult. This lack of transparency raises concerns in sensitive applications such as healthcare and autonomous driving.

Third, increasing model complexity has raised concerns regarding energy consumption and environmental sustainability. Training very large models requires substantial computational resources, which increases both financial costs and carbon emissions.

Finally, deep learning models often perform poorly when exposed to conditions that differ from those encountered during training. Variations in weather, image quality, lighting conditions, and unexpected objects can significantly reduce performance. As a result, excellent benchmark results do not always translate into reliable real-world deployment.

Despite these challenges, deep learning remains the foundation of modern computer vision. Future research must focus not only on improving accuracy but also on enhancing efficiency, explainability, robustness, and trustworthiness.

3. Transfer Learning and Foundation Models

Transfer learning has become one of the most important techniques in modern computer vision. Training deep neural networks from scratch often requires large datasets, expensive hardware, and considerable time. In many practical applications, these resources are not readily available. Transfer learning addresses this problem by adapting pre-trained models to new tasks.

Most transfer learning approaches make use of models trained on large datasets such as ImageNet. Instead of building a model from the beginning, researchers fine-tune existing networks using smaller domain-specific datasets. This approach reduces training time and computational cost while improving performance.

Transfer learning has found applications in many areas. In healthcare, pre-trained models have been successfully adapted for disease diagnosis and medical image analysis. Similar approaches are used in agriculture, manufacturing, remote sensing, and environmental monitoring.

Despite these advantages, transfer learning faces several challenges. One major issue is domain shift. Features learned from natural images may not perform well when applied to specialised data such as medical images or thermal imagery. Differences between the source and target domains can lead to reduced accuracy and poor generalisation.

Another concern is bias transfer. If the original training data contain imbalances or annotation errors, these problems may be inherited by the new model. Consequently, predictions may become unreliable or unfair.

Recent advances have introduced foundation models such as CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2023), Segment Anything Model (SAM) (Kirillov et al., 2023), and GPT-4V (OpenAI, 2023). More recent iterations, such as SAM 2 (Ravi et al., 2024), have extended these capabilities from static images to video segmentation. These models are trained on massive datasets and can perform a variety of tasks with minimal additional training. Their flexibility has attracted significant attention from both researchers and industry.

However, foundation models introduce their own challenges. Their training processes are computationally intensive, they are often difficult to interpret, and they may inherit biases present in internet-scale datasets. Furthermore, updating these models requires substantial resources and expertise.

Transfer learning and foundation models have made advanced computer vision techniques more accessible. Nevertheless, careful evaluation and responsible deployment are necessary to ensure reliable performance across different domains.

4. Three-Dimensional Vision and Sensor Fusion

Three-dimensional vision has become increasingly important in applications such as autonomous vehicles, robotics, drones, and industrial automation. Unlike traditional two-dimensional images, three-dimensional vision provides information about depth and spatial structure, allowing machines to understand their environments more effectively.

Three-dimensional information is commonly obtained using LiDAR sensors, stereo cameras, and RGB-D cameras. These devices generate point clouds that describe objects and surfaces in three-dimensional space.

Processing point clouds presents several challenges. Unlike images, point clouds are irregular and sparse, making them difficult to analyse using conventional convolutional neural networks. Architectures such as PointNet and PointNet++ (Qi et al., 2017a; Qi et al., 2017b) have been developed to address these difficulties and have improved the understanding of three-dimensional structures.

Sensor fusion represents another important area of research. Combining information from multiple sensors allows systems to exploit the strengths of different sensing technologies. For example, cameras provide rich visual information, while LiDAR sensors offer accurate depth measurements. Combining these data sources improves reliability and accuracy.

Sensor fusion plays a crucial role in autonomous driving systems, where different sensors complement one another. However, integrating multiple sensors introduces additional complexity. Calibration errors, synchronization issues, and environmental conditions such as rain, fog, and bright sunlight can affect system performance.

Occlusion is another important challenge. Objects may be partially hidden, making detection and recognition more difficult. Real-time processing requirements further increase the complexity because systems must make accurate decisions within very short time intervals.

