



Article

Intelligent visual perception beyond classification: a survey of deep learning and foundation models

Sunita Dixit*

Department of Computer Science & Engineering, St. Andrews Institute of Technology & Management, Gurgaon, India

ARTICLE INFO	ABSTRACT
<i>Article history:</i> Received 10 August 2026 Received in revised form 14 September 2026 Accepted 01 October 2026 Keywords: Intelligent visual perception, Deep learning, Vision transformers, Multimodal learning, Foundation models *Corresponding author Email address: bhardwajsunita23@gmail.com DOI: 10.55670/fpjl.fdtai.2.2.3	Intelligent visual perception has advanced from traditional image classification to thorough comprehension of objects, scenes, relations, spatial context, and semantic information. This paper surveys the development of intelligent visual perception, particularly focusing on DL and foundation models. This section provides a comprehensive overview of visual recognition, detection, segmentation, and scene understanding, including key methods such as convolutional neural networks, recurrent neural networks, attention mechanisms, encoder-decoder architectures, generative models, and Vision Transformers. The study also explores feature attribution, visual attention, visual reasoning, and multimodal vision-language models, which combine visual and text information for tasks such as visual question answering, captioning, and visual generation. This study examines recent foundation models incorporating vision-language architectures for their transferability, contextual understanding, and generalized perception capabilities. The study also discusses challenges related to computational efficiency, environmental variability, data requirements, robustness, explainability, and real-time deployment. Finally, it contrasts current developments to identify new approaches to scalable, interpretable, multimodal, and human-centric visual intelligence systems.

1. Introduction

Manufacturing Healthcare increasingly relies on intelligent technology to enable accurate diagnosis, clinical decision-making, patient monitoring, and medical analysis, creating a growing need for dependable systems that can efficiently interpret complex visual information. Humans perceive and interact with the world mostly through visual information, which accounts for more than 80% of all external information humans gain [1]. Visual perception is a key element of intelligent systems, allowing robots to see and interpret their surroundings, recognize things, and make educated decisions [2]. Traditional visual perception systems rely on passive processing of visual input and have been widely used across fields, focusing on object identification, segmentation, and recognition. Such systems often rely on predefined algorithms to extract information from fixed perspectives or sensors and perform well in controlled, predictable contexts. But in the actual world, environments are typically complex, dynamic, and congested, making it difficult for passive systems to collect the entire breadth of essential information [3].

Deep Learning (DL) is a subfield of AI that has revolutionized image processing [4], emulating how the human brain processes and analyses information. Historically, classical ML techniques and manual feature extraction were used for image processing [5]. These approaches required substantial domain expertise and often struggled with the complexity and unpredictability of visual input. These systems performed well for tightly stated aims, but they couldn't handle the diversity and high dimensionality of real-world photographs because they weren't scalable or flexible enough [6]. DL's ability to learn hierarchical representations directly from raw data has helped mitigate many limitations of previous techniques [7]. With the development of multi-layered neural networks (NNs) it became possible for models to learn complicated patterns and features automatically that were not possible with human procedures [8]. Automated feature learning is a milestone in the history of image processing, replacing manually created feature engineering. This change enabled enormous improvements in accuracy and generalizability across many applications [9].

Foundation model building blocks like deep neural networks and self-supervised learning have been present for a while, but the current boom, especially in LLMs, is mostly due to data and model size scaling up. Significant research on large foundation models for various

sensory problems has coincided with LLMs in natural language processing. For example, CLIP and other pre-trained vision-language (VL) models have shown strong zero-shot performance on several downstream vision tasks, such as object identification and image categorization. The representations provided by these VL foundation models, usually trained on millions of image-text pairings retrieved from the web, transfer well and generalize strongly. By providing these VL foundation models with a natural language task description and instructions, they can be trained to perform a downstream task. The groundbreaking CLIP paradigm, for example, uses thoughtfully crafted prompts to accomplish various downstream tasks, such as zero-shot classification- a process in which the text encoder builds classifiers on the fly using class names or other free-form text. Customized prompt templates help here; for instance, "A photo of a {label}" can help ensure the language matches the image's visual content [10]. A lot of recent work has also looked at ways to fine-tune VL models on a given set of instructions such that they can conduct conversations [11]. This paper structure is as follows: Section II introduces Intelligent Visual Perception. Section III explores Deep Learning for Intelligent Visual Perception. Section IV reviews recent literature, and Section V concludes the paper while outlining potential future research directions.