Recent developments in Neural Radiance Fields (NeRFs) and implicit neural representations have created new opportunities for three-dimensional scene reconstruction. These approaches can produce highly detailed models of complex environments. However, their high computational requirements currently limit their application in real-time systems.

Three-dimensional vision remains an active area of research, and improving efficiency, robustness, and scalability continues to be a major challenge.

5. Generative Models: From GANs to Diffusion Models

Generative models have emerged as one of the most exciting areas in computer vision. These models can create realistic images, improve image quality, and generate synthetic data for machine learning applications.

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), consist of two competing networks. The generator produces synthetic images, while the discriminator attempts to distinguish between real and generated images. Through this competition, both networks gradually improve their performance.

GANs have been applied to image restoration, super-resolution, style transfer, and synthetic data generation. Models such as StyleGAN have demonstrated the ability to generate highly realistic images that are often difficult to distinguish from real photographs.

More recently, diffusion models (Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022) have emerged as powerful alternatives to GANs. Models such as DALL-E, Stable Diffusion, and Imagen are capable of producing high-quality images from textual descriptions. Compared with GANs, diffusion models generally provide more stable training and better diversity in generated outputs.

Despite their impressive capabilities, generative models present several technical challenges. GANs are often difficult to train and may suffer from mode collapse, where the generator produces limited variations of images. Hyperparameter tuning can also be complex and computationally expensive.

Generative models have also raised important ethical concerns. Deepfake technology has enabled the creation of highly realistic synthetic images and videos that can be used for misinformation, fraud, and identity theft. As these systems become more sophisticated, distinguishing between authentic and manipulated content becomes increasingly difficult.

Another challenge concerns the evaluation of generated images. Existing metrics, such as the Fréchet Inception Distance (FID) (Heusel et al., 2017), provide useful measures of image quality but do not fully capture realism or diversity.

Researchers are currently exploring techniques for watermarking synthetic content, detecting deepfakes, and improving transparency. Regulatory frameworks and ethical guidelines are also becoming increasingly important as generative artificial intelligence continues to evolve.

Generative models offer enormous potential for future applications. However, their responsible development and deployment will be essential to ensure that their benefits outweigh their associated risks.

6. Data Quality, Curation and Governance

Data are at the centre of every computer vision system. Regardless of how sophisticated a model may be, its performance largely depends on the quality, diversity, and reliability of the training data. Poor-quality datasets often lead to inaccurate predictions and poor generalisation.

One of the biggest challenges in computer vision is data annotation. Creating labelled datasets requires considerable time, effort, and financial resources. Human errors during annotation may introduce incorrect labels, which can negatively affect model performance. In some cases, different annotators may interpret the same image differently, leading to inconsistencies within the dataset.

Class imbalance is another common problem. Some classes may contain thousands of examples, whereas others are represented by only a small number of samples. As a result, models often become biased towards majority classes and perform poorly on minority categories.

Bias within datasets has attracted increasing attention in recent years. Studies have shown that facial recognition systems may exhibit different levels of accuracy across demographic groups when training data fail to represent the population adequately. Such biases may lead to unfair decisions and undermine public trust in artificial intelligence systems.

Another challenge is dataset ageing. Real-world conditions constantly evolve, and datasets collected several years ago may no longer reflect present environments. For example, images captured under controlled laboratory conditions may differ significantly from those encountered in real-world applications.

To address these issues, researchers have increasingly adopted a data-centric approach to artificial intelligence. Rather than focusing exclusively on model architecture, greater emphasis is now placed on improving data quality,

documentation, and governance. Frameworks such as Datasheets for Datasets and Model Cards provide detailed information about how datasets are collected, their intended use, and their limitations.

Privacy has also become an important aspect of data governance. Many applications involve sensitive information, particularly in healthcare and surveillance. Techniques such as federated learning and differential privacy aim to protect personal information while maintaining model performance.

Ultimately, improving data quality is just as important as developing more advanced algorithms. Reliable and representative datasets are essential for building trustworthy computer vision systems.

7. Explainable Artificial Intelligence and Interpretability

As computer vision systems are increasingly used in critical applications, understanding how these systems arrive at their decisions has become a major research challenge. Explainability is particularly important in domains such as healthcare, autonomous driving, finance, and law enforcement, where incorrect predictions may have serious consequences.

Modern deep learning models often achieve excellent performance but operate as black boxes. Although they can provide highly accurate predictions, understanding the reasoning behind these predictions is often difficult. This lack of transparency can reduce trust and limit the adoption of artificial intelligence systems.

Several techniques have been developed to improve model interpretability. Saliency maps highlight image regions that contribute to predictions. Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) provides visual explanations by identifying important regions within an image. Other approaches, such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017), estimate the contribution of individual features to model outputs. Samek et al. (2021) provide a comprehensive overview of these and related explainability techniques.

Although these methods provide useful insights, they also have limitations. Explanations may vary between different runs and may not always reflect the true internal reasoning of the model. In some situations, explanations that appear convincing to humans may fail to represent how predictions were actually generated.

Another challenge is the lack of standard evaluation metrics for explainability. Different techniques may produce different explanations for the same prediction, making it difficult to determine which explanation is most reliable.

Researchers have therefore begun to explore inherently interpretable models. Rudin (2019) argues that for high-stakes decisions, transparent models should be preferred over post-hoc explanations. Although such models may sacrifice some degree of accuracy, they can provide greater confidence in safety-critical applications.

Recent studies have also highlighted the importance of causal reasoning. Traditional explainability methods often describe correlations rather than causal relationships. Future research is expected to combine explainable artificial intelligence with causal inference to provide more meaningful and reliable explanations.

Explainability remains one of the most active areas of research in artificial intelligence. As computer vision systems continue to influence important decisions, transparency and accountability will become increasingly essential.

8. Robustness and Adversarial Attacks

Although computer vision systems have achieved impressive performance under controlled conditions, they often struggle when exposed to unexpected situations. Ensuring that models remain reliable under diverse circumstances is known as robustness.

One of the most important discoveries in deep learning research is the existence of adversarial attacks. Goodfellow et al. (2015) demonstrated that very small, nearly imperceptible changes to images can cause models to produce completely incorrect predictions. Such vulnerabilities raise serious concerns regarding the security and reliability of artificial intelligence systems.

Adversarial attacks can be divided into different categories. Some attacks require knowledge of the internal structure of the model, while others can be performed without access to the target system. Researchers have also demonstrated physical attacks, where specially designed stickers or patterns can deceive computer vision systems operating in real-world environments.

Apart from malicious attacks, natural environmental changes can also affect model performance. Hendrycks & Dietterich (2019) showed that variations in lighting, rain, fog, blur, compression artefacts, and unfamiliar objects may significantly reduce accuracy. Models that perform exceptionally well on benchmark datasets often struggle when deployed in practical settings.

Distribution shift presents another important challenge. Data encountered during deployment frequently differ from those used during training. Changes in camera settings, weather conditions, or population characteristics can result in substantial performance degradation.

To improve robustness, researchers have proposed several defence mechanisms. Adversarial training incorporates manipulated examples into the training process to increase resistance to attacks. Ensemble methods combine multiple models to improve stability and reliability. Other approaches rely on formal verification techniques and certified defences to provide mathematical guarantees regarding model behaviour.

Another growing area of interest is uncertainty estimation. Ideally, models should recognise when they are uncertain instead of making highly confident but incorrect predictions. Bayesian methods and Monte Carlo techniques have been developed to provide more reliable confidence estimates.

Despite considerable progress, achieving truly robust computer vision systems remains difficult. Defences that are effective against one type of attack may fail against others, and many defence mechanisms introduce additional computational costs.

Robustness therefore remains one of the most important challenges in modern computer vision research. Future developments will require a balance between accuracy, security, efficiency, and reliability.