2. Intelligent visual perception overview

Intelligent visual perception refers to artificial intelligence systems' ability to interpret and understand visual information beyond simple image classification. It integrates object detection, segmentation, attention mechanisms, feature attribution, visual reasoning, and multimodal learning to capture objects, relationships, spatial context, and semantic information. Modern vision transformers, vision-language models, and foundation models further enhance these capabilities by enabling contextual understanding and reasoning across visual and textual modalities. Intelligent visual perception lays the groundwork for future visual intelligence systems that are more adaptable, contextually aware, interpretable, and human-focused.

2.1 Transformers for intelligent visual perception

Transformers are primarily used for text encoding, with subsequent applications including merging visual and linguistic elements [12]. Rather than significantly altering the central focus block shown in Figure 1, these models often use the vanilla multi-layer Transformer with multimodal inputs.

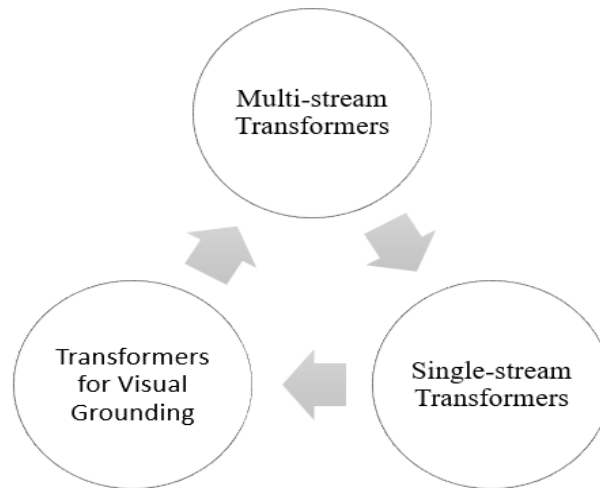


Figure 1. Transformers for intelligent visual perception

Multi-stream transformers: The initial attempt to adapt the BERT concept for use in a multi-modal setting was ViLBERT [13]. The objective was to learn representations that model both visual data and natural language. To achieve this goal, ViLBERT designed a two-stream architecture, with one stream designed to mimic visual inputs and the other to mimic language inputs.

Single-stream transformers: VisualBERT jointly represents the visual and textual worlds with a single stack of Transformers. The Transformer is trained to automatically learn relations between the two domains from an input sequence of text (e.g., a caption) and the visual attributes that match the object suggestions.

Transformers for visual grounding: The CNN and BERT backbones of a modulated DETR each extract features from their respective input types: images and text, respectively [14]. Afterward, the visual and textual data are linearly projected into a common space, combined, and input into a transformer model. This model then predicts the object bounding boxes that match the queries in the grounded text.

2.2 Feature attribution for intelligent visual perception

An attribution approach provides relevance ratings to input variables based on their contribution to a model's output. Higher ratings indicate more relevance, whereas lower scores suggest less relevance. In vision models, these scores are visualized as saliency maps, where red denotes important parts, while colors fade to yellow and green for less important regions.

Attention: Rollout and Attention Flow are early approaches for measuring attention flow across many layers and heads in transformer systems. Rollout computes attention weights by multiplying the attention matrices of each layer sequentially, resulting in a cumulative assessment of how input tokens affect the final output.

Gradient: Gradients of the output with respect to input features help identify important areas that contribute to the model's predictions [15]. Foundational techniques such as saliency maps, Layer-wise Relevance Propagation (LRP), DeepTaylor Decomposition (DTD), and Integrated Gradients (IG) form the basis for specialized approaches for ViTs.