Table 1 Comparison of Major Computer Vision Architectures

Architecture	**Strengths**	**Limitations**	**Typical Applications**
Convolutional Neural Networks (CNNs)	Efficient feature extraction and strong performance	Limited ability to capture long-range relationships	Image classification, object detection
ResNet	Enables very deep networks and improves training stability	High computational complexity	Medical imaging, image recognition
EfficientNet	High accuracy with fewer parameters	Requires careful scaling	Mobile vision applications
Vision Transformers (ViTs)	Excellent global feature representation	Requires large datasets and high computational power	Large-scale image analysis
Foundation Models (CLIP, DINOv2, SAM)	Strong transferability and generalisation	Expensive training and limited interpretability	Multimodal learning, zero-shot classification

Table 2 Comparison of Explainability Techniques

Method	**Main Idea**	**Advantages**	**Limitations**
Saliency Maps	Highlight important image regions	Simple and intuitive	Sensitive to noise
Grad-CAM	Visual explanation based on gradients	Easy to interpret	Limited spatial precision
LIME	Local model approximation	Model-independent	Computationally expensive
SHAP	Feature contribution analysis	Strong theoretical basis	Slow for large models
Inherently Interpretable Models	Transparent model structures	Improved trust and accountability	May sacrifice accuracy

Table 3 Major Adversarial Attacks and Their Effects

Attack Type	Description	Impact
FGSM (Fast Gradient Sign Method)	Small perturbations added to images	Misclassification
PGD (Projected Gradient Descent)	Iterative attack technique	Reduced robustness
CW Attack (Carlini-Wagner)	Optimisation-based attack	High attack success rate
Physical Attacks	Real-world modifications such as stickers	Failure of object recognition systems
Distribution Shift	Environmental changes and unseen conditions	Performance degradation

9. Ethical Considerations and Regulatory Frameworks

Ethical considerations have become increasingly important as computer vision technologies continue to advance. While these systems offer significant benefits, they also raise concerns regarding privacy, fairness, accountability, and misuse. As a result, ethical issues are now regarded as an essential component of responsible artificial intelligence.

One of the major concerns relates to privacy. Modern computer vision systems are capable of analysing faces, tracking individuals, and monitoring activities in real time. Although these capabilities have many practical applications, they may also threaten personal privacy if appropriate safeguards are not implemented. The widespread use of surveillance systems has therefore generated concerns regarding civil liberties and the protection of personal information.

Bias and fairness represent another important challenge. If training datasets fail to represent different populations adequately, computer vision models may produce unequal outcomes across demographic groups. Such biases can have serious consequences when artificial intelligence systems are applied in healthcare, law enforcement, recruitment, and border security. Ensuring fairness requires careful dataset design, continuous monitoring, and regular auditing.

The rapid development of generative artificial intelligence has introduced additional ethical concerns. Deepfake technologies can create highly realistic synthetic images and videos that may be used for misinformation, fraud, and identity theft. As these technologies become more sophisticated, distinguishing between authentic and manipulated content becomes increasingly difficult.

Accountability also remains a major issue. When artificial intelligence systems make incorrect decisions, questions arise regarding who should be held responsible. Developers, organisations, regulators, and users all share responsibility for ensuring that these systems operate safely and ethically. Effective governance frameworks are therefore necessary to promote transparency and accountability.

Several regulatory initiatives have emerged to address these concerns. The European Commission (2024) introduced the Artificial Intelligence Act, which classifies AI systems according to their level of risk and establishes requirements for high-risk applications. The National Institute of Standards and Technology (2023) has published an AI Risk Management Framework to guide responsible AI development. Similarly, the General Data Protection Regulation (GDPR) establishes rules regarding privacy, data protection, and automated decision-making. UNESCO (2021) has also issued a Recommendation on the Ethics of Artificial Intelligence providing global guidance for responsible development.

Responsible deployment of computer vision systems requires more than technological innovation. It also demands transparency, fairness, human oversight, and compliance with legal and ethical standards. Future progress in computer vision will depend not only on improving technical performance but also on maintaining public trust.

10. Conceptual Framework: From Pixels to Trustworthy Perception

The challenges discussed throughout this paper are closely interconnected. Improvements in one area often influence several others. To illustrate these relationships, this study proposes a conceptual framework that organises the computer vision pipeline into five interconnected layers.

The first layer is the Input Layer, which consists of images, videos, point clouds, and sensor streams. The quality and diversity of these inputs directly affect the overall performance of the system.

The second layer is the Representation Layer. At this stage, deep learning architectures such as Convolutional Neural Networks, Vision Transformers, and foundation models learn meaningful features from visual data. Transfer learning and self-supervised learning techniques also contribute to this layer.

The third layer is the Reasoning Layer, where models perform tasks such as image classification, object detection, segmentation, depth estimation, and scene understanding. Generative models and three-dimensional vision systems also contribute to this stage.

The fourth layer is the Trustworthiness Layer. This layer focuses on explainability, robustness, fairness, privacy, and uncertainty estimation. These properties determine whether model predictions can be trusted in practical applications.

The final layer is the Governance Layer, which includes ethical principles, regulatory frameworks, audit mechanisms, and human oversight. This layer ensures that computer vision systems are developed and deployed responsibly.

Figure 1 Conceptual Framework – From Pixels to Trustworthy Perception

The proposed framework highlights an important observation. Weaknesses in data quality may affect model performance, explainability, robustness, and ultimately public trust. Similarly, ethical and regulatory requirements influence the design and deployment of computer vision systems.

Consequently, the development of trustworthy computer vision systems requires a balanced approach that considers technical performance together with transparency, fairness, security, and accountability.

10.1. Discussion of the Framework

The proposed framework demonstrates that trustworthy computer vision systems depend on several interconnected layers. High-quality input data support effective feature representation, which in turn enables accurate reasoning and

decision-making. However, technical performance alone is insufficient. Explainability, robustness, fairness, and privacy are necessary to establish trust. These characteristics are further supported by governance mechanisms that promote accountability and ethical compliance.

This layered view emphasises that weaknesses in one component can affect the entire system. Consequently, developing reliable computer vision systems requires a holistic approach that balances accuracy with transparency, security, and responsible governance.

10.2. Future Research Directions

Computer vision continues to evolve rapidly, and several emerging areas are expected to shape future developments. One important direction is multimodal artificial intelligence, where visual information is combined with text, audio, and other forms of data to enable more comprehensive understanding. Models such as vision-language foundation models and other multimodal systems (Wang et al., 2024) demonstrate the potential of integrating multiple modalities for more intelligent decision-making.

Continual learning represents another promising area. Traditional deep learning systems often require retraining whenever new data become available. Future computer vision systems are expected to learn continuously and adapt to changing environments without forgetting previously acquired knowledge.

Foundation models are also likely to play an increasingly important role. Models trained on massive datasets can perform multiple tasks with minimal additional training. However, improving their efficiency, explainability, and fairness remains a major challenge.

Edge artificial intelligence is expected to expand significantly in applications such as autonomous vehicles, robotics, wearable devices, and smart cities. Developing lightweight models that can operate under limited computational resources will therefore become increasingly important.

Finally, future research must focus on trustworthy artificial intelligence. Explainability, robustness, privacy, fairness, and ethical governance will remain central requirements for ensuring that computer vision systems are accepted and deployed responsibly.

11. Conclusion

Computer vision has experienced remarkable progress over the past decade. Advances in deep learning, transfer learning, foundation models, generative models, and three-dimensional vision have significantly improved the ability of machines to understand visual information. These developments have enabled numerous applications in healthcare, autonomous driving, robotics, agriculture, manufacturing, and many other fields.

Despite these achievements, important challenges remain. High computational requirements, limited explainability, vulnerability to adversarial attacks, data quality issues, and ethical concerns continue to limit the reliability of computer vision systems in real-world environments. Strong performance on benchmark datasets alone does not guarantee dependable operation in practical applications.