Perturbation: Evaluate feature relevance by varying input characteristics and monitoring corresponding changes in the model's output. These strategies are based on theoretical approaches like LIME, which approximates the model locally via an interpretable surrogate model, or SHAP, which computes Shapley values from game theory for fair feature attribution. Recent techniques have generalized these frameworks to ViTs.

2.3 Attention mechanisms for visual intelligence

Attention mechanisms allow visual models to focus on informative areas, features, or tokens in an image. They enhance contextual awareness and feature representation by selectively emphasizing pertinent visual information while suppressing less relevant areas.

Channel attention: Channel Attention (CA) models inter-channel dependencies and recalibrates feature responses by adaptively weighting features from distinct channels.

Spatial attention: One approach to train a model to differentiate between similar images is spatial attention (SA), which aims to describe feature saliency fluctuations across the spatial dimension [16]. Spatial attention focuses on “where to focus” by weighting spatial locations, unlike channel attention.

2.4 Multimodal visual understanding and generation

Combine visual and verbal information so models can understand images, answer queries, reason about scenarios, and provide meaningful descriptions or visual content. These skills provide enhanced vision–language interaction, contextual reasoning, and cross-modal content production beyond standard visual categorization, as seen in Figure 2.

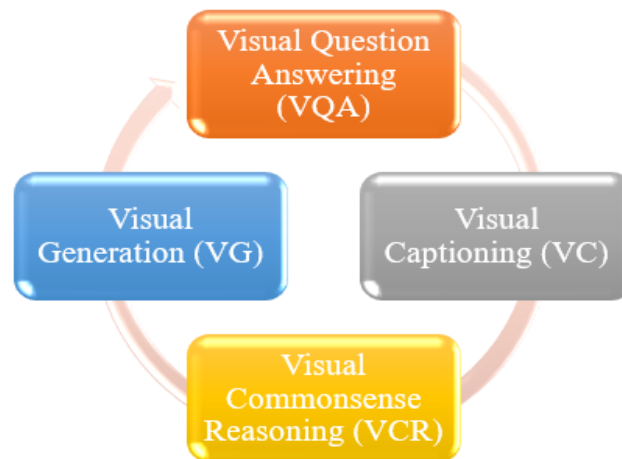


Figure 2. Multimodal visual understanding and generation

Visual Question Answering (VQA): This field focuses on enabling machines to understand visual material and respond appropriately to queries about the provided image. The model's accuracy depends on its ability to understand visual data and extract the logical connections relevant to the question.

Visual Captioning (VC): This method aims to automatically generate syntactically and semantically correct descriptions of visual material. To produce effective captions, substantial language information must be combined with knowledge of the entities, scenes, and interactions in the visual input.

Visual Commonsense Reasoning (VCR): This area focuses on robots' ability to infer cognitive and commonsense information from visual material [17]. This kind of reasoning includes answering questions about an image while providing proper arguments drawn from the visual context.

Visual Generation (VG): This approach includes creating visual material from written descriptions. To produce outputs that are visually relevant, cohesive, and context-rich, one must understand the semantic information in the input text.

2.5 Benefits of intelligent visual perception

Intelligent visual perception enhances artificial intelligence (AI) systems by providing better visual knowledge, contextual reasoning, and correct interpretation of complicated situations. It also enables multimodal interaction, better decision-making, interpretability, and generalization across many real-world visual tasks.

- **Enhanced Visual Understanding:** Captures objects, scenes, relationships, and contextual information beyond basic classification.
- **Improved Decision-Making:** Provides more visual information for correct and context-aware decisions.
- **Better Feature Representation:** Attention and advanced DL models identify the most relevant visual features.
- **Contextual Reasoning:** Allows models to grasp interactions, spatial linkages, and complicated visual situations.

Table 1 summarizes the main components of intelligent visual perception, including typical approaches and models and their main competencies. It includes visual recognition, segmentation, interpretability, reasoning, generation, and multimodal foundation models.

Table 1. Key components of intelligent visual perception

Component	Representative Techniques/Models	Key Capability
Visual Recognition & Detection	CNNs, ResNet, YOLO, DETR	Object and scene understanding
Segmentation & Scene Understanding	U-Net, Mask R-CNN, SAM	Pixel-level and contextual perception
Attention & Feature Attribution	Self-Attention, Attention Rollout, IG, SHAP	Focused and interpretable perception
Visual Reasoning & Generation	VQA, VCR, NLVR, Diffusion Models	Reasoning and content generation
Multimodal & Foundation Models	CLIP, BLIP, ViT, LLaVA, SAM	Vision–language and generalized intelligence

3. Deep learning for intelligent visual perception

DL has dramatically advanced intelligent visual perception by allowing computers to automatically learn hierarchical visual properties from complex images and videos. CNNs, attention mechanisms, and Vision Transformers assist object identification, detection, segmentation, scene comprehension, and visual reasoning. These models help machines better understand spatial, semantic, and contextual information. Recent advances in multimodal learning better integrate visual and linguistic information, enabling more complete, human-like visual comprehension. DL Models for Intelligent Visual Perception.

3.1 Convolutional neural network

CNNs for the recognition of handwritten numerals. This neural network model was developed using backpropagation and gradient-based techniques. A CNN is a neural network commonly used for difficult image-driven pattern recognition tasks [18]. CNNs are structured to mirror how the human visual cortex processes visual information, drawing significant inspiration from this structure and its organization.

3.2 Recurrent neural network-based model

RNNs can (hopefully) enhance the segmentation map estimate by simulating the short- and long-term relationships among pixels. To represent global context and improve semantic segmentation, RNNs allow sequential processing of connected pixels.

3.3 Attention-based model

Several studies have used attention techniques for semantic segmentation, which is not unexpected given their extensive exploration in computer vision over the years.

3.4 Generative model and adversarial training

GANs have been used for picture segmentation and many other computer vision tasks [19]. The team trained a convolutional semantic segmentation network and an adversarial network to separate the segmentation network's output from the ground-truth maps. This method was used for semantic segmentation.

3.5 Encoder-decoder based model

The convolutional encoder-decoder architecture underpins another popular family of deep models for image segmentation. An encoder-decoder paradigm is used by the majority of DL-based segmentation operations.

3.6 Deep learning for visual reasoning

DL for Visual Reasoning enables models to interpret visual information and perform logical, relational, and contextual reasoning to answer complex questions. Attention mechanisms, relation networks, and neural-symbolic approaches help identify relevant visual features and understand relationships between objects, scenes, and language.

- **Attention-based visual reasoning:** In the context of Image QA, attention mechanisms mostly mean spatial attention to obtain region-specific CNN features or selective terms in a given question relevant to expected answers.
- **Relation-based visual reasoning:** The output of the conditional relation network is eventually used for answer prediction [20]. Although these relation networks for machine reasoning are relatively simple to design, the concept is generic and thus applicable to a variety of tasks.
- **Neural-symbolic visual reasoning:** There is a family of works called Neural Module Networks (NMN) that address compositional visual reasoning based on the structures of language constituents and co-attention.

- Visual question answering: Visual Question Answering (VQA) as a representative task of language and visual reasoning. VQA is a task to assess whether a machine can understand and reason about the visual content present in either a static image or a video through a question-answering process.

3.7 Challenges in deep learning-based visual perception

Several factors limit the effectiveness and practical deployment of DL-based visual perception systems. These include computational complexity, environmental variability, limited and costly training data, difficulties in model interpretability, and the need for robust, safe, and real-time performance in complex environments.

- Real-time performance and computational efficiency: The main challenge in autonomous robots during implementation of DL models for dynamic scene understanding is to provide real-time operation on embedded devices with limited computing units [21].
- Robustness to illumination and environmental variations: Autonomous robots usually face diverse and challenging environments, such as illumination changes and severe weather conditions, all of which severely affect the perception system [22].
- Explainability and interpretability: To build trust and facilitate successful human-robot interaction, having the rationale for the decisions that a robot makes, especially in dynamic and safety-critical scenarios, is most important.
- Data efficiency and synthetic data generation: One of the major issues of DL in robotics is the sheer amount of annotated data for training that is costly and may be laborious to obtain, particularly with heterogeneous and changing real-world situations [23].

3.8 Visual perception computational model for deep learning

The benefits of developing visual perception computational models oriented toward artificial intelligence are two-fold. First, biological visual perception provides a rich source of inspiration for new types of algorithms and architectures [24]. Second, it can be independent of, and complementary to, the mathematical and logic-based methods that have largely dominated traditional approaches to deep neural networks. Second, visual perception can validate existing deep neural network techniques. The main application areas of visual perception computational models focus on artificial intelligence.

- Attention-based computational models: Visual attention refers to the cognitive operations of selectively focusing on crucial information and simultaneously filtering out irrelevant information from visual scenes [25].
- Visual Memory-based Computational Models: Visual memory plays a critical role in senior cognitive functions, since visual memory is the received visual information that is encoded and then restored in the brain, and finally can be used to recall, recognize, categorize, and identify objects.
- Spatio-Temporal Coordination-based Computational Models: Although they belong to different dimensions, they are processed and encoded collaboratively for human perception and recognition. Here, spatial information and temporal information denote a broad sense.

4. Literature review

This section highlights related studies on intelligent visual perception, demonstrating the role of DL and foundation models in enhancing visual understanding, semantic interpretation, multimodal reasoning, explainability, and perception capabilities across diverse application domains.

Yuru Liu et al. [26] proposed a hierarchical visual-semantic alignment framework that improves image interpretation by jointly modeling multi-scale visual features, local-global context, and cross-modal semantic correspondence. It integrates multi-scale feature extraction, token representation, spatial-channel attention, dual embedding projection, and cross-layer feature interaction, along with joint constraint optimization for classification, semantic alignment, and perturbation consistency. Experiments on MS-COCO2017 and NUS-WIDE achieved 85.1% COCO-F1, 87.2% COCO-mAP, 84.2% NUS-F1, and 87.8% NUS-mAP, outperforming ResNet50, Swin-T, CLIP-B/16, and ALBEF.

Haozhong Zhao et al. [27] proposed a Multi-Scale Mixed Attention Network (MMAN) for enhancing underwater visual perception. The network employs a three-branch feature-extraction structure to model multi-scale features and a dual-branch reconstruction architecture to separately restore structural information and enhance fine-grained details. In addition, the authors introduce a Multi-scale Mixed Attention Block (MMAB) to strengthen collaboration among different feature representations.

Guo Yang et al. [28] critically reviewed key advances and trends in visual perception for autonomous driving, focusing on four dimensions: visual scene generation using generative AI, particularly diffusion models, to provide high-fidelity data for system testing, 3D occupancy prediction for fine-grained scene understanding and its benchmark performance; end-to-end driving models that integrate perception and planning, evaluated in complex scenarios; and large-scale models, including world models and multimodal foundation models, and their potential to enhance cognitive intelligence in autonomous systems.

Chi Zhang et al. [29] address explainable evaluation for visual intelligence, which faces challenges from sophisticated, heterogeneous, or inaccessible AI algorithms. The authors propose a latent attribute dictionary learning method constrained by mapping consistency to explain performance variations in visual perception intelligence across test samples. By jointly learning latent concept representations and regressing latent concept-generalization performance, the method establishes relationships among deep representations, semantic attributes, and generalization performance to predict the influence of semantic attributes. The method obtains the optimal solution through alternating optimization.

Changde Du et al. [30] proposed BraVL, a generic neural decoding method that uses multimodal learning of brain, visual, and linguistic features. It models their relationships through multimodal deep generative models and employs a mixture-of-product-of-experts formulation to infer a latent code for coherent joint generation across the three modalities. Intra- and inter-modality mutual information maximization

regularization improves representation consistency and data efficiency with limited brain activity data. BraVL also supports semi-supervised training by incorporating visual and textual features from extra categories.

Byoung-Hee Kwon et al. [31] proposed a visual perception-guided network training approach for decoding visual imagery. Visual perception decreases alpha-band power in the visual cortex over time, whereas visual imagery increases it, resulting in opposite brain activity patterns. The proposed network uses these characteristics. Using visual perception data, the proposed method achieves an average classification performance of 0.7008.

Table 2 summarizes key studies on intelligent visual perception, including their study emphasis, techniques, key results, limitations, and future prospects within the framework of DL- and foundation model-based approaches.

Table 2. Comparative analysis of deep learning and foundation model-based intelligent visual perception

Reference	Study Focus	Approach / Methodology	Key Findings	Challenges and Limitations	Future Directions
Y. Liu et al. [26]	Image interpretation and visual-semantic alignment	Hierarchical alignment using multi-scale features, local-global tokens, attention, dual projection, and cross-layer interaction	Achieved 85.1% COCO-F1, 87.2% COCO-mAP, 84.2% NUS-F1, and 87.8% NUS-mAP, outperforming ResNet50, Swin-T, CLIP-B/16, and ALBEF	Complex architecture, computational cost, and cross-modal alignment challenges	Develop efficient and scalable visual-semantic alignment models
H. Zhao et al. [27]	Underwater visual perception	MMAN with multi-scale feature extraction, dual-branch reconstruction, and mixed attention blocks	Improves structural restoration, fine-grained details, and multi-scale feature collaboration	Complex underwater degradation and high computational cost	Extend MMAN for real-time and diverse underwater perception
G. Yang et al. [28]	Visual perception for autonomous driving	Review of generative AI, diffusion models, 3D occupancy, end-to-end driving, world models, and foundation models	Highlights improvements in scene understanding and autonomous driving intelligence	High computation, complex scenarios, reliability, safety, and generalization	Develop robust world models, multimodal foundation models, and safer driving systems
C. Zhang et al. [29]	Explainable visual intelligence	Latent attribute dictionary learning with mapping-consistency constraints and alternating optimization	Explains the influence of semantic attributes on visual perception generalization	Inaccessible AI mechanisms and dependence on latent concepts	Develop transparent, general, and model-agnostic explainability methods
C. Du et al. [30]	Multimodal brain-visual-linguistic learning	BraVL with multimodal generative models, mixture-of-product-of-experts, and mutual-information regularization	Enables joint generation across modalities and improves data efficiency	Limited brain-activity data and complex multimodal relationships	Improve neural decoding using larger datasets and stronger cross-modal representations
B.-H. Kwon et al. [31]	Visual imagery decoding	Visual perception-guided training using differences in alpha-frequency brain activity	Achieved 0.7008 average visual-imagery classification performance	Moderate performance, brain-signal variability, and dependence on perception/imagery characteristics	Improve decoding using advanced deep learning and personalized models

5. Conclusion

Visual intelligence has advanced well beyond simple image classification to provide a holistic grasp of objects, scenes, connections, geographical information, and semantic context. The study explored the contributions of DL models such as CNNs, RNNs, attention-based models, generative models, and encoder-decoder architectures in improving visual recognition, segmentation, and reasoning. The research also examined feature attribution and attention mechanisms, which improve model interpretability and help systems focus on important visual information. Moreover, multimodal tasks such as visual question answering, visual captioning, visual commonsense reasoning, and visual production all show the increasing integration of vision and language. Recent foundation models and vision-language architectures have enabled transferability, contextual comprehension, and generalized perception capabilities. However, computational complexity, environmental unpredictability, limited training data, interpretability, resilience, and real-time deployment remain key issues. Intelligent visual perception in general is an essential route toward scalable, explainable, multimodal, adaptive, and human-centric visual intelligence systems. Future work should aim to establish efficient, strong foundation models for real-time multimodal perception across a variety of environmental conditions. Greater emphasis is needed on explainable and trustworthy visual reasoning, data-efficient learning, lightweight architectures, and improved cross-modal alignment. Integrating visual memory, spatio-temporal reasoning, and human-centered interaction can further advance adaptive intelligent perception systems.

Ethical issue

The author is aware of and complies with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, figure manipulation, competing interests, and compliance with research ethics policies. The author adheres to publication requirements that the submitted work is original and has not been published elsewhere in any language.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the corresponding author.

Conflict of interest

The author declares no potential conflict of interest.

References

- [1] Y. Hu, H. Chi, and H. Duan, "Metaoptics merging computational optics and optical computing toward intelligent visual perception," *Sci. Adv.*, vol. 12, no. 1, p. eaca8941, 2026, doi: 10.1126/sciadv.aea8941.
- [2] M. N. Paralath, "Towards Interpretable and Attentive Deep Visual Perception Through Iterative Refinement in Convolutional Neural Networks," in 2025 International Conference on Advances in Next-Gen Computer Science (ICANCS), IEEE, Nov. 2025, pp. 1–6. doi: 10.1109/ICANCS65819.2025.11376668.
- [3] Y. Li, X. Guo, H. Zhang, S. Li, and X. Dai, "Active Visual Perception: Opportunities and Challenges," pp. 1–8, 2025, doi: 10.48550/arXiv.2512.03687.
- [4] A. Banan, A. Nasiri, and A. Taheri-Garavand, "Deep learning-based appearance features extraction for automated carp species identification," *Aquac. Eng.*, vol. 89, p. 102053, May 2020, doi: 10.1016/j.aquaeng.2020.102053.
- [5] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern Recognit. Lett.*, vol. 141, pp. 61–67, Jan. 2021, doi: 10.1016/j.patrec.2020.07.042.
- [6] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, 2021, doi: 10.1109/MSP.2020.3016905.
- [7] X. Li et al., "SARPointNet: An Automated Feature Learning Framework for Spaceborne SAR Image Registration," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 6371–6381, 2022, doi: 10.1109/JSTARS.2022.3196383.
- [8] M. Alshayegi, J. Al-Buloushi, A. Ashkanani, and S. Abed, "Enhanced brain tumor classification using an optimized multi-layered convolutional neural network architecture," *Multimed. Tools Appl.*, vol. 80, no. 19, pp. 28897–28917, 2021, doi: 10.1007/s11042-021-10927-8.
- [9] M. Trigka and E. Dritsas, "A Comprehensive Survey of Deep Learning Approaches in Image Processing," *Sensors*, vol. 25, no. 2, p. 531, Jan. 2025, doi: 10.3390/s25020531.
- [10] M. R. C. Mukkolakal, "1.5 Million Messages Per Second on 3 Machines: Benchmarking and Latency Optimization of Apache Pulsar at Enterprise Scale," in 22nd International Conference on Network and Service Management, 2026, pp. 1–4, Mar. doi: 10.48550/arXiv.2603.29113.
- [11] M. Awais et al., "Foundation Models Defining a New Era in Vision: A Survey and Outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2245–2264, 2025, doi: 10.1109/TPAMI.2024.3506283.
- [12] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.
- [13] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," no. 2, pp. 1–14, 2019, doi: 10.48550/arXiv.1908.03557.
- [14] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "MDETR - Modulated Detection for End-to-End Multi-Modal Understanding," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1760–1770, 2021, doi: 10.1109/ICCV48922.2021.00180.
- [15] W. El Amrani, "Attention and Beyond: Explainability Techniques for Vision Transformers. 2025. Conference: Knowledge Extraction and ManagementAt: Strasbourg, FranceVolume: Explain'AI
- [16] Z. Liu and B. Jiang, "Revisiting the Ordering of Channel and Spatial Attention : A Comprehensive Study on Sequential and Parallel Designs," vol. 20, pp. 1–26.
- [17] S. Uppal et al., "Multimodal research in vision and language: A review of current and emerging trends," *Inf. Fusion*, vol. 77, pp. 149–171, Jan. 2022, doi: 10.1016/j.inffus.2021.07.009.
- [18] A. Bose, J. Bhumireddy, and N. Naveen, "A Study on Real-time Object Detection using Deep Learning," pp. 1–34, 2026, doi: 10.48550/arXiv.2602.15926.
- [19] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2021, doi: 10.1109/TPAMI.2021.3059968.
- [20] T. M. Le, "Deep Neural Networks for Visual Reasoning. 2022. doi: 10.48550/arXiv.2209.11990.
- [21] S. Li, X. Chen, Y. Liu, D. Dai, C. Stachniss, and J. Gall, "Multi-Scale Interaction for Real-Time LiDAR Data Segmentation on an Embedded Platform," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 738–745, Apr. 2022, doi: 10.1109/LRA.2021.3132059.
- [22] A. Maham and D. E. N. Tashfa, "Deep Learning Perspective of Scene Understanding in Autonomous Robots," 2025, doi: 10.48550/arXiv.2512.14020.
- [23] T. M. Strat, R. Chellappa, and V. M. Patel, "Vision and Robotics," *AI Mag.*, vol. 41, no. 2, pp. 49–65, Jun. 2020, doi: 10.1609/aimag.v41i2.5299.
- [24] B. Wei, Y. Zhao, K. Hao, and L. Gao, "Visual Sensation and Perception Computational Models for Deep Learning : State of the art , Challenges and Prospects," 2021, doi: 10.48550/arXiv.2109.03391.
- [25] X. Zhu, K. Hao, R. Xie, and B. Huang, "Soft sensor based on eXtreme gradient boosting and bidirectional converted gates long short-term memory self-attention network," *Neurocomputing*, vol. 434, pp. 126–136, Apr. 2021, doi: 10.1016/j.neucom.2020.12.028.
- [26] Y. Liu, Z. Zhao, C. Wang, and N. Du, "A Study on Image Understanding and Visual-Semantic Alignment Modeling Methods for Intelligent Perception," in 2026 International Conference on Intelligent Perception and Autonomous Control (IPAC), 2026, pp. 380–383. doi: 10.1109/IPAC69021.2026.11542612.
- [27] H. Zhao, X. Ji, X. Li, Y. Zhang, and L.-Y. Hao, "MMAN: Multi-scale Mixed Attention Network for Underwater Visual Perception Enhancement," in 2026 41st Youth Academic Annual Conference of Chinese Association of Automation (YAC), IEEE, May 2026, pp. 169–174. doi: 10.1109/YAC71005.2026.11616203.
- [28] G. Yang, Y. Wang, B. Qiu, J.-P. Wu, S. Zheng, and J.-W. Liu, "A Review of Visual Perception and Generation Technologies in Intelligent Transportation Scenarios," in 2025 8th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI), 2025, pp. 1–7. doi: 10.1109/ACAI68217.2025.11406561.

- [29] C. Zhang et al., “Global-Mapping-Consistency-Constrained Visual-Semantic Embedding for Interpreting Autonomous Perception Models,” IEEE Open J. Intell. Transp. Syst., vol. 5, pp. 393–408, 2024, doi: 10.1109/OJITS.2024.3418552.
- [30] C. Du, K. Fu, J. Li, and H. He, “Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 9, pp. 10760–10777, Sep. 2023, doi: 10.1109/TPAMI.2023.3263181.
- [31] B.-H. Kwon, B.-H. Lee, J.-H. Cho, and J.-H. Jeong, “Decoding Visual Imagery from EEG Signals using Visual Perception Guided Network Training Method,” in 2022 10th International Winter Conference on Brain-Computer Interface (BCI), IEEE, Feb. 2022, pp. 1–5. doi: 10.1109/BCI53720.2022.9735014.

Disclaimer/Publisher’s Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Future Publishing LLC (Future) and/or the editor(s). Future and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.



This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).