This review has examined eight major challenge areas and explored how they influence one another. The proposed conceptual framework demonstrates that data quality, model architecture, robustness, explainability, and governance should not be viewed as independent problems but as interconnected components of trustworthy artificial intelligence.

Future research should focus on developing systems that are not only accurate but also efficient, interpretable, secure, and fair. Emerging directions such as multimodal learning, continual learning, privacy-preserving artificial intelligence, and transparent foundation models are expected to play important roles in the next generation of computer vision systems.

Ultimately, the future of computer vision extends beyond achieving higher accuracy. The next generation of intelligent systems must also be transparent, robust, fair, and trustworthy. Progress in this field will require collaboration among researchers, industry practitioners, policymakers, and society to ensure that technological advances deliver benefits that are both responsible and sustainable.

Compliance with ethical standards

Disclosure of conflict of interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Funding

This research received no external funding. The study was conducted through self-funding by the author.

Data Availability Statement

No new datasets were created or analysed during this study. All information used in this review was obtained from publicly available sources cited in the reference list.

References

[1] Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11), 2189–2202.

[2] Bay, H., Tuytelaars, T., & Van Gool, L. (2008). SURF: Speeded up robust features. Computer Vision and Image Understanding, 110(3), 346–359.

[3] Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems.

[4] Carion, N., Massa, F., Synnaeve, G., et al. (2020). End-to-end object detection with transformers. European Conference on Computer Vision.

[5] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1251–1258.

[6] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[7] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

[9] European Union. (2024). Artificial Intelligence Act. European Commission.

[10] Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI benchmark suite. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[11] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672–2680.

[13] Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations.

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

[15] Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. International Conference on Learning Representations.

[16] Heusel, M., Ramsauer, H., Unterthiner, T., et al. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems.

[17] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems.

[18] Howard, A., Sandler, M., Chu, G., et al. (2019). Searching for MobileNetV3. Proceedings of the IEEE International Conference on Computer Vision, 1314–1324.

[19] Kingma, D., & Welling, M. (2014). Auto-encoding variational Bayes. International Conference on Learning Representations.

[20] Kirillov, A., Mintun, E., Ravi, N., et al. (2023). Segment Anything. Proceedings of the IEEE/CVF International Conference on Computer Vision.

[21] Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

[22] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

[23] Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE International Conference on Computer Vision, 10012–10022.

[24] Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110.

[25] Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[26] National Institute of Standards and Technology. (2023). AI Risk Management Framework 1.0. NIST.

[27] OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

[28] Oquab, M., Darcet, T., Moutakanni, T., et al. (2023). DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

[29] Qi, C., Su, H., Mo, K., & Guibas, L. (2017a). PointNet: Deep learning on point sets for 3D classification and segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 652–660.

[30] Qi, C., Yi, L., Su, H., & Guibas, L. (2017b). PointNet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems.

[31] Radford, A., Kim, J., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning.

[32] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[33] Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD Conference, 1135–1144.

[34] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 10684–10695.

[35] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention, 234–241.

[36] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

[37] Saharia, C., Chan, W., Saxena, S., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems.

[38] Samek, W., Montavon, G., Vedaldi, A., Hansen, L., & Müller, K. (2021). Explainable AI: Interpreting, explaining and visualizing deep learning. Springer.

[39] Selvaraju, R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision.

[40] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.

[41] Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–9.

[42] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, 6105–6114.

[43] UNESCO. (2021). Recommendation on the ethics of artificial intelligence. UNESCO.

[44] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998–6008.

[45] Zhou, B., Khosla, A., Lapedriza, A., et al. (2016). Learning deep features for discriminative localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[46] Ravi, N., Gabeur, V., Hu, Y., et al. (2024). *SAM 2: Segment Anything in Images and Videos*. arXiv preprint arXiv:2408.00714.

[47] Wang, W., Chen, C., Ding, M., et al. (2024). *Vision-language foundation models: A survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence.