technologies

MDPI

Review

A Review of Machine Learning and Deep Learning for Object Detection, Semantic Segmentation, and Human Action Recognition in Machine and Robotic Vision

Nikoleta Manakitsa [1], George S. Maraslidis [1], Lazaros Moysis [2,3] and George F. Fragulis [1,*]

[1] Department of Electrical and Computer Engineering, University of Western Macedonia, 50100 Kozani, Greece; nikoleta_m@hotmail.gr (N.M.); dece00079@uowm.gr (G.S.M.)
[2] Laboratory of Nonlinear Systems-Circuits and Complexity, Physics Department, Aristotle University of Thessaloniki, 54624 Thessaloniki, Greece; lmousis@physics.auth.gr
[3] Department of Mechanical Engineering, University of Western Macedonia, ZEP Campus, 50100 Kozani, Greece
* Correspondence: gfragulis@uowm.gr

Abstract: Machine vision, an interdisciplinary field that aims to replicate human visual perception in computers, has experienced rapid progress and significant contributions. This paper traces the origins of machine vision, from early image processing algorithms to its convergence with computer science, mathematics, and robotics, resulting in a distinct branch of artificial intelligence. The integration of machine learning techniques, particularly deep learning, has driven its growth and adoption in everyday devices. This study focuses on the objectives of computer vision systems: replicating human visual capabilities including recognition, comprehension, and interpretation. Notably, image classification, object detection, and image segmentation are crucial tasks requiring robust mathematical foundations. Despite the advancements, challenges persist, such as clarifying terminology related to artificial intelligence, machine learning, and deep learning. Precise definitions and interpretations are vital for establishing a solid research foundation. The evolution of machine vision reflects an ambitious journey to emulate human visual perception. Interdisciplinary collaboration and the integration of deep learning techniques have propelled remarkable advancements in emulating human behavior and perception. Through this research, the field of machine vision continues to shape the future of computer systems and artificial intelligence applications.

Keywords: machine vision; computer vision; image processing; object classification; object detection; object segmentation; pattern recognition; artificial intelligence; machine learning; deep learning; robotics; mechatronics

Citation: Manakitsa, N.; Maraslidis, G.S.; Moysis, L.; Fragulis, G.F. A Review of Machine Learning and Deep Learning for Object Detection, Semantic Segmentation, and Human Action Recognition in Machine and Robotic Vision. *Technologies* **2024**, *12*, 15. https://doi.org/10.3390/technologies12020015

Academic Editor: Francesco Aggogeri

Received: 9 October 2023
Revised: 27 December 2023
Accepted: 10 January 2024
Published: 23 January 2024

1. Introduction

Computer vision, through digital image processing, empowers machines to map surroundings, identify obstacles, and determine their positions with high precision [1,2]. This multidisciplinary field integrates computer science, artificial intelligence, and image analysis to extract meaningful insights from the physical world, empowering computers to make informed decisions [3]. Real-time vision algorithms, applied in domains like robotics and mobile devices, have yielded significant results, leaving a lasting impact on the scientific community [4].

The study of computer vision presents numerous complex challenges and inherent limitations. Developing algorithms for tasks such as image classification, object detection, and image segmentation requires a deep understanding of the underlying mathematics. However, it is important to acknowledge that each computer vision task requires a unique approach, which adds complexity to the study itself. Therefore, a combination of theoretical knowledge and practical skills is crucial in this field, as it leads to advancements in artificial intelligence and the creation of impactful real-world applications.

The field of computer vision has been greatly influenced by earlier research efforts. In the 1980s, significant advancements were made in digital image processing and the analysis of algorithms related to image understanding. Prior to these breakthroughs, researchers worked on mathematical models to replicate human vision and explored the possibilities of integrating vision into autonomous robots. Initially, the term "machine vision" was primarily associated with electrical engineering and industrial robotics. However, over time, it merged with computer vision, giving rise to a unified scientific discipline. This convergence of machine vision and computer vision has led to remarkable growth, with machine learning techniques playing a pivotal role in accelerating progress. Today, real-time vision algorithms have become ubiquitous, seamlessly integrated into everyday devices like mobile phones equipped with cameras. This integration has transformed how we perceive and interact with technology [4].

Machine vision has revolutionized computer systems, empowering them with advanced artificial intelligence techniques that surpass human capabilities in various specific tasks. Through computer vision systems, computers have gained the ability to perceive and comprehend the visual world [3].

The overarching goals of computer vision are to enable computers to see, recognize, and comprehend the visual world in a manner analogous to human vision. Researchers in machine vision have dedicated their efforts to developing algorithms that facilitate these visual perception functions. These functions include image classification, which determines the presence of specific objects in image data; object detection, which identifies instances of semantic objects within predefined categories; and image segmentation, which breaks down images into distinct segments for analysis. The complexity of each computer vision task, coupled with the diverse mathematical foundations involved, poses significant challenges to their study. However, understanding and addressing these challenges holds great theoretical and practical importance in the field of computer vision.

The contribution of this work is a presentation of the literature that showcases the current state of research of machine learning and deep learning methods for object detection, semantic segmentation, and human action recognition in machine and robotic vision. In this paper, we present a comprehensive overview of the key elements that constitute machine vision and the technologies that enhance its performance. We discuss innovative scientific methods extensively utilized in the broad field of machine and deep learning in recent years, along with their advantages and limitations. This review not only adds new insights into machine learning and deep learning methods in machine/robotic vision but also features real-world applications of object detection, semantic segmentation, and human action recognition. Additionally, it includes a critical discussion aimed at advancing the field.

This paper's organizational structure is as follows. Section 2 offers an overview of machine learning/deep learning algorithms and methods. Section 3 comprehensively covers object detection, image, and semantic segmentation algorithms and methods, with a specific focus on human action recognition methods. Section 4 introduces detailed notions regarding robotic vision. Section 5 presents Hubel and Wiesel's electrophysiological insights, Van Essen's map of the brain, and their impact on machine/robotic vision. Section 6 presents a discussion regarding the aforementioned topics. Lastly, Section 7 addresses the current challenges and future trends in the field.

2. Machine Learning/Deep Learning Algorithms

Various AI algorithms facilitate pattern recognition in machine vision and can be broadly categorized into supervised and unsupervised types. Supervised algorithms, which leverage labeled data to train models predicting the class of input images, can be further divided into parametric (assuming data distribution) and non-parametric methods. Examples include k-nearest neighbors, support vector machines, and neural networks. Unsupervised algorithms, which lack labeled data, unveil patterns or structures and can be categorized into clustering and dimensionality reduction methods. Comparative analyses

assess technologies based on metrics like accuracy and scalability, with the optimal choice dependent on the specific problem and resources. Initially met with skepticism, public perception of AI's benefits has shifted positively over time. Artificial intelligence aims to replicate human intelligence, with vision being a crucial aspect. Exploring the link between computer vision and AI, the latter comprises machine learning and deep learning subsets, essential for understanding machine vision's progress (see Figure 1).

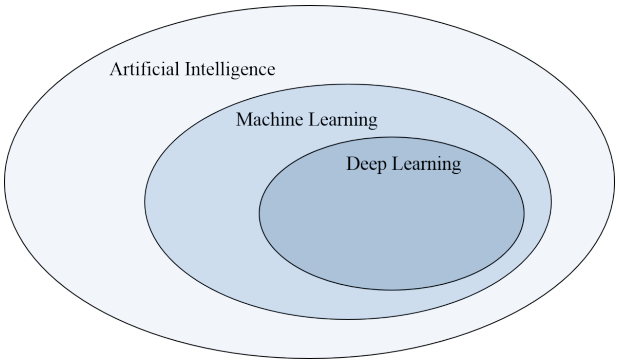

Figure 1. Relationship between artificial intelligence, machine learning, and deep learning.

The terms artificial intelligence, machine learning, and deep learning are often mistakenly used interchangeably. To grasp their relationship, it is helpful to envision them as concentric circles. The outermost circle represents artificial intelligence, which was the initial concept. Machine learning, which emerged later, forms a smaller circle that is encompassed by artificial intelligence. Deep learning, the driving force behind the ongoing evolution of artificial intelligence, is represented by the smallest circle nested within the other two.

2.1. Machine Learning

Human nature is marked by the innate ability to learn and progress through experiences. Similarly, machines possess the capacity for improvement through data acquisition, a concept known as machine learning (ML). ML, a subset of artificial intelligence, empowers computers to autonomously detect patterns and make decisions with minimal human intervention. Algorithms undergo training through exposure to diverse situations, refining understanding with more data, leading to enhanced accuracy. Organizations adopt ML for automated, efficient operations. Computer vision applications, like facial recognition and image detection, showcase ML's impact. Image analysis identifies facial features for applications such as smartphone unlocking and security systems. In autonomous vehicles, image detection recognizes objects in real time, enabling informed decisions. ML embraces supervised learning, making inferences based on past data, and unsupervised learning, identifying patterns without labeled guidance, offering versatility in various domains.

2.2. Deep Learning

Deep learning, an evolution of machine learning, surpasses shallow neural networks by employing intricate algorithms that mirror human cognitive processes. These algorithms, forming deep neural networks, emulate the logical structure of the human brain, enabling them to draw conclusions by analyzing data. Unlike traditional machine learning, which relies on manually extracted features, deep learning operates on an end-to-end learning framework, minimizing human intervention. The architecture of deep neural networks consists of multiple interconnected layers with non-linearity, enhancing their capacity to learn complex patterns. In contrast, traditional machine learning, represented by shallow neural networks, involves step-by-step feature extraction and model construction with human-designed features. Computer vision utilizes "manual features" for precise identification within images, a process distinct from the automatic feature learning of deep neural networks. The comparison presented in Figures 2 and 3 underscores the automatic

nature of deep learning, driven by data and minimal user involvement, whereas traditional machine learning relies on human-crafted features and a more manual, stepwise process.

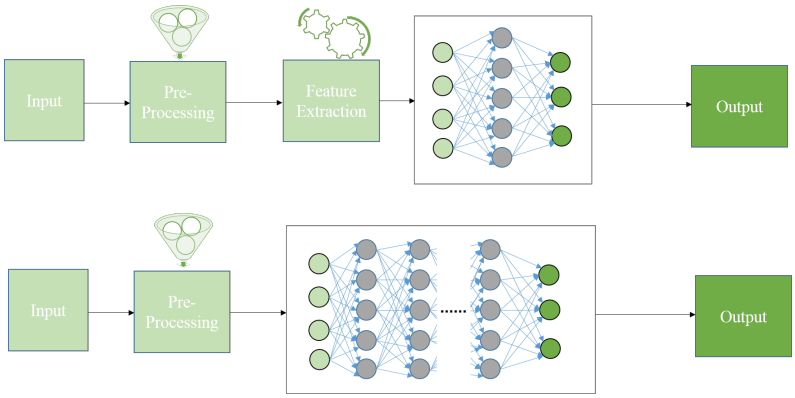

Figure 2. Comparison between a shallow neural network (first image above) and deep learning (second image below).

	Way of Training			
Traditional Machine Learning	Feature extraction from humans	Process performed step-by-step	Works with less data	Automatic classification
Deep Learning	Data train the neurons	Fully automatic process	Large amount of data required	Possible error helps to adjust the model

Figure 3. Supervised machine learning vs. supervised deep learning.

2.3. Vision Applications Using Deep Learning Methods

Although the term "deep learning" initially referred to the depth of the neural network (number of hidden layers), it has evolved to encompass a broader class of machine learning techniques that utilize neural networks with multiple layers to model and solve complex problems. The relevance of deep learning spans various domains and applications. Here are some of the most relevant problems and applications where deep learning has demonstrated a significant impact:

- Image Recognition and Classification: Deep learning, especially convolutional neural networks (CNNs), excels in tasks like image classification, object recognition, facial recognition, and medical image analysis [4].
- Autonomous Vehicles: Deep learning, particularly CNNs, plays a crucial role in perception tasks for autonomous vehicles, enabling object detection, segmentation, and recognition [5].
- Medical Image Analysis: Deep learning, especially CNNs, is applied in tasks such as tumor detection, pathology recognition, and organ segmentation in medical image analysis [6].
- Generative Modeling: Generative models like GANs and VAEs are used for image synthesis, style transfer, and the generation of realistic data samples [7].
- Reinforcement Learning: Deep reinforcement learning successfully trains agents for game playing, robotic control, and optimizing complex systems through interaction with the environment [8].
- Human Activity Recognition: Deep learning models, especially RNNs and 3D CNNs, recognize and classify human activities from video or sensor data, with applications in healthcare, surveillance, and sports analytics [9].

These are just a few examples, and the versatility of deep learning continues to expand as researchers and practitioners explore new applications and architectures. The success of deep learning in these domains is attributed to its ability to automatically learn hierarchical representations from data, capturing complex patterns and relationships.

3. Object Detection, Semantic Segmentation, and Human Action Recognition Methods

Digital image processing algorithms have been transformative in machine vision and computer vision, reshaping visual perception and enabling machines to comprehend and

analyze images [10,11]. Originating from image processing, these algorithms have driven progress in pattern recognition, object detection, and image classification, ushering in a paradigm shift. Machine vision leverages intricate techniques and mathematical models, bridging the gap between human visual systems and machine intelligence. By extracting meaning from visual stimuli, computer vision has transformed our understanding of artificial intelligence's visual realms. Images convey diverse information, including colors, shapes, and recognizable objects, analogous to how the human brain interprets emotions and states. In machine vision, algorithms analyze digital images to extract information based on user-defined criteria. Object detection, face detection, and color recognition are some examples, illustrating the system's dependence on specific patterns for information extraction [12]. The process involves detecting patterns representing objects, with the detailed steps outlined in Figure 4.

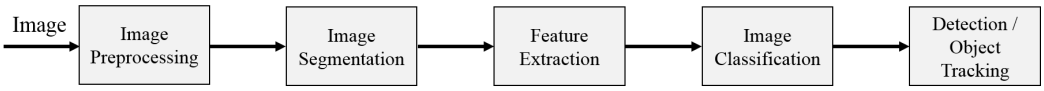

Figure 4. Steps in machine vision.

3.1. Image Preprocessing

Image preprocessing plays a vital role in refining images before applying pattern recognition algorithms, aiming to enhance quality, reduce noise, correct illumination, and extract relevant features [13]. Common techniques include filtering, histogram equalization, edge detection, and morphological operations. A robust mathematical foundation is essential for effective image analysis, laying the groundwork for the subsequent steps. This foundation determines the color space and model, representing colors mathematically. Color models like RGB, HSI, and HSV define colors precisely using variables, forming color spaces. The RGB model is composed of red, green, and blue components. It combines the intensity levels of these components to create colors. The full strength of all three yields white, whereas their absence results in black [14,15]. The process ensures a comprehensive understanding of image content and sets the stage for employing algorithms in image analysis. Figures 5 and 6 illustrate the RGB color model's primary colors, their combinations, and a sample RGB model color space.

Figure 5. An RGB image, its red , green and blue component [15].

The HSI and HSV models aim to approximate human perception by considering characteristics such as hue (H), saturation (S), intensity (I), brightness (B), and value (V). In the HSI model, the hue component ranges from 0° to 360°, determining the color's hue, whereas saturation (S) expresses the mixing degree of a primary color with white (Figure 7). The intensity (I) component denotes light intensity without conveying color information [16]. The HSI model, depicted as a double cone, exhibits upper and lower peaks corresponding to white (I = 1) and black (I = 0), with maximum purity (S = 1) at I = 0.5 (Figure 8).

Figure 9 showcases the HSI model in a real photo, depicting HSV channels as grayscale images, revealing color saturation and modified color intensity for a clearer representation. The HSV model calculates the brightness component differently from HSI, primarily managing hue and chromatic clarity components for digital tasks like histogram balancing.

The HSV color space positions black at the inverted cone's top (V = 0) and white at the base (V = 1). The hue component (hue) for red is 0°, differing by 180° from the

complementary colors. Saturation (S) is determined by the distance from the cone's base, simplifying color representation and extraction in object detection compared to the RGB color space [14,17].

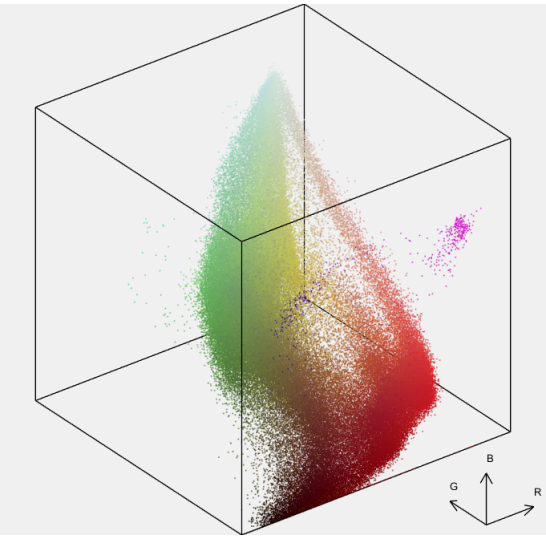

Figure 6. RGB model's color space in a pepper image.

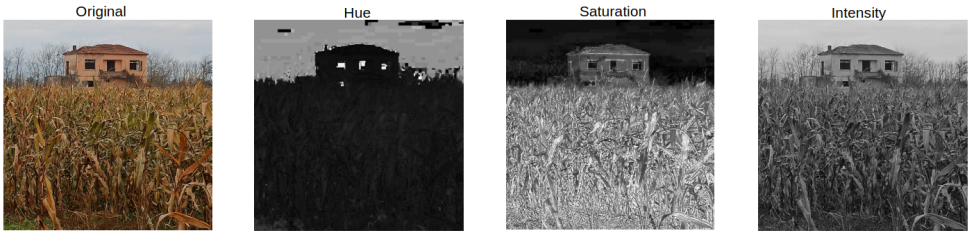

Figure 7. Color components of the HSI model on a face: hue, saturation, and intensity.

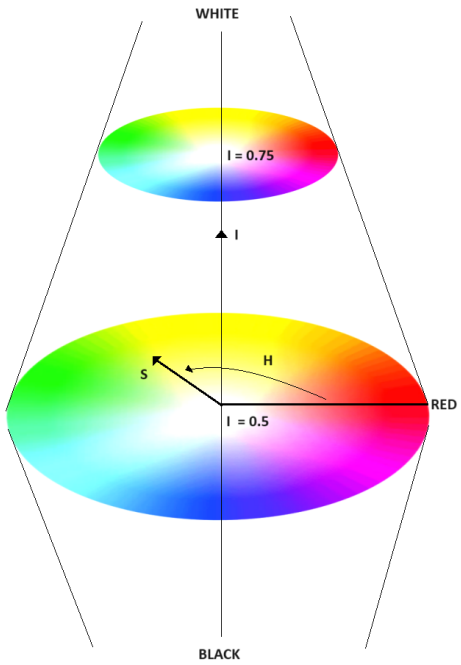

Figure 8. HSI model's color space [18].

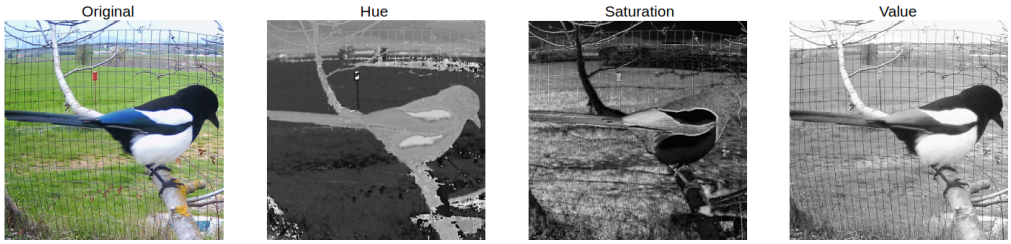

Figure 9. Color components of the HSI model.

3.2. Image Segmentation

Image segmentation is the pivotal process of partitioning an image into segments or regions with similar characteristics, aiding in foreground-background separation, object grouping, boundary location, and more. The mathematical methods usually used in image segmentation are given in Table 1. Techniques such as thresholding, clustering, region growing, and watershed are common for this purpose. The segmentation's primary aim is to simplify image information, thereby facilitating subsequent analyses and reducing complexity. Successful segmentation, crucial for efficient image analysis, involves dividing the image into homogeneous regions, ideally corresponding to objects like faces. It plays a vital role in object identification and boundary delineation, assigning labels to pixels with common visual characteristics. The algorithms for image segmentation fall into two groups: boundary-based (edge and object detection) and region-based (thresholding, expansion, division, merging, watershed) algorithms (see Table 2 for main algorithms) . Robust mathematical foundations, incorporating clustering, edge detection, and graph-based algorithms, are imperative for successful image segmentation, object detection, and image classification [19].

Table 1. Mathematical methods used in image segmentation.

Mathematical Method	Application in Image Segmentation
Graph Theory [20]	Graph cuts: partitions images into segments by minimizing an energy function.
Probability and statistics [21,22]	Bayesian methods: model pixel likelihood based on statistical properties. Gaussian Mixture Models (GMMs): represent pixel intensity distribution.
Partial Differential Equations (PDEs) [23,24]	Chan–Vese model: level-set method for segmentation, particularly for smooth object boundaries. Active contour models (Snakes): PDE-based models that evolve contours for boundary identification.
Clustering algorithms [25]	K-means clustering: unsupervised clustering for grouping pixels based on similarity. Mean-shift clustering: adaptive clustering method for image segmentation.
Fourier transform [26]	Frequency domain segmentation: transforms images for segmentation based on frequency characteristics.
Markov Random Fields (MRFs) [27]	MRF-based segmentation: models pixel dependencies for improved segmentation.
Distance metrics [28]	Watershed algorithm: segments images into regions based on distance metrics.
Convolutional neural networks (CNNs) [29]	Fully Convolutional Network (FCN): adapts CNNs for pixel-wise classification. U-Net: specialized architecture for biomedical image segmentation.
Level-set methods [30]	Geodesic active contour: combines level-set methods with geodesic active contours for accurate segmentation.
Fuzzy logic [31]	Fuzzy c-means clustering: fuzzy logic-based algorithm for uncertain boundary images.

Table 2. Image segmentation algorithms.

Image Segmentation Algorithms	
Area Boundary Detection Algorithms	**Algorithms Based on Regions**
Edge detection [32]	Downgrading [33]
Detection of objects [34]	Expansion of areas [35]
	Division and merging of areas [36]
	Segmentation based on Watershed [28]

Image segmentation, crucial for analyzing objects through mathematical models, results in a binary image based on features like texture and color. The process can utilize color or color intensities, and the histogram-based method, which constructs a histogram from all pixels, aids in identifying common pixels in the image. In medical applications, such as chest X-rays, histogram-based segmentation is prevalent [37]. When segmenting based on color, one-dimensional histograms are obtained for monochrome images, whereas color images require three histograms for each channel. Peaks or valleys in the histogram assist in identifying objects and backgrounds. Multicolor images involve processing individual RGB histograms, combining results to select a robust segmentation hypothesis. Segmentation based on pixel intensity is less complex, as evidenced in black-and-white images where relatively large-sized objects create pixel distributions around their average intensity values [38].

3.2.1. Feature Extraction

In machine learning, pattern recognition, and image processing, feature extraction is vital. Following image segmentation, this process transforms a large input dataset into a reduced set of features or a feature vector. Distinct image components, like lines and shapes, are identified, each of which is assigned a normalized value (e.g., perimeter, pixel coverage). Algorithms consider that each pixel has an 8-bit value, reducing information in a 640 × 320 image to a focused feature vector. Feature extraction facilitates generalized learning by isolating relevant information and describing the image in a structured manner (Figure 10) [39,40].

Figure 10. Original image (**left**); representation of the boundaries of the regions on the original image (**top-right image**); segmentation result, where each uniform region is described by an integer and all pixels in the region have this value (**bottom-right image**).

Some algorithms commonly used for feature extraction include:

- Histogram of Oriented Gradients (HOG);
- Scale-Invariant Feature Transform (SIFT);
- Speeded-Up Robust Features (SURF) ([39,41]).

3.2.2. Image Classification

Humans effortlessly perceive the three-dimensional world, distinguishing objects and recognizing emotions. In computer vision, recognition tasks involve feature extraction, such as identifying a cat. However, distinguishing between a cat and a dog requires a database and specific classification mechanisms. Learning is crucial, with algorithms like Learning and Classification Algorithms (LCAs) being employed based on the application domain. Image classification categorizes images into predefined classes by utilizing mathematical techniques and neural networks. Challenges persist in emulating human visual system complexities, and LCAs can adapt to factors like class linearity. Image classification relies on a multidisciplinary approach, integrating feature extraction, data representation, and model training (see Table 3).

Some commonly used classification algorithms include [42]:

- Naive Bayes classifier;
- Decision trees;
- Random forests;
- Neural networks;
- Nearest neighbor or k-means.

Table 3. Mathematical methods used in image classification.

Mathematical Method	Application in Image Classification
Linear algebra [43]	Vectors and matrices: represent images and perform matrix operations for pixel manipulation. Eigenvalue decomposition: dimensionality reduction (e.g., PCA, SVD).
Statistics [44]	Probability and statistics: model feature distributions, probabilistic models (e.g., Naive Bayes, Gaussian Mixture Models).
Calculus [45]	Gradient descent: Optimization during machine learning model training.
Machine learning algorithms [46]	Support vector machine (SVM): Finds hyperplanes in feature space. Decision trees and random forests: use mathematical decision rules for classification.
Neural networks [47]	Backpropagation: updates weights during neural network training. Activation functions: introduce non-linearity in neural networks.
Signal processing [48]	Fourier transform: extracts features in the frequency domain. Wavelet transform: captures high- and low-frequency components.
Distance metrics [49]	Euclidean distance: measures similarity between feature vectors. Mahalanobis distance: considers correlations between features.

3.2.3. Object Detection

Object detection in computer vision identifies and labels objects in images or videos, employing algorithms that extract and process information, with an emphasis on specific

aspects. This technique enables the counting, tracking, and precise labeling of objects. The process, termed 'object detection' or 'recognition', employs mathematical techniques like convolution, spatial transformations, and machine learning algorithms. Specific instances, such as 'face detection' or 'car detection', focus on extracting information related to faces or cars. Mathematical concepts essential for object detection include convolutions, spatial transformations, and machine learning algorithms like support vector machines and decision trees (see Table 4) [39,50] (Figure 11).

Figure 11. Example of detecting objects in an image [51].

Table 4. Mathematical methods used in object detection.

Mathematical Method	Application in Object Detection
Linear algebra [52]	Vectors and matrices: represent images and use linear algebra operations for processing. Geometric transformations: perform translation, rotation, and scaling for object manipulation.
Calculus [53]	Gradient descent: used for optimization during model training. Partial derivatives: used for position determination in localization tasks.
Statistics [54]	Probability and statistics: employ probabilistic models and statistical methods for feature analysis. Non-maximum suppression: removes duplicate or low-confidence bounding boxes.
Machine learning algorithms [55]	Region-based CNNs (R-CNNs): propose and classify bounding boxes. Fast R-CNN, Faster R-CNN: enhance speed and accuracy.
Geometry and trigonometry [56]	Trigonometric functions: use geometric calculations for object poses and orientations. Homography: uses transformations for image rectification or feature matching.
Optimization techniques [57]	Integer Linear Programming (ILP): refines bounding-box selection.
Deep learning architectures [58]	Feature Pyramid Networks (FPNs): capture features at multiple scales. Single-Shot Multibox Detector (SSD), You Only Look Once (YOLO): efficiently detect objects in real time.

3.2.4. Object Tracking

Following object detection, visual object tracking involves recognizing and estimating the states of moving objects within a visual scene. It plays a vital role in machine vision applications, tracking entities like people and faces. Despite recent advancements, challenges persist due to factors like environmental conditions, object characteristics, and non-linear motion. Challenges include variations in background, lighting conditions, rigid object parts, and interactions with other objects. Effectively predicting future trajectories requires assessing and determining the most suitable tracking algorithms (see Figure 12) [59,60].

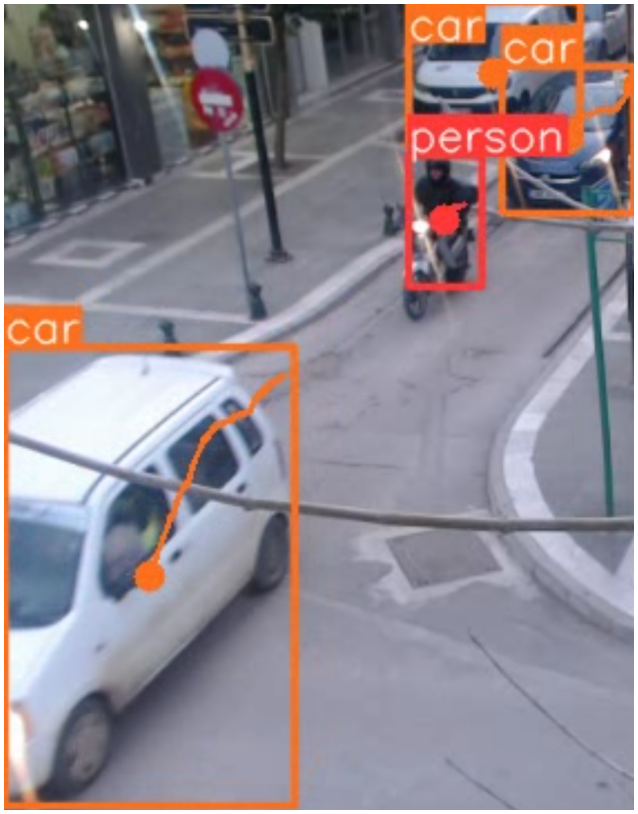

Figure 12. Example of tracking objects in an image.

3.3. Human Action Recognition

Human action recognition in computer vision, crucial for surveillance, human–computer interaction, and sports analysis, benefits from advanced deep learning techniques. By leveraging CNNs and RNNs like LSTM, these models can effectively capture spatial and temporal information from video sequences. By extracting features representing motion and appearance patterns, they can enhance accuracy. The fusion of modalities like RGB and depth information further refines recognition. Recent strides in attention mechanisms and metaheuristic algorithms have optimized network architectures, emphasizing relevant regions for improved performance [9,61–72].

There are also various other approaches regarding HAR.

In [73], a two-stream attention-based LSTM network was proposed for deep learning models, enhancing feature distinction determination. The model integrates a spatiotemporal saliency-based multi-stream network with an attention-aware LSTM, utilizing a saliency-based approach to extract crucial features and incorporating an attention mechanism to prioritize relevance. By introducing an LSTM network, temporal information and long-term dependencies are captured, improving accuracy in distinguishing features and enhancing action differentiation. In [74], a hybrid deep learning model for human action recognition was introduced, achieving 96.3% accuracy on the KTH dataset. The model focuses on precise classification through robust feature extraction and effective learning,

leveraging the success of deep learning in various contexts. The authors of [75] proposed a framework for action recognition, utilizing multiple models to capture both global and local motion features. A 3D CNN captures overall body motion, whereas a 2D CNN focuses on individual body parts, enhancing recognition by incorporating both global and local motion information. Furthermore, [76] drew inspiration from deep learning achievements, proposing a CNN-Bi-LSTM model for human activity recognition. Through end-to-end training, the model refines pre-trained CNN features, demonstrating exceptional accuracy in recognizing single- and multiple-person activities on RGB-D datasets. In [77], a novel hybrid architecture for human action recognition was introduced, combining four pre-trained network models through an optimized metaheuristic algorithm. The architecture involves dataset creation, a deep neural network (DNN) design, training optimization, and performance evaluation. The results demonstrate its superiority over existing architectures in accurately predicting human actions. The authors of [78] presented a key contribution with temporal-spatial mapping, capturing video frame evolution. The proposed temporal attention model within a convolutional neural network achieved remarkable performance, surpassing a competing baseline method by 4.2% in accuracy on the challenging HMDB51 dataset. In [79] authors tackled still image-based human action recognition challenges using transfer learning and data augmentation and by fine-tuning CNN architectures. The proposed model outperformed prior benchmarks on the Stanford 40 and PPMI datasets, showcasing its robustness. Finally, [80] introduced the cooperative genetic algorithm (CGA) for feature selection, employing a cooperative approach that enhances accuracy, reduces overfitting, and improves resilience to noise and outliers. CGA offers superior feature selection outcomes across various domains.

The main human action recognition methods are presented in Table 5, while their characteristics, including their advantages, disadvantages/limitations, and complexities, are given in Appendix A.

Table 5. Human action recognition methods.

Method	References
Deep learning (CNNs and RNNs) addresses the critical task of human action recognition in computer vision, enhancing accuracy and optimizing performance.	[9,61–72]
Attention-based LSTM for feature distinctions, incorporating a spatiotemporal saliency-based multi-stream network.	[73]
A hybrid deep learning model for human action recognition.	[74]
Utilizes multiple models to capture global and local motion features for action recognition.	[75]
Uses RGB frames, Bi-LSTM, and a CNN for action recognition.	[76]
A novel hybrid architecture combining four pre-trained network models, predicting human actions.	[77]
Uses a temporal-spatial mapping operation for action recognition.	[78]
Use of image-based HAR through transfer learning.	[79]
A cooperative approach for feature selection.	[80]

3.4. Semantic Segmentation

In computer vision, the fundamental challenges are image classification, object detection, and segmentation, each escalating in complexity [12,81–85]. Object detection involves labeling objects and determining their locations, whereas image segmentation delves deeper, precisely delineating object boundaries. Image segmentation can be classified into two techniques: semantic segmentation, which assigns each pixel to a specific label, and instance segmentation, which uniquely labels each instance of an object. Semantic segmentation plays a vital role in perceiving and interpreting images, crucial for appli-

cations like autonomous driving and medical imaging. Convolutional neural networks, especially in deep learning, have significantly advanced semantic segmentation, providing high-resolution mapping for various applications, including YouTube stories and scene understanding [86–92]. This technique finds applications in diverse areas, such as document analysis, virtual makeup, self-driving cars, and background manipulation in images, showcasing its versatility and importance. Semantic segmentation architectures typically involve an encoder network, which utilizes pre-trained networks like VGG or ResNet, and a decoder network, which projects learned features onto the pixel space, enabling dense pixel-level classification [86–92].

The three main approaches are :

1. Region-Based Semantic Segmentation

Typically, region-based approaches use the "segmentation using recognition" pipeline. In this method, free-form regions are extracted from an image and described before being subjected to region-based classification. The region-based predictions are transformed into pixel predictions during testing by giving each pixel a label based on the region with the highest score to which it belongs [86,87,93–96].

2. Fully Convolutional Network-Based Semantic Segmentation

The original Fully Convolutional Network (FCN) does not require region proposals because it learns a mapping from pixels to pixels. By enabling it to handle images of any size, the FCN expands the capabilities of a conventional CNN. FCNs only use convolutional and pooling layers, as opposed to traditional CNNs, which use fixed fully connected layers, allowing predictions on inputs of any size [92,97–100].

3. Weakly Supervised Semantic Segmentation

Many semantic segmentation methods depend on pixel-wise segmentation masks, which are laborious and costly to annotate. To address this challenge, weakly supervised methods have emerged. These approaches leverage annotated bounding boxes to achieve semantic segmentation, providing a more efficient and cost-effective solution [50,63,90–92,101–107].

Some other approaches are discussed below.

In [108], the authors discussed the application of deep learning for the semantic segmentation of medical images. They outlined crucial steps for constructing an effective model and addressing challenges in medical image analysis. Deep convolutional neural networks (DCNNs) in semantic segmentation were explored in [109], where models like UNet, DeepUNet, ResUNet, DenseNet, and RefineNet were reviewed. DCNNs proved effective in semantic segmentation, following a three-phase procedure: preprocessing, processing, and output generation. Ref. [110] introduced CGBNet, a segmentation network that enhanced performance through context encoding and multi-path decoding. The network intelligently selects relevant score maps and introduces a boundary delineation module for competitive scene segmentation results.

The main semantic segmentation methods are presented in Table 6, while their characteristics, including their advantages, disadvantages/limitations, and complexities, are given in Appendix B.

Table 6. Semantic segmentation methods.

Summary	References
Identify fundamental computer vision problems: image classification, object detection, and segmentation.	[12,81–85]
Semantic segmentation assigns labels to every pixel, significantly enhanced by deep learning, particularly CNNs.	[86–92]
Describe components of a semantic segmentation architecture and three main approaches: region-based, FCN-based, and weakly supervised.	[50,63,86,87,90–92,97–107]
Semantic segmentation, focusing on medical image analysis and DCNNs.	[108–110]

3.5. Automatic Feature Extraction

Automatic feature extraction methods play a crucial role in robotic vision by helping neural networks (NNs) effectively process and understand visual information. Following is an overview of how these methods are used in conjunction with neural networks:

1. Preprocessing:Image Enhancement.Methods like histogram equalization and noise reduction improve image quality, aiding neural networks in extracting meaningful features.
2. Feature Extraction:
   - Traditional Techniques: Edge and corner detection and texture analysis extract relevant features, capturing crucial visual information.
   - Deep Learning-Based Techniques: Convolutional neural networks (CNNs) learn hierarchical features directly from raw pixel data, covering both low-level and high-level features.
3. Data Augmentation: Automatic feature extraction is integrated into data augmentation, applying techniques like rotation and scaling to diversify the training dataset.
4. Hybrid Models: Hybrid models combine traditional computer vision methods with neural networks, leveraging the strengths of both for feature extraction and classification.
5. Transfer Learning: Pre-trained neural networks, especially in computer vision tasks, can be fine-tuned for specific robotic vision tasks, saving training time and resources.
6. Object Detection and Recognition: Automatic feature extraction contributes to object detection, as seen in region-based CNNs (R-CNNs) using region proposal networks and subsequent feature extraction for classification.
7. Semantic Segmentation: In tasks like semantic segmentation, automatic feature extraction aids the neural network in understanding context and spatial relationships within an image.

By integrating automatic feature extraction methods with neural networks, robotic vision systems can efficiently process visual information, understand complex scenes, and perform tasks such as object recognition, localization, and navigation. This combination of techniques allows for more robust and accurate vision-based applications in robotics.

4. Robotic Vision Methods

Robotic vision algorithms serve three primary functions in visual perception. In this subsection, we explore and examine examples of each of these functions [3].

4.1. Pattern Recognition—Object Classification

Pattern recognition in machine vision is the process of identifying and classifying objects or patterns in images or videos using machine learning algorithms. Pattern recognition can be used for various applications, such as object detection, face recognition, optical character recognition, biometric authentication, etc. [111,112]. Pattern recognition can also be used for image preprocessing and image segmentation, which are essential steps for many computer vision tasks [113–115].

Robotic vision is based on pattern recognition. It is necessary to classify the data into different categories to make it easier to use appropriate algorithms to select the right decisions.

Originally, two approaches were founded for the implementation of a pattern recognition system. Statistical pattern recognition is based on underlying statistical models to describe the patterns and their classes. The first pattern is the theoretical decision. In the second approach, the classes are represented by formal structures such as grammar and strings. This approach is called syntactic pattern recognition, otherwise defined as a structural approach. The third approach was developed later, and it has experienced rapid development in recent years. It is called neural pattern recognition. In this approach, the classifier is depicted as a network of small autonomous units that perform a small number of specific actions, i.e., "cells" that mimic the neurons of the human brain.

Classifying objects belongs to a biological capacity of the human system that refers to visual perception. It is a very important function in the field of computer vision, aiming

to automatically classify images into predefined categories. For decades, researchers have developed advanced techniques to improve the quality of classification. Traditionally, classification models can only perform well on small datasets, such as CIFAR-10 [116] and MNIST [117]. The biggest leap forward in the development of image classification occurred when the large-scale image dataset "ImageNet" was created by Feifei Li in 2009 [106].

An equally important and challenging task in computer vision is object detection, which involves identifying and localizing objects from either a large number of predefined categories in natural images or for a specific object. Object detection and image classification face a similar technological challenge: both need to handle a wide variety of objects. However, object detection is more challenging compared to image classification because it requires identifying the exact target object being searched for [19]. Most research efforts have focused on detecting a single class of object data, such as pedestrians or faces, by designing a set of suitable features. In these studies, objects are detected using a set of predefined patterns, where the features correspond to a location in the image or a feature pyramid.

Object classification identifies the objects present in the visual scene, whereas object detection reveals their locations. Object segmentation is defined as the pixel-level categorization of pixels, aiming to divide an image into significant regions by classifying each pixel into a specific layer. In classical object segmentation, the method of uncontrolled merging and region segmentation has been extensively investigated based on clustering, general feature optimization, or user intervention. It is divided into two primary branches based on object partitioning. In the first branch, semantic segmentation is employed, where each pixel corresponds to a semantic object classification. In the second branch, instance segmentation is utilized, providing different labels for different object instances as a further improvement of semantic segmentation [19].

In [72], the authors presented a comprehensive survey of the literature on human action recognition, with a specific focus on the fusion of vision and inertial sensing modalities. The surveyed papers were categorized based on fusion approaches, features, classifiers, and multimodality datasets. The authors also addressed challenges in real-world deployment and proposed future research directions. The work contributed a thorough overview, categorization, and insightful discussions of the fusion-based approach for human action recognition.

The authors of [118] evaluated some Kinect-based algorithms for human action recognition using multiple benchmark datasets. Their findings revealed that most methods excelled in cross-subject action recognition compared to cross-view action recognition. Additionally, skeleton-based features exhibited greater resilience in cross-view recognition, while deep learning features were well-suited for large datasets.

The authors of [119] offered a comprehensive review of recent advancements in human action recognition systems. They introduced hand-crafted representation-based methods, as well as deep learning-based approaches, for this task. A thorough analysis and a comparison of these methods and datasets used in human action recognition were presented. Furthermore, the authors suggested potential future research directions in the field.

In [120], a comprehensive review of recent progress made in semantic segmentation was presented. The authors specifically examined and compared three categories of methods: those relying on hand-engineered features, those leveraging learned features, and those utilizing weakly supervised learning. The authors presented the descriptions, as well as a comparison, of prominent datasets used in semantic segmentation. Furthermore, they conducted a series of comparisons between various semantic segmentation models to showcase their respective strengths and limitations.

In [121], a comprehensive examination of semantic segmentation techniques employing deep neural networks was presented. The authors thoroughly analyzed the leading approaches in this field, highlighting their strengths, weaknesses, and key challenges. They concluded that deep convolutional neural networks have demonstrated remarkable

effectiveness in semantic segmentation. The review encompassed an in-depth assessment of the top methods employed for semantic segmentation using deep neural networks. The strengths, weaknesses, and significant challenges associated with these approaches were carefully summarized. Semantic segmentation has played a vital role in enhancing and expanding our understanding of visual data, providing valuable insights for various computer vision applications.

The authors of [122] presented a comprehensive review of deep learning-based methods for semantic segmentation. They explored the common challenges faced in current research and highlighted emerging areas of interest in this field. Deep learning techniques have played a pivotal role in enhancing the performance of semantic segmentation tasks. Research on semantic segmentation can be categorized based on the level of supervision, namely fully supervised, weakly supervised, and semi-supervised approaches. The current research faces challenges such as limited data availability and class imbalance, which necessitate further exploration and innovation.

In [123], the latest advancements in semantic image segmentation were explored. The authors conducted a comparative analysis of different models and concluded by discussing the model that exhibited the best performance. They suggested that semantic image segmentation is a rapidly evolving field that has involved the development and application of numerous models across various domains. A performance evaluation of each semantic image segmentation model was carried out using the Intersection-over-Union (IoU) method. The results of the IoU were used to facilitate a comprehensive comparison of the different semantic image segmentation models.

Key deep learning architectures in robotic vision include CNNs, RNNs, and Generative Adversarial Networks (GANs). These innovations have wide-ranging applications in robotic vision, encompassing tasks such as object detection, pose estimation, and semantic segmentation. Convolutional neural networks (CNNs) play a central role in tasks like object detection, image classification, and scene segmentation. They excel in extracting intricate features from raw image data, enabling precise object identification and tracking. In situations demanding temporal insights, recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) networks, are essential. They excel in tracking moving objects and predicting future actions based on historical data. Some deep learning applications in robotic vision include object grasping and pick-and-place operations [124]. Offline reinforcement learning algorithms have also surfaced, facilitating continuous learning in robots without erasing previous knowledge [125]. In flower removal and pollination, a 3D perception module rooted in deep learning has emerged, elevating detection and positioning precision for robotic systems [126]. Additionally, deep learning has found utility in fastener detection within computer vision-based robotic disassembly and servicing, excelling in performance and generalization [127]. Neural networks have wielded a pivotal influence in robot vision, making strides in image segmentation, drug detection, and military applications. Deep learning methods, as demonstrated [128], stand as versatile and potent tools for augmenting robotic vision capabilities.

Semantic segmentation entails delineating a specific object or region within an image. This task finds applications in diverse industries, including filmmaking and augmented reality. In the era of deep learning, the convolutional neural network (CNN) has emerged as the main method for semantic segmentation. Rather than attempting to discern object boundaries through traditional visual cues like contrast and sharpness, deep CNNs reframe the challenge as a classification problem. By assigning a class to each pixel in the image, the network inherently identifies object boundaries. This transformation involves adapting the final layers of a conventional CNN classification network to produce H*W values, representing pixel classes, in lieu of a single value representing the entire image's class. The DeepLab series, following the FCN paradigm, spans four iterations: V1, V2, V3, and V3+, developed from 2015 to 2018. DeepLab V1 laid the foundation, while subsequent versions introduced incremental improvements. These iterations harnessed innovations from recent

image classification advancements to enhance semantic segmentation, thereby serving as a catalyst for research endeavors in the field [92,129,130].

4.2. Mathematical Foundations of Deep Learning Methods in Robotic Vision

Deep learning in robotic vision reveals a plethora of promising approaches, each with its own unique strengths and characteristics. To utilize the full potential of this technology, it is crucial to identify the most promising methods and consider several combinations to tackle specific challenges.

Convolutional neural networks (CNNs): Among the most promising approaches are CNNs, which excel in image recognition tasks. They have revolutionized object detection, image segmentation, and scene understanding in robotic vision. Their ability to automatically learn hierarchical features from raw pixel data is a game-changer.

Recurrent neural networks (RNNs): RNNs are vital for tasks requiring temporal understanding. They are used in applications like video analysis, human motion tracking, and gesture recognition. Combining CNNs and RNNs can address complex tasks by leveraging spatial and temporal information.

Reinforcement learning (RL) in robotic vision involves algorithms for robots to learn and decide via environmental interaction, utilizing a Markov Decision Process (MDP) framework. RL algorithms like Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO) use neural networks to approximate mappings between states, actions, and rewards, improving robots' understanding and navigation. The integration of RL and robotic vision is promising for applications like autonomous navigation and human–robot collaboration, which rely on well-designed reward functions.

Generative Adversarial Networks (GANs): GANs offer transformative potential in generating synthetic data and enhancing data augmentation. This is especially valuable when dealing with limited real-world data. Their combination with other models can enhance training robustness.

Transfer Learning: Using pre-trained models is a promising strategy. By fine-tuning models on robotic vision data, we can benefit from knowledge transfer and accelerate model convergence. This approach is particularly useful when data are scarce.

Multi-Modal Fusion: Combining information from various sensors, such as cameras, LiDARS, and depth sensors, is crucial for comprehensive perception. Techniques like sensor fusion, including vision and LiDAR or radar data, are increasingly promising.

The convergence of these approaches holds tremendous potential. For instance, combining CNNs, RNNs, and GANs for real-time video analysis or fusing multi-modal data with transfer learning can address complex robotic vision challenges. The promise lies in the thoughtful integration of these approaches to create holistic solutions that can empower robots to effectively perceive, understand, and interact with their environments.

The main deep learning methods are presented in Table 7.

Table 7. Deep learning methods in robotic vision.

Method	Key Features
CNNs [131]	CNNs revolutionize robotic vision by automatically learning hierarchical features from raw pixel data.
RNNs [132]	RNNs are essential for temporal tasks like video analysis and gesture recognition.
RL [133]	RL uses neural networks to approximate mappings between states, actions, and rewards, improving robots' understanding and navigation.
GANs [134]	GANs generate synthetic data and enhance data augmentation, especially beneficial for limited real-world data.
Transfer learning [135]	Transfer learning accelerates model convergence in robotic vision by leveraging pre-trained models.
Multi-modal fusion [136]	Multi-modal fusion combines information from various sensors through sensor fusion, including vision and LiDAR or radar data.

4.2.1. Convolutional Neural Networks (CNNs)

1. Convolution Operation: The convolution operation in CNNs involves the element-wise multiplication of a filter (kernel) with a portion of the input image, followed by summing the results to produce an output feature map. Mathematically, it can be represented as:

$$(I * K)(x, y) = \sum_i \sum_j I(x + i, y + j) \cdot K(i, j)$$

where I is the input image, K is the convolutional kernel, (x, y) represents the spatial position in the output feature map, and (i, j) iterates over the kernel dimensions.

2. Activation Functions: Activation functions, such as the Rectified Linear Unit (ReLU), introduce non-linearity into the network. The ReLU function is defined as:

$$f(x) = \max(0, x)$$

and is applied to the output of the convolutional and fully connected layers.

3. Pooling: Pooling layers reduce the spatial dimensions of feature maps. Max pooling, for example, retains the maximum value in a specified window. Mathematically, it can be represented as:

$$P(x, y) = \max(I(x, y))$$

where P is the pooled output and $I(x, y)$ is the input.

4. Fully Connected Layers: In the final layers of a CNN, fully connected layers perform traditional neural network operations. A fully connected layer computes the weighted sum of all inputs and passes it through an activation function, often a softmax, for classification tasks.

5. Backpropagation: The training of CNNs relies on backpropagation, a mathematical process for adjusting network weights and biases to minimize a loss function. This process involves the chain rule to compute gradients and update model parameters.

4.2.2. Recurrent Neural Networks (RNNs)

RNNs are a type of neural network designed for processing sequences of data. They have a dynamic and recurrent structure that allows them to maintain hidden states and process sequential information. The core mathematical components of RNNs include:

1. Hidden State Update: At each time step t , the hidden state h_t is updated using the current input x_t and the previous hidden state h_{t-1} through a set of weights and activation functions. Mathematically, this can be expressed as:

$$h_t = f(W_h \cdot h_{t-1} + W_x \cdot x_t + b_h)$$

where h_t is the hidden state at time step t; f is the activation function, typically the hyperbolic tangent (tanh) or sigmoid; W_h and W_x are the weight matrices; and b_h is the bias term.

2. Output Calculation: The output at each time step can be computed based on the current hidden state. For regression tasks, the output y_t is often calculated as:

$$y_t = W_y \cdot h_t + b_y$$

where y_t is the output at time step t, W_y is the weight matrix for the output, and b_y is the bias term.

3. Backpropagation Through Time (BPTT): RNNs are trained using the backpropagation through time (BPTT) algorithm, which is an extension of backpropagation. BPTT calculates gradients for each time step and updates the network's weights and biases accordingly.

RNNs are well-suited for sequence data, time-series analysis, and natural language processing tasks. They can capture dependencies and contexts in sequential information, making them a valuable tool in machine learning and deep learning.

4.2.3. Reinforcement Neural Networks (RNNs)

Reinforcement learning (RL) is a machine learning paradigm focused on training agents to make sequential decisions in an environment to maximize cumulative rewards. The fundamental mathematical components of RL include:

1. Markov Decision Process (MDP): RL problems are often formalized as MDPs. An MDP consists of a tuple (S, A, P, R), where S is the state space representing the possible environmental states; A is the action space consisting of the possible actions the agent can take; P is the transition probability function, defining the probability of transitioning from one state to another after taking a specific action; and R is the reward function, which provides a scalar reward signal to the agent for each state-action pair.

2. Policy (π): A policy defines the agent's strategy for selecting actions in different states. It can be deterministic or stochastic. Mathematically, a policy π maps states to actions: $\pi : S \rightarrow A$.

3. Value Functions: Value functions evaluate the desirability of states or state-action pairs. The most common value functions are:

- State-Value Function (V): $V^{\pi}(s)$ estimates the expected cumulative reward when starting from a state s and following policy π.
- Action-Value Function (Q): $Q^{\pi}(s, a)$ estimates the expected cumulative reward when starting from a state s, taking action a, and following policy π.

4. Bellman Equations: The Bellman equations express the relationship between the value of a state or state-action pair and the values of the possible successor states. They are crucial for updating the value functions during RL training.

5. Optimality: RL aims to find an optimal policy π^* that maximizes the expected cumulative reward. This can be achieved by maximizing the value functions:

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

$$Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a)$$

Reinforcement learning algorithms, such as Q-learning, SARSA, and various policy gradient methods, use these mathematical foundations to train agents in a wide range of applications, from game playing to robotics and autonomous systems.

4.2.4. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are a class of deep learning models that consist of two neural networks: a generator (G) and a discriminator (D). The mathematical foundations of GANs include:

1. Generator (G): The generator maps random noise z from a prior distribution ($p(z)$) to generate data samples. This process can be represented as $G(z)$.

2. Discriminator (D): The discriminator evaluates whether a given data sample is real (x) or generated by the generator ($G(z)$). It produces a scalar value representing the probability that the input is real ($D(x)$).

3. Objective Function: GANs are trained using a minimax game between G and D. The objective function to be minimized by G and maximized by D is defined as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(D(x))] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]$$

where $p_{\text{data}}(x)$ is the real data distribution, $p(z)$ is the prior distribution of noise, and $\mathbb{E}$ represents the expectation.

4. Optimal Generator: At optimality, the generator produces samples that are indistinguishable from real data, meaning $D(G(z)) = 0.5$. This occurs when the objective function $V(D, G)$ reaches its global minimum.

5. Training: GANs are trained using techniques like stochastic gradient descent. The generator updates its parameters to minimize the objective function, whereas the discriminator updates its parameters to maximize it.

6. Generated Data: The generator produces synthetic data samples $G(z)$ that closely resemble real data.

GANs are widely used in various applications, including image generation, style transfer, and data augmentation.

4.2.5. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) [137] is a type of recurrent neural network (RNN) architecture designed to overcome the vanishing gradient problem and capture long-range dependencies in sequential data. The key equations and components of LSTM include:

1. Gates: LSTMs have three gates: the forget gate (f_t), the input gate (i_t), and the output gate (o_t). These gates regulate the flow of information within a cell.

2. Cell State (C_t): LSTMs maintain a cell state, which serves as a memory unit. The cell state is updated using the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

where σ is the sigmoid activation function and $\odot$ represents the element-wise multiplication.

3. Hidden State (h_t): The hidden state is derived from the cell state and is updated using the output gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot \tanh(C_t)$$

4. Training: LSTMs are trained using the backpropagation through time (BPTT) and gradient descent algorithms. The gradients are computed with respect to the cell state, hidden state, and parameters.

LSTMs are known for their ability to capture long-term dependencies and are widely used in natural language processing, speech recognition, and various sequential data tasks.

In Table 8, we can see a detailed comparison of deep learning algorithms and methods and their integration in robotic vision.

Table 8. Comparison of deep learning algorithms and methods in robotic vision.

Neural Network	Details
Convolutional neural network (CNN) [87]	**Mathematical model:** – Utilizes convolutional layers for feature extraction. – Applies pooling layers for downsampling and spatial hierarchies. – Employs fully connected layers for classification. **Performance:** – Well suited for image-related tasks: object detection, image classification, and segmentation. – Well suited for processing static visual information.

Table 8. *Cont.*

Neural Network	Details
Recurrent neural network (RNN) [81]	**Mathematical model:** – Designed for sequential data processing and capturing temporal dependencies. – Uses hidden states and gates to model sequential information. **Performance:** – Suitable for tasks like tracking objects over time and recognizing sequential patterns in data. – Effective in understanding the dynamic nature of robotic vision. – Challenges with vanishing gradients in capturing long-range dependencies.
Reinforcement learning (RL) [133]	**Mathematical model:** – Markov Decision Process (MDP) framework. – Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO). **Performance:** – Extracts features from visual input, enhancing robots' understanding and navigation. – Utilizes algorithms, enabling robots to learn and decide through environmental interaction, and approximates mappings between states, actions, and rewards.
Generative Adversarial Network (GAN) [7]	**Mathematical model:** – Consists of a generator and discriminator engaged in a minimax game to generate and evaluate data. – Aims to generate realistic data by improving the generator over time. **Performance:** – Valuable for generating synthetic data, which can be used for training in various environments or simulating conditions in robotic vision tasks. – Valuable for image-to-image translation and style transfer.
Long Short-Term Memory (LSTM) [138]	**Mathematical model:** – A type of RNN designed to capture long-term dependencies in sequential data. – Uses gates, cell states, and hidden states to model memory. **Performance:** – Effective for tasks requiring memory of past states. – Valuable for tracking objects with complex trajectories and understanding long-term patterns.

4.3. Combining Approaches for Robotic Vision

To address the complexity of robotic vision tasks, a combination of these neural network architectures can be powerful. For instance:

- Using CNNs for initial image feature extraction to identify objects and their positions.
- Integrating RNNs to process temporal data and track object movement and trajectories over time.
- Implementing GANs to generate synthetic data for training in various environments and conditions.
- Employing LSTMs to remember past states and recognize long-term patterns in robot actions and sensor data.

By combining these approaches, robotic vision systems can leverage the strengths of each architecture to improve object detection, tracking, and the understanding of complex visual scenes in dynamic environments (see Table 9).

Table 9. Combined approaches in robotic vision.

Approach	Strategy	Benefits
CNN-RNN fusion [139]	Utilizes CNNs for initial image feature extraction and integrates RNNs to process temporal data.	– Improved object tracking by capturing both spatial and temporal features. – Enhanced understanding in dynamic scenes, combining spatial and temporal context.
GAN-based data augmentation [140]	Applies GANs to generate synthetic data, diversifying training datasets.	– Diversifying training datasets by adding synthetic data. – Enhancing robustness by training on various simulated environments.
Hybrid CNN-LSTM models [141]	Combine CNNs for static feature extraction with LSTMs for sequential understanding.	– Improved object recognition by capturing both static and sequential features. – Enhanced tracking in dynamic scenes, understanding both spatial and temporal aspects.
Triplet network with GANs [142]	Implements GANs for generating realistic variations of images and uses a triplet network (embedding CNN) to enhance similarity comparisons.	– Improved recognition through realistic image variations. – Better understanding of similar objects in varied conditions facilitated by the triplet network.

4.4. Big Data, Federated Learning, and Vision

Big data and federated learning play significant roles in advancing the field of computer vision. Big data provides a wealth of diverse visual information, which is essential for training deep learning models that power computer vision applications. These datasets enable more accurate object recognition, image segmentation, and scene understanding.

Federated learning, on the other hand, enhances privacy and efficiency. It allows multiple devices to collaboratively train models without sharing sensitive data. In computer vision, this means that the collective intelligence of various sources can be used while preserving data privacy, making it a game-changer for applications like surveillance, healthcare, and autonomous vehicles or drones.

4.4.1. Big Data

Big data uses vast and complex datasets arising from diverse origins and applications, such as social media, sensors, and cameras. Within machine vision, big data proves invaluable for pattern recognition, offering a plethora of information like images, videos, texts, and audio.

The advantages of big data are numerous: it can facilitate the creation of more accurate and resilient pattern recognition models by supplying ample samples and variations; it can display latent patterns and insights inaccessible to smaller datasets; and it can support pattern recognition tasks necessitating multiple modalities or domains. However, big data also has certain drawbacks: it can present challenges in data collection, storage, processing, analysis, and visualization; it can create ethical and legal concerns surrounding data privacy, security, ownership, and quality; and it can introduce noise, bias, or inconsistency that may impede the performance and reliability of pattern recognition models.

Big data and machine vision find a lot of applications. In athlete training, they aid behavior recognition. By combining machine vision with big data, the actions of athletes can be analyzed using cameras, providing valuable information for training and performance improvements [143].

In image classification, spatial pyramids can enhance the bag-of-words approach. Machine vision-driven big data analysis can improve speed and precision in microimage surface defect detection or be used to create intelligent guidance systems in large exhibition

halls, enhancing the visitor experience. In the context of category-level image classification, the use of spatial pyramids based on 3D scene geometry has been proposed to improve classification accuracy [144]. Data fusion techniques with redundant sensors have been used to boost robotic navigation. Big data and AI have been used to optimize communication and navigation within robotic swarms in complex environments. They have also been applied in robotic platforms for navigation and object tracking using redundant sensors and Bayesian fusion approaches [145]. Additionally, the combination of big data analysis and robotic vision has been used to develop intelligent calculation methods and devices for human health assessment and monitoring [146].

4.4.2. Federated Learning

Federated learning, a distributed machine learning technique, facilitates the collaborative training of a shared model among multiple devices or clients while preserving the confidentiality of their raw data. In the context of machine vision, federated learning proves advantageous when dealing with sensitive or dispersed data across various domains or locations. Federated learning offers several benefits: it can safeguard client data privacy and security by preserving data locally; it can minimize communication and computation costs by aggregating only model updates; and it can harness the diversity and heterogeneity of client data to enhance model generalization. Nonetheless, federated learning entails certain drawbacks: it may encounter challenges pertaining to coordination, synchronization, aggregation, and evaluation of model updates; it may be subject to communication delays or failures induced by network bandwidth limitations or connectivity issues; and it may confront obstacles in model selection, optimization, or regularization due to non-iidness or data imbalance. In computer vision and image processing, "IID" stands for "Independent and Identically Distributed". It refers to a statistical assumption about the data used in vision-related tasks.

Federated learning can be used to improve the accuracy of machine vision models. It enables training a machine learning model in a distributed manner using local data collected by client devices, without exchanging raw data among clients [147]. This approach is effective in selecting relevant data for the learning task, as only a subset of the data is likely to be relevant, whereas the rest may have a negative impact on model training. By selecting the data with high relevance, each client can use only the selected subset in the federated learning process, resulting in improved model accuracy compared to training with all data [148]. Additionally, federated learning can handle real-time data generated from the edge without consuming valuable network transmission resources, making it suitable for various real-world embedded systems [149].

LEAF is a benchmarking framework for learning in federated settings. It includes open-source federated datasets, an evaluation framework, and reference implementations. The goal of LEAF is to provide realistic benchmarks for developments in federated learning, meta-learning, and multi-task learning. It aims to capture the challenges and intricacies of practical federated environments [150].

Federated learning (FL) offers several potential benefits for machine vision applications. Firstly, FL allows multiple actors to collaborate on the development of a single machine learning model without sharing data, addressing concerns such as data privacy and security [151]. Secondly, FL enables the training of algorithms without transferring data samples across decentralized edge devices or servers, reducing the burden on edge devices and improving computational efficiency [152]. Additionally, FL can be used to train vision transformers (ViTs) through a federated knowledge distillation training algorithm called FedVKD, which reduces the edge-computing load and improves performance in vision tasks [153].

Finally, FL algorithms like FedAvg and SCAFFOLD can be enhanced using momentum, leading to improved convergence rates and performance, even with varying data heterogeneity and partial client participation [154]. The authors of [155] introduced personalized federated learning (pFL) and demonstrated its application in tailoring models for

diverse users within a decentralized system. Additionally, they introduced the Contextual Optimization (CoOp) method for fine-tuning pre-trained vision-language models.

5. Hubel and Wiesel's Electrophysiological Insights, Van Essen's Map of the Brain, and Their Impact on Robotic Vision

5.1. Hubel and Wiesel's Contribution

Deep learning's impact on robotic vision connects insights from neuroscience and computer science. Hubel and Wiesel's electrophysiological research revealed the fundamental mechanisms of human visual perception, laying the foundation for understanding how neural networks process visual information in deep learning. Similarly, Van Essen's brain map serves as a critical reference for comprehending neural pathways and functions, elucidating connections within the visual cortex for developing deep learning algorithms. The synergy between neuroscientific revelations and computer science has redefined robotic vision. Deep learning algorithms, inspired by the neural architectures discovered by Hubel and Wiesel and refined through insights from Van Essen's map, have empowered robots to decipher visual data with precision. This fusion of understanding and innovation has accelerated the development of autonomous robots capable of perceiving, interpreting, and reacting to their surroundings. By embracing the neural foundations of visual perception, deep learning has surpassed human abilities, allowing robots to navigate, interact, and make knowledgeable decisions.

Hubel and Wiesel's groundbreaking contributions in their electrophysiological studies mixed neuroscience, artificial neural networks (ANNs), and computer vision, shaping the very foundation of modern AI. Their exploration of the cat and monkey visual systems unearthed fundamental insights into sensory processing, establishing vital connections between biological mechanisms and computational paradigms. Understanding the receptive fields of cells in the cat's striate cortex shed light on brain visual processing. The authors of [156] enriched the comprehension of visual pathways from the retina to the cortex, influencing perception. Notably, moving stimuli trigger robust responses, suggesting motion's key role in cortical activation. This insight has led to advances in fields like computer vision and robotics, refining motion detection. Specific shapes, sizes, and orientations that activate cortical cells have impacted experimental design. Moreover, intricate properties within the striate cortex units hint at deeper complexities necessitating exploration. Such insights contribute to a holistic understanding of the brain's visual processing mechanisms. Studying a cat's visual cortex unveils complex receptive fields, surpassing lower visual levels. This involves receptive fields and binocular interaction and overcomes the limitations of slow-wave recording. A new approach studies individual cells using micro-electrodes, correlating responses with cell location. This method has enhanced the understanding of functional anatomy in smaller cortex areas [157–159].

Hubel and Wiesel's pioneering revelation of "feature detectors" is another cornerstone that resonates within ANNs and computer vision. These specialized neurons, responsive to distinct visual attributes, resemble the artificial neurons that define the core architecture of ANNs. Just as Hubel and Wiesel studied layers of neurons processing features like edges, ANNs harness a similar hierarchy to progressively grasp more complex patterns, enriching our understanding of both brain and machine vision. Moreover, Hubel and Wiesel's discovery of "ocular dominance columns" and "orientation columns" mirrors the hierarchical arrangement of ANNs, creating structured systems for pattern recognition. The layer-wise organization they elucidated forms the multi-layer architecture of ANNs, maximizing their capacity to decipher complex data patterns. Hubel and Wiesel's legacy also extends to computer vision, infusing it with a deeper understanding of visual processing. Their identification of critical periods in visual development aligns with the iterative "training" stages of ANNs. By synthesizing their discoveries, ANNs can autonomously learn and recognize complex patterns from images, revolutionizing fields like image classification, object detection, and facial recognition.

Many research papers have built upon the contributions of Hubel and Wiesel. Here, we examine a few of these papers. The VLSI binocular vision system simulates the primary visual cortex disparity computation in robotics and computer vision [160]. It employs silicon retinas, orientation chips, and an FPGA, enabling real-time disparity calculation with minimal hardware. Complex cell responses and a disparity map assist in depth perception and 3D reconstruction. This blend of analog and digital circuits ensures efficient computation. However, the authors solely addressed the primary visual cortex disparity emulation, overlooking other visual aspects. In [161], the authors introduced a practical vergence eye control system for binocular robot vision. The system is rooted in the primary visual cortex (V1) disparity computation and comprises silicon retinas, simple cell chips, and an FPGA. Silicon retinas mimic vertebrate retinal fields, while simple cell chips emulate orientation-selective fields like Hubel and Wiesel's model. The system generates real-time complex cell outputs for five disparities, enabling reliable vergence movement, even in complex scenarios. This development has paved the way for accurate eye control in binocular robot vision, with potential applications in robotics, computer vision, and AI. In [162], the authors introduced a hierarchical machine vision system based on the primate visual model, thereby enhancing pattern recognition in machines. It involves invariance transforms and an adaptive resonance theory network, focusing on luminance, not color, motion, or depth. The system mirrors network-level biological processes, without biochemical simulation. This system can enhance machine vision algorithms, aiding tasks like object recognition and image classification.

The authors of [163] studied visual mechanisms like automatic gain control and nonuniform processing. They suggested that these biological processes, if applied to machine vision, could reduce data and enhance computational efficiency, particularly in wide-view systems. Implementing these mechanisms could boost machine vision's processing power and effectiveness. In [164], the growth of cognitive neuroscience and the merging of psychology and neurobiology were explored. In addition, the authors examined memory, perception, action, language, and awareness, bridging behavior and brain circuits. Cognitive psychologists emphasized information flow and internal representations. The authors also touched on the molecular aspects of memory, delving into storage and neural processes, and underscored the progress in memory research within cognitive neuroscience and the value of comprehending both behavioral and molecular memory facets. The authors of [165] explored how the human visual cortex processes complex visual stimuli. They discussed the event-related potentials (ERPs) generated when viewing faces, objects, and letters. Specific ERPs revealed different stages of face processing. The study revealed distinct regions used for the recognition of objects and letters, along with bilateral and right hemisphere-specific face activity. These findings have enhanced our understanding of the neural mechanisms involved in face perception and object recognition in the human brain.

Individuals with autism exhibit challenges in recognizing faces, often due to reduced attention to eyes and unusual processing methods [166]. Impairments start early, at around 3 years old, affecting both structural encoding and recognition memory stages. Electrophysiological studies have highlighted disruptions in the face-processing neural systems from an early age that persist into adulthood. Slower face processing has been linked to more severe social issues. Autism also impacts the brain's specialization for face processing. These insights have deepened our comprehension of social cognition impairments in autism, aiding early identification and interventions. Other research papers on the use of machine learning methods for classifying autism include [167–179].

Table 10 shows the main articles discussing the above methods, while their characteristics, including their advantages, disadvantages/limitations, and complexities, are given in Appendix C.

Table 10. Methods related to Hubel and Wiesel's main methods.

Summary	References
Discuss Hubel and Wiesel's electrophysiological studies connecting neuroscience, ANNs, and computer vision.	[156–159]
Build upon Hubel and Wiesel's work, exploring VLSI binocular vision systems, practical vergence eye control systems, hierarchical machine vision systems, and visual mechanisms	[160–163]
Explores cognitive neuroscience by merging psychology and neurobiology, with a focus on memory, perception, action, language, and awareness.	[164]
Explore how the human visual cortex processes complex stimuli, revealing distinct regions for object and letter recognition and face processing. Individuals with autism face challenges in recognizing faces, with disruptions in neural systems linked to social issues.	[165,166]
Machine learning methods for classifying autism	[167–179]

5.2. Van Essen's Functional Mapping of the Brain

Van Essen's work on brain mapping serves as a bridge between our brain's complex networks and advanced deep neural networks (DNNs) in modern AI. His methodical approach to understanding how different brain regions work together during thinking and sensing is like solving a puzzle. This is similar to how DNNs, with their layers of connected artificial neurons, learn from data. Van Essen's study of the human connectome, mapping brain pathways, is similar to how DNNs are structured. Both systems process information step by step, finding patterns and understanding them better. By combining Van Essen's brain mapping with the DNN architecture, we can connect natural and artificial networks. This could lead to a better understanding of how we think and improve AI. This mix of neuroscience and AI could inspire new ways of thinking and improve what machines can do. Looking ahead, this blend could help create AI systems that work more like brains, giving us a deeper understanding of thinking and pushing AI to new levels. It is like a partnership between human creativity and machine learning, a place where what we learn about the brain can help improve technology. This mix of science and technology shows the potential of connecting our natural thinking with the digital thinking we are building.

The new map of the human cerebral cortex from Van Essen's studies has important practical implications for researchers and medical professionals. It helps researchers understand brain disorders like autism, schizophrenia, dementia, and epilepsy, potentially leading to better treatments. This map was created with precise boundaries and a well-designed algorithm, allowing researchers to more accurately compare results from different brain studies. It can also facilitate personalized brain maps for surgeries and treatments, which is especially helpful for neurosurgeons. Additionally, the map identifies specific brain areas for tasks like language processing and sensory perception, benefiting both cognitive neuroscience and interventions for people with impairments. Overall, the findings of Van Essen's study connect brain research with practical applications in medicine and neuroscience. Specifically, the authors of [180] outlined the cortical areas tied to vision and other senses and presented a database of connectivity patterns. They analyzed the cortex's hierarchy, focusing on visual and somatosensory/motor systems, and highlighted the interconnectedness and distinct processing streams. The study uncovered visual area functions and suggested that the organization allows for both divergence and convergence. The research deepened our knowledge of primate cortex hierarchy and connectivity, particularly in vision and somatosensory/motor functions.

The authors of [181] explored surface-based visualization for mapping the cerebral cortex's functional specialization. They employed an atlas to show the link between specialized regions and topographic areas in the visual cortex. The surface-based warping enhanced data mapping, thereby reducing distortions. These methods advanced high-resolution brain mapping, improving our comprehension of cerebral cortex organization and function across species, especially in humans. The authors of [182] revealed that the brain's activation–deactivation balance during tasks is naturally present, even at rest. Two

networks, linked by correlations and anticorrelations, show ongoing brain organization. This intrinsic structure showcases the brain's dynamic functionality and supports the understanding of coherent neural fluctuations' impact on brain function. The authors of [183] presented a comprehensive map of the human cerebral cortex's divisions, using magnetic resonance images and a neuroanatomical method. They identified 97 new areas, confirmed 83 previously known areas, and developed a machine learning classifier for automated identification. This tool enhanced our understanding of cortical structure and function, aiding research in diverse contexts.

Table 11 presents the main articles regarding Van Essen's mapping, while their characteristics, including their advantages, disadvantages/limitations, and complexities, are given in Appendix D.

Table 11. Methods related to Van Essen's functional mapping of the brain.

Summary	Reference
Outlines cortical areas tied to vision and other senses and presents a database of connectivity patterns.	[180]
Explores surface-based visualization for mapping the cerebral cortex's functional specialization.	[181]
Reveals the brain's activation–deactivation balance during tasks, showcasing ongoing brain organization and supporting an understanding of neural fluctuations' impact on function.	[182]
Presents a comprehensive map of the human cerebral cortex's divisions, identifies new areas, and develops a machine learning classifier for automated identification.	[183]

6. Discussion

In machine vision, there exist numerous contemporary technologies pertaining to pattern recognition, each harboring its own merits and demerits. Presented below are several recent technologies alongside their respective advantages and disadvantages.

Deep learning leverages neural networks comprising multiple layers to extract intricate and high-level features from data. Remarkable achievements have been witnessed in diverse pattern recognition tasks through deep learning, such as image classification, object detection, face recognition, and semantic segmentation, among others. Deep learning possesses certain advantages: it can autonomously learn from extensive datasets without substantial human intervention or feature engineering; it can adeptly capture non-linear and hierarchical relationships within the data; and it can reap the benefits of hardware and software advancements like GPUs and frameworks. However, deep learning also entails certain drawbacks: it demands substantial computational resources and time for training and deployment; it may be susceptible to issues of overfitting or underfitting, hinging upon network architecture selection, hyperparameter tuning, regularization techniques, and more; it may lack interpretability and explainability concerning learned features and decisions; and it may prove vulnerable to adversarial attacks or data poisoning.

Challenges and Limitations

Deep learning techniques in robotic vision offer distinct advantages, enabling high-level tasks like image recognition and segmentation, vital for robust robot vision systems. Deep learning algorithms and neural networks find diverse applications, spanning domains such as drug detection and military applications. These methods facilitate the acquisition of data-driven representations, features, and classifiers, thereby enhancing the perceptual capabilities of robotic systems. However, inherent challenges exist in employing deep learning for robotic vision. The limitations in robot hardware and software pose efficiency challenges for vision systems, and deep learning alone may not resolve all issues in industrial robotics. Furthermore, designing deep learning-based vision systems necessitates specific methodologies and tools tailored to the field's unique demands.

In the case of deep learning, several challenges demand attention. One challenge involves the design and optimization of network architectures and hyperparameters tai-

lored to distinct pattern recognition tasks and datasets. A universal or optimal solution to this quandary remains elusive, often necessitating trial-and-error or heuristic approaches. Another challenge lies in ensuring the robustness and dependability of learned models, particularly when deployed in real-world scenarios. Numerous factors can influence model performance and behavior, such as data quality, distribution shifts, and adversarial examples. Lastly, deep learning confronts the task of enhancing the interpretability and explainability of learned features and decisions, particularly when faced with intricate and high-dimensional data. Striking a balance between model accuracy and interpretability poses a challenge, as comprehending the reasoning behind model predictions or classifications is no easy feat.

Other challenges in using deep learning methods for robotic vision include the complexity and entanglement of optical parameters in wide-angle systems, which require data-driven prediction models to overcome. Another challenge is the need for robust 3D object detection, which is crucial for decision making in autonomous intelligent systems. Although deep learning has shown potential in this area, a lack of critical review and comparison among various methods makes it challenging to select the most suitable approach for specific applications. Achieving non-adversarial robustness in deep learning models is also challenging, as it is difficult to predict the types of distribution shifts that may occur. Researchers have proposed various approaches to address this challenge, but there is a need for further improvement and evaluation of model performance under data distribution shifts. Additionally, applying visual algorithms developed from computer vision datasets to robotic vision poses unique challenges due to the assumption of fixed categories and time-invariant task distributions.

Although big data holds immense potential for pattern recognition in machine vision, it possesses certain limitations that merit consideration. One limitation is the potential unavailability or inaccessibility of big data for analysis. Legal or ethical regulations may restrict access to certain data sources, such as personal or medical data. Additionally, data obtained from user-generated or crowd-sourced platforms can be unreliable or incomplete. Another limitation arises from the incompatibility or inconsistency of data derived from diverse modalities or domains. Another facet of big data's limitations lies in its varying usefulness and informativeness for pattern recognition in machine vision. Redundant or irrelevant data, such as noisy or corrupted samples, may hinder effective analysis. Moreover, biased or unrepresentative data, including imbalanced or skewed datasets, can undermine the accuracy of pattern recognition models. Furthermore, misleading or deceptive data, such as manipulated or fabricated information, can introduce additional challenges.

Federated learning presents several challenges that warrant attention. One challenge revolves around effectively coordinating and synchronizing model updates from diverse clients in a distributed and dynamic environment. Communication and computation efficiency within federated learning are influenced by various factors such as network latency, bandwidth, connectivity, and heterogeneity. Another challenge lies in striking a balance between model privacy and accuracy. Different privacy protection levels and methods, including differential privacy, secure aggregation, and encryption, exist in federated learning. However, these methods may introduce noise or distortion into model updates, potentially degrading accuracy or convergence. A third challenge pertains to addressing non-iidness and data imbalance among clients. Variations in data distributions or characteristics stemming from client preferences, behaviors, or contexts may arise. This imbalance can result in certain clients exerting greater influence or weight on the model, leading to suboptimal generalization or fairness.

7. Conclusions

Machine vision is arguably the most crucial pillar for supporting and creating functional artificial intelligence. Vision, as one of the five senses, plays a key role in proper sensory perception among humans and has significantly influenced our social and technological evolution. Considering this, the academic and research community is investing

tremendous efforts to pave new paths in machine vision development and optimize existing algorithms and methods.

In this paper, we presented a comprehensive overview of the key elements that constitute machine vision and the technologies that enhance its performance. We discussed innovative scientific methods extensively utilized in the broad field of AI in recent years, along with their advantages and limitations.

Specific attention and research focus must be directed toward understanding the aspects of how the human brain recognizes and categorizes objects. This knowledge can then be transferred to robotic models. Robotic vision, coupled with robotic touch, presents a significant challenge in robotics. Achieving a robotic hand that adapts to tasks in a manner similar to the human hand's behavior will greatly contribute to the evolution of the mechatronics scientific field and bring us closer to achieving AI with human-like characteristics. Such AI systems would be capable of successfully performing arduous, repetitive, and hazardous tasks that pose challenges for humans. Moreover, they would have the ability to tackle complex problems across various domains, ranging from astronomy to biomedicine.

It is crucial to note that careful attention should be given to the subsequent steps of technological development. Establishing an appropriate regulatory framework is necessary to ensure responsible management of these new findings and experiments by countries worldwide, thereby mitigating any potential adverse effects on humans. We are currently experiencing a period of significant change, often referred to as a new Technological and Industrial Revolution, which may rival, if not surpass, the transformative impact of the Internet. Therefore, the forthcoming steps are pivotal for human evolution, as they will shape the trajectory of our species.

In conclusion, machine vision has made remarkable progress in replicating human visual perception in computers. This survey provided a comprehensive overview of robot vision with a detailed review of papers published in the past 3–5 years Because of its interdisciplinary nature and integration with computer science, mathematics, and robotics, machine vision has become widely used in daily gadgets. The subject has advanced significantly thanks, in large part, to deep learning.

Future research directions for using deep learning methods in robotic vision include addressing challenges such as insufficient and inaccurate annotations, recognizing pathology images with different data distributions, and training AI models based on decentralized data sources. Another important area of research is the development of self-supervised learning methods and domain-adaptation techniques for medical image analysis, which can help overcome the limitations of labeled data. Additionally, there is a need for comprehensive analysis and validation of 3D object detection methods using benchmark datasets and validation matrices. Furthermore, exploring the applications of deep learning algorithms and deep nets in various areas of robot vision, such as image segmentation and drug detection, is an important research direction. Overall, the field of robotic vision is constantly evolving, and future research should focus on improving the performance and automation capabilities of deep learning-based systems

Future pathways for computer vision could include:

- Improved object detection: Overcoming challenges with small or occluded objects.
- Real-time 3D reconstruction: Creating 3D models of environments in real time.
- Automated image labeling: Automatically tagging images with descriptive and accurate labels.
- Visual reasoning and understanding: Developing algorithms that can reason and make decisions based on visual input.
- Robustness to adversarial attacks: Creating computer vision models that are robust to adversarial attacks, suitable for security applications, and capable of preventing image manipulation.
- Integration with other technologies: Finding new ways to integrate computer vision with other technologies, such as robotics, virtual reality, and augmented reality.

- Improved facial recognition: Developing more accurate and reliable methods for facial recognition that can be used in security and identification applications.
- More efficient deep learning models: Developing deep learning models that require less computation and can run faster on mobile devices.
- Enhanced video analysis: Improving the ability of computer vision to analyze video data, including object tracking and activity recognition.
- Expanding applications: Finding new and innovative ways to apply computer vision technology in fields such as healthcare, agriculture, and transportation.
- Detection of hidden/camouflaged objects, with applications in surveillance and biology.
- A challenging task will be the detection of objects that are intentionally designed to blend into their environments, like camouflaged ones [184]. This will be especially interesting in monitoring natural environments but also has potential in military applications.
- Depth perception and 3D object detection are also very interesting, as they have applications in depth perception, navigation, action recognition, and more [185,186]. This topic was also identified as a future challenge in [187].
- Finally, emergency rescue missions would be a highly impactful application to consider [188].

Author Contributions: Conceptualization, N.M. and G.F.F.; methodology, N.M.; validation, L.M. and G.S.M.; writing—original draft preparation, N.M.; writing—review and editing, L.M., G.S.M. and G.F.F.; supervision, G.F.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No data available.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Characteristics of methods presented in Table 5.

Method	Performance and Advantages	Disadvantages and Limitations	Complexities
Deep learning (CNNs and RNNs) [9,61–72]	High accuracy in human action recognition. Comprehensive solution with spatial and temporal modeling. End-to-end learning enhances accuracy. Transfer learning improves performance. Robustness to variations.	Requires large labeled datasets. Computational complexity. "Black-box" nature. Prone to overfitting. Difficulty in capturing long-term dependencies.	Designing effective architectures. Hyperparameter tuning. Integration of modalities. Real-time processing. Adapting to dynamic environments.
Attention-based LSTM [73]	Selective feature attention. Enhanced discriminative power. Beneficial for sequential data. Multi-stream integration. Robust feature extraction.	Computational complexity. Limited generalization. Reduced interpretability.	Designing effective architectures. Hyperparameter tuning. Addressing convergence challenges. Integrating saliency information. Handling temporal misalignments.

Table A1. *Cont.*

Method	Performance and Advantages	Disadvantages and Limitations	Complexities
Hybrid deep learning model [74]	Combined strengths for improved action recognition. Enhanced accuracy and generalization. Effective in capturing spatial and temporal features. Potential for end-to-end learning.	Complexity in architecture design. Increased computational requirements. Potential interpretability challenges.	Balancing architecture complexity. Hyperparameter tuning. Handling data variability.
Utilization of multiple models [75]	Captures global and local motion features. Improved accuracy through diversified learning. Potential for handling complex actions.	Increased computational demands. Model integration challenges. Complexity in handling diverse motion patterns.	Coordinating multiple models. Addressing computational efficiency. Handling variations in action patterns.
Utilization of RGB frames, Bi-LSTM, and CNN [76]	Leverages RGB frames for visual information. Bi-LSTM captures temporal dependencies. CNN extracts spatial features. Comprehensive approach for accurate action recognition.	Requires labeled RGB frame data. Challenges in handling long-term dependencies. Computational complexity with multiple components.	Integrating RGB frames, Bi-LSTM, and CNN. Optimal parameter tuning for each component. Addressing computational demands.
Novel hybrid architecture [77]	Combines knowledge from four pre-trained network models. Leverages pre-trained models for feature extraction. Potential for improved accuracy through model fusion.	Dependence on availability of pre-trained models. Challenges in model fusion and coordination. Interpretability may be compromised.	Integrating outputs from multiple pre-trained models. Handling domain shifts between pre-trained models. Addressing interpretability challenges.
Temporal-spatial mapping operation [78]	Utilizes temporal-spatial mapping for capturing intricate action patterns. Effective in handling spatial and temporal dynamics. Potential for improved accuracy in recognizing complex actions.	Computational demands for mapping operations. Models interpretability challenges. Dependence on effective mapping algorithms.	Designing and optimizing temporal-spatial mapping operations. Addressing computational complexity through efficient algorithms. Ensuring model interpretability through transparent mapping.
Image-based HAR through transfer learning [79]	Applies transfer learning for leveraging pre-existing image-based models. Reduces the need for extensive labeled data. Potential for improved accuracy through knowledge transfer.	Limited by the quality and diversity of pre-existing models. Challenges in transferring knowledge across different domains. May require additional adaptation for specific tasks.	Ensuring effective transfer of knowledge. Addressing domain-specific challenges in transfer learning. Fine-tuning for optimal performance in target tasks.
Cooperative approach for feature selection [80]	Cooperative approach for effective feature selection. Utilizes collaboration for identifying relevant features. Potential for enhanced accuracy through feature synergy.	Coordination challenges in cooperative feature selection. Dependence on effective collaboration mechanisms. May require additional computational resources.	Designing and implementing effective cooperative feature selection mechanisms. Addressing coordination challenges in a cooperative approach. Ensuring scalability and efficiency in the feature selection process.

Appendix B

Table A2. Characteristics of methods presented in Table 6.

Method	Performance and Advantages	Disadvantages/Limitations	Complexities
Identification of fundamental computer vision problems [12,81–85]	Establish a foundation for computer vision tasks. Address image classification, object detection, and segmentation.	Assume a clear understanding of the identified problems. May oversimplify the challenges faced by each problem.	Defining problem-specific criteria for identification. Ensuring a comprehensive understanding of each identified problem.
Semantic segmentation with CNNs [86–92]	Deep learning, particularly CNNs, significantly enhances semantic segmentation. Assign labels to every pixel in an image.	Require large labeled datasets for effective training. Computational complexity, especially with deep networks.	Leveraging pre-trained CNNs for transfer learning. Addressing challenges in dataset acquisition and annotation.
Semantic segmentation architectures [50,63,86,87,90–92,97–107]	Describe components of semantic segmentation architectures. Introduce three main approaches: region-based, FCN-based, and weakly supervised.	Variability in performance across different architectures. Difficulty in choosing the most suitable approach for a given task.	Understanding the components and trade-offs of each architecture. Task-specific evaluation to select the most appropriate approach.
Semantic segmentation in medical image analysis with DCNNs [108–110]	Focus on semantic segmentation in medical image analysis. Highlight the use of deep convolutional neural networks (DCNNs) in this context.	Limited availability of labeled medical imaging datasets. Ethical considerations and privacy concerns in medical data usage.	Developing strategies for obtaining or generating labeled medical datasets. Adhering to ethical guidelines and regulations in medical image analysis.

Appendix C

Table A3. Characteristics of methods presented in Table 10.

Method	Performance and Advantages	Disadvantages and Limitations	Complexities
Hubel and Wiesel's electrophysiological studies [156–159]	Connect neuroscience, artificial neural networks (ANNs), and computer vision.	Limited to the understanding provided by electrophysiological studies.	Integrating findings from neuroscience into ANNs and computer vision.
Works building upon Hubel and Wiesel's work [160–163]	Explore VLSI binocular vision systems, practical vergence eye control systems, hierarchical machine vision systems, and visual mechanisms.	Require expertise in various domains (VLSI design, eye control, machine vision). Practical implementation challenges in building complex systems.	Collaborating across multidisciplinary fields. Overcoming technical challenges in system design and integration.
Exploring cognitive neuroscience [164]	Merges psychology and neurobiology, with a focus on memory, perception, action, language, and awareness.	Complexity in studying and understanding cognitive processes. Interdisciplinary nature may result in varied perspectives.	Developing models that bridge psychological and neurobiological concepts. Ensuring a holistic understanding of cognitive processes.
Human visual cortex processing [165,166]	Explore how the human visual cortex processes complex stimuli. Reveal distinct regions for object recognition, letters, and face processing. Investigate challenges faced by individuals with autism in recognizing faces.	Limited to observational and correlational findings. Ethical considerations in studying neurological conditions.	Developing interventions based on understanding neural processing. Adhering to ethical guidelines in neuroscience research.
Machine learning methods for classifying autism [167–179]	Apply machine learning for classifying autism based on identified neural disruptions.	Rely on the availability of relevant and diverse datasets. Challenges in generalization to diverse populations.	Ensuring representative and unbiased datasets for training. Addressing the complexity of individual variations in autism.

Appendix D

Table A4. Characteristics of methods presented in Table 11.

Method	Performance and Advantages	Disadvantages and Limitations	Complexities
Cortical areas and connectivity patterns [180]	Outlines cortical areas tied to vision and other senses. Presents a database for connectivity patterns.	Limited to observational and correlational findings. The database may not capture dynamic changes over time.	Enhancing the database to incorporate temporal connectivity patterns. Ensuring accurate mapping of cortical areas for different sensory functions.
Surface-based visualization for functional specialization [181]	Explores surface-based visualization for mapping the cerebral cortex's functional specialization.	Interpretation challenges in surface-based visualization. Limited to the visible cortex surface, potentially missing deeper structures.	Developing advanced visualization techniques for deeper structures. Validating functional specialization findings through complementary methods.
Brain activation–deactivation balance [182]	Reveals the brain's activation–deactivation balance during tasks. Showcases ongoing brain organization and the impact of neural fluctuations on function.	Challenges in precisely quantifying activation–deactivation balance. Limited to the understanding provided by observational studies.	Developing quantitative measures for activation–deactivation balance. Integrating findings with computational models to understand neural dynamics.
Comprehensive map of human cerebral cortex [183]	Presents a comprehensive map of the human cerebral cortex's divisions. Identifies new areas and develops a machine learning classifier for automated identification.	Limited by the quality and diversity of available datasets. Challenges in interpreting the functional significance of newly identified areas.	Ensuring diversity and representativeness in dataset collection. Collaborating with neuroscientists to validate functional roles of newly identified areas.

References

1. Bayoudh, K.; Knani, R.; Hamdaoui, F.; Mtibaa, A. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.* **2021**, *38* , 2939–2970. [CrossRef] [PubMed]
2. Robinson, N.; Tidd, B.; Campbell, D.; Kulić, D.; Corke, P. Robotic Vision for Human-Robot Interaction and Collaboration: A Survey and Systematic Review. *ACM Trans. Hum.-Robot. Interact.* **2023**, *12*, 1–66. [CrossRef]
3. Anthony, E.J.; Kusnadi, R.A. Computer Vision for Supporting Visually Impaired People: A Systematic Review. *Eng. Math. Comput. Sci. (Emacs) J.* **2021**, *3*, 65–71. [CrossRef]
4. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
5. Gupta, A.; Anpalagan, A.; Guan, L.; Khwaja, A.S. Deep Learning for Object Detection and Scene Perception in Self-Driving Cars: Survey, Challenges, and Open Issues. *Array* **2021**, *10*, 100057. [CrossRef]
6. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Sánchez, C.I. A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
7. Huang, H.; Yu, P.S.; Wang, C. An Introduction to Image Synthesis with Generative Adversarial Nets. *arXiv* **2018**, arXiv:1803.04469.
8. Ibarz, J.; Tan, J.; Finn, C.; Kalakrishnan, M.; Pastor, P.; Levine, S. How to Train Your Robot with Deep Reinforcement Learning: Lessons We Have Learned. *Int. J. Robot. Res.* **2021**, *40*, 698–721. [CrossRef]
9. Ganesh, D.; Teja, R.R.; Reddy, C.D.; Swathi, D. Human Action Recognition based on Depth maps, Skeleton and Sensor Images using Deep Learning. In Proceedings of the 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 7–9 October 2022. [CrossRef]
10. Wu, D.; Sharma, N.; Blumenstein, M. Recent advances in video-based human action recognition using deep learning: A review. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 2865–2872.
11. Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; Liu, Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *Acm Comput. Surv. (Csur)* **2021**, *54*, 1–40. [CrossRef]
12. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *arXiv* **2017**, arXiv:1703.06870.
13. Jin, L.; Zhu, Z.; Song, E.; Xu, X. An Effective Vector Filter for Impulse Noise Reduction Based on Adaptive Quaternion Color Distance Mechanism. *Signal Process.* **2019**, *155*, 334–345. [CrossRef]
14. Chernov, V.; Alander, J.; Bochko, V. Integer-Based Accurate Conversion between RGB and HSV Color Spaces. *Comput. Electr. Eng.* **2015**, *46*, 328–337. [CrossRef]

15. Tsapatsoulis, N. Digital Image Processing Lecture Notes. 2023. Available online: https://www.studocu.com/in/document/jawaharlal-nehru-technological-university-hyderabad/ece/digital-image-processing-lecture-notes-2022-2023/56139343 (accessed on 10 March 2023).
16. Arunpandian, M.; Arunprasath, T.; Vishnuvarthanan, G.; Rajasekaran, M.P. Thresholding Based Soil Feature Extraction from Digital Image Samples–A Vision Towards Smarter Agrology. In *Information and Communication Technology for Intelligent Systems (ICTIS 2017)-Volume 1*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 458–465.
17. Kolkur, S.; Kalbande, D.; Shimpi, P.; Bapat, C.; Jatakia, J. Human Skin Detection Using RGB, HSV and YCbCr Color Models. *arXiv* **2017**, arXiv:1708.02694.
18. BlackIce. HSI Color Conversion—Imaging Toolkit Feature. 2023. Available online: https://www.blackice.com/colorspaceHSI.htm (accessed on 10 March 2023).
19. Feng, X.; Jiang, Y.; Yang, X.; Du, M.; Li, X. Computer vision algorithms and hardware implementations: A survey. *Integration* **2019**, *69*, 309–320. [CrossRef]
20. Boykov, Y.; Funka-Lea, G. Graph Cuts and Efficient ND Image Segmentation. *Int. J. Comput. Vis.* **2006**, *70*, 109–131. [CrossRef]
21. Pizurica, A.; Philips, W.; Lemahieu, I.; Acheroy, M. A Joint Inter-and Intrascale Statistical Model for Bayesian Wavelet Based Image Denoising. *IEEE Trans. Image Process.* **2002**, *11*, 545–557. [CrossRef] [PubMed]
22. Shi, X.; Li, Y.; Zhao, Q. Flexible Hierarchical Gaussian Mixture Model for High-Resolution Remote Sensing Image Segmentation. *Remote Sens.* **2020**, *12*, 1219. [CrossRef]
23. Wang, X.F.; Huang, D.S.; Xu, H. An Efficient Local Chan–Vese Model for Image Segmentation. *Pattern Recognit.* **2010**, *43*, 603–618. [CrossRef]
24. Bresson, X.; Esedoğlu, S.; Vandergheynst, P.; Thiran, J.P.; Osher, S. Fast Global Minimization of the Active Contour/Snake Model. *J. Math. Imaging Vis.* **2007**, *28*, 151–167. [CrossRef]
25. Aytaç, E. Unsupervised Learning Approach in Defining the Similarity of Catchments: Hydrological Response Unit Based k-Means Clustering, a Demonstration on Western Black Sea Region of Turkey. *Int. Soil Water Conserv. Res.* **2020**, *8*, 321–331. [CrossRef]
26. Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.K.; Ren, F. Learning in the Frequency Domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1740–1749.
27. Dubes, R.C.; Jain, A.K.; Nadabar, S.G.; Chen, C.C. MRF Model-Based Algorithms for Image Segmentation. In Proceedings of the 10th International Conference on Pattern Recognition, Atlantic City, NJ, USA, 16–21 June 1990; Volume 1, pp. 808–814.
28. Bleau, A.; Leon, L.J. Watershed-Based Segmentation and Region Merging. *Comput. Vis. Image Underst.* **2000**, *77*, 317–370. [CrossRef]
29. Wu, Z.; Gao, Y.; Li, L.; Xue, J.; Li, Y. Semantic Segmentation of High-Resolution Remote Sensing Images Using Fully Convolutional Network with Adaptive Threshold. *Connect. Sci.* **2019**, *31*, 169–184. [CrossRef]
30. Gout, C.; Le Guyader, C.; Vese, L. Segmentation under Geometrical Conditions Using Geodesic Active Contours and Interpolation Using Level Set Methods. *Numer. Algorithms* **2005**, *39*, 155–173. [CrossRef]
31. Das, P.; Das, A. A Fast and Automated Segmentation Method for Detection of Masses Using Folded Kernel Based Fuzzy C-Means Clustering Algorithm. *Appl. Soft Comput.* **2019**, *85*, 105775. [CrossRef]
32. Ziou, D.; Tabbone, S. Edge Detection Techniques-an Overview. *Pattern Recognit. Image Anal. C/C Raspoznavaniye Obraz. Anal. Izobr.* **1998**, *8*, 537–559.
33. Kurak, C.W., Jr.; McHugh, J. A Cautionary Note on Image Downgrading. In Proceedings of the Annual Computer Security Applications Conference, San Antonio, TX, USA, 30 November–4 December 1992; pp. 153–159.
34. Hussin, R.; Juhari, M.R.; Kang, N.W.; Ismail, R.C.; Kamarudin, A. Digital Image Processing Techniques for Object Detection from Complex Background Image. *Procedia Eng.* **2012**, *41*, 340–344. [CrossRef]
35. Cruz, D.J.; Amaral, R.L.; Santos, A.D.; Tavares, J.M.R. Application of Digital Image Processing Techniques to Detect Through-Thickness Crack in Hole Expansion Test. *Metals* **2023**, *13*, 1197. [CrossRef]
36. Wang, R.; Lei, T.; Cui, R.; Zhang, B.; Meng, H.; Nandi, A.K. Medical Image Segmentation Using Deep Learning: A Survey. *Iet Image Process.* **2022**, *16*, 1243–1267. [CrossRef]
37. Yadav, S.S.; Jadhav, S.M. Deep Convolutional Neural Network Based Medical Image Classification for Disease Diagnosis. *J. Big Data* **2019**, *6*, 113. [CrossRef]
38. Giuliani, D. Metaheuristic Algorithms Applied to Color Image Segmentation on HSV Space. *J. Imaging* **2022**, *8*, 6. [CrossRef]
39. Mallick, S. Image Recognition and Object Detection: Part 1; Learn OpenCV. 2016. Available online: https://learnopencv.com/image-recognition-and-object-detection-part1/ (accessed on 9 March 2023).
40. Xylourgos, N. Segmentation of Ultrasound Images for Finding Anatomical References. Bachelor's Thesis, Technological Educational Institute of Crete, Heraklion, Greece, 2009.
41. Nixon, M.; Aguado, A. *Feature Extraction and Image Processing for Computer Vision*; Academic Press: Cambridge, MA, USA, 2019.
42. Wang, P.; Fan, E.; Wang, P. Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning. *Pattern Recognit. Lett.* **2021**, *141*, 61–67. [CrossRef]
43. Uddin, M.P.; Mamun, M.A.; Hossain, M.A. PCA-based Feature Reduction for Hyperspectral Remote Sensing Image Classification. *Iete Tech. Rev.* **2021**, *38*, 377–396. [CrossRef]
44. Wan, H.; Wang, H.; Scotney, B.; Liu, J. A Novel Gaussian Mixture Model for Classification. In Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 3298–3303.

45. Haji, S.H.; Abdulazeez, A.M. Comparison of Optimization Techniques Based on Gradient Descent Algorithm: A Review. *Palarch's J. Archaeol. Egypt/Egyptol.* **2021**, *18*, 2715–2743.
46. Chandra, M.A.; Bedi, S.S. Survey on SVM and Their Application in Image Classification. *Int. J. Inf. Technol.* **2021**, *13*, 1–11. [CrossRef]
47. Frenkel, C.; Lefebvre, M.; Bol, D. Learning without Feedback: Fixed Random Learning Signals Allow for Feedforward Training of Deep Neural Networks. *Front. Neurosci.* **2021**, *15*, 629892. [CrossRef] [PubMed]
48. Zhao, X.; Huang, P.; Shu, X. Wavelet-Attention CNN for Image Classification. *Multimed. Syst.* **2022**, *28*, 915–924. [CrossRef]
49. Venkataramanan, A.; Benbihi, A.; Laviale, M.; Pradalier, C. Gaussian Latent Representations for Uncertainty Estimation Using Mahalanobis Distance in Deep Classifiers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 4488–4497.
50. Tong, K.; Wu, Y.; Zhou, F. Recent Advances in Small Object Detection Based on Deep Learning: A Review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
51. Barazida, N. YOLOv6: Next Generation Object Detection—Review and Comparison. 2022. Available online: https://www.linkedin.com/posts/dagshub_yolov6-next-generation-object-detection-activity-6947577684583456768-06KJ?trk=public_profile_like_view (accessed on 9 March 2023).
52. Zoph, B.; Cubuk, E.D.; Ghiasi, G.; Lin, T.Y.; Shlens, J.; Le, Q.V. Learning Data Augmentation Strategies for Object Detection. In *Computer Vision–ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12372, pp. 566–583. [CrossRef]
53. Masita, K.L.; Hasan, A.N.; Shongwe, T. Deep Learning in Object Detection: A Review. In Proceedings of the 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 6–7 August 2020; pp. 1–11.
54. Shepley, A.J.; Falzon, G.; Kwan, P.; Brankovic, L. Confluence: A Robust Non-IoU Alternative to Non-Maxima Suppression in Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 11561–11574 . [CrossRef]
55. Zhang, L.; Zhang, Y.; Zhang, Z.; Shen, J.; Wang, H. Real-Time Water Surface Object Detection Based on Improved Faster R-CNN. *Sensors* **2019**, *19*, 3523. [CrossRef]
56. Zhong, F.; Quan, C. Stereo-Rectification and Homography-Transform-Based Stereo Matching Methods for Stereo Digital Image Correlation. *Measurement* **2021**, *173*, 108635. [CrossRef]
57. Jin, S.; Liu, W.; Xie, E.; Wang, W.; Qian, C.; Ouyang, W.; Luo, P. Differentiable Hierarchical Graph Grouping for Multi-person Pose Estimation. In *Computer Vision–ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12352, pp. 718–734. [CrossRef]
58. Min, K.; Lee, G.H.; Lee, S.W. Attentional Feature Pyramid Network for Small Object Detection. *Neural Netw.* **2022**, *155*, 439–450. [CrossRef] [PubMed]
59. Ciaparrone, G.; Sánchez, F.L.; Tabik, S.; Troiano, L.; Tagliaferri, R.; Herrera, F. Deep Learning in Video Multi-Object Tracking: A Survey. *Neurocomputing* **2020**, *381*, 61–88. [CrossRef]
60. Mehul. Object Tracking in Videos: Introduction and Common Techniques. 2020. Available online: https://aidetic.in/blog/2020/10/05/object-tracking-in-videos-introduction-and-common-techniques/ (accessed on 11 March 2023).
61. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3D joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012. [CrossRef]
62. Fathi, A.; Mori, G. Action recognition by learning mid-level motion features. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008. [CrossRef]
63. Huang, C.P.; Hsieh, C.H.; Lai, K.T.; Huang, W.Y. Human Action Recognition Using Histogram of Oriented Gradient of Motion History Image. In Proceedings of the 2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control, Beijing, China, 21–23 October 2011. [CrossRef]
64. Chun, S.; Lee, C. Human action recognition using histogram of motion intensity and direction from multiple views. *Iet Comput. Vis.* **2016**, *10*, 250–257. [CrossRef]
65. Hassan, M.; Ahmad, T.; Farooq, A.; Ali, S.A.; hassan, S.R.; Liaqat, N. A Review on Human Actions Recognition Using Vision Based Techniques. *J. Image Graph.* **2014**, *2*, 28–32. [CrossRef]
66. Al-Ali, S.; Milanova, M.; Al-Rizzo, H.; Fox, V.L. Human Action Recognition: Contour-Based and Silhouette-Based Approaches. In *Computer Vision in Control Systems-2*; Springer International Publishing: Cham, Switzerland, 2014; pp. 11–47. [CrossRef]
67. Chang, M.J.; Hsieh, J.T.; Fang, C.Y.; Chen, S.W. A Vision-based Human Action Recognition System for Moving Cameras Through Deep Learning. In Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning, New York, NY, USA, 27–29 November 2019. [CrossRef]
68. Chiang, M.L.; Feng, J.K.; Zeng, W.L.; Fang, C.Y.; Chen, S.W. A Vision-Based Human Action Recognition System for Companion Robots and Human Interaction. In Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 7–10 December 2018. [CrossRef]
69. Hoshino, S.; Niimura, K. Robot Vision System for Real-Time Human Detection and Action Recognition. In *Intelligent Autonomous Systems 15*; Springer International Publishing: Cham, Switzerland, 2018; pp. 507–519. [CrossRef]

70. Hoshino, S.; Niimura, K. Robot Vision System for Human Detection and Action Recognition. *J. Adv. Comput. Intell. Intell. Inform.* **2020**, *24*, 346–356. [CrossRef]
71. Chen, Q.; Tang, H.; Cai, J. Human Action Recognition Based on Vision Transformer and L2 Regularization. In Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition, New York, NY, USA, 17–19 November 2022. [CrossRef]
72. Majumder, S.; Kehtarnavaz, N. Vision and Inertial Sensing Fusion for Human Action Recognition: A Review. *IEEE Sens. J.* **2021**, *21*, 2454–2467. [CrossRef]
73. Dai, C.; Liu, X.; Lai, J. Human Action Recognition Using Two-Stream Attention Based LSTM Networks. *Appl. Soft Comput.* **2020**, *86*, 105820. [CrossRef]
74. Jaouedi, N.; Boujnah, N.; Bouhlel, M.S. A New Hybrid Deep Learning Model for Human Action Recognition. *J. King Saud-Univ.-Comput. Inf. Sci.* **2020**, *32*, 447–453. [CrossRef]
75. Gu, Y.; Ye, X.; Sheng, W.; Ou, Y.; Li, Y. Multiple Stream Deep Learning Model for Human Action Recognition. *Image Vis. Comput.* **2020**, *93*, 103818. [CrossRef]
76. Singh, T.; Vishwakarma, D.K. A Deeply Coupled ConvNet for Human Activity Recognition Using Dynamic and RGB Images. *Neural Comput. Appl.* **2021**, *33*, 469–485. [CrossRef]
77. Yilmaz, A.A.; Guzel, M.S.; Bostanci, E.; Askerzade, I. A Novel Action Recognition Framework Based on Deep-Learning and Genetic Algorithms. *IEEE Access* **2020**, *8*, 100631–100644. [CrossRef]
78. Song, X.; Lan, C.; Zeng, W.; Xing, J.; Sun, X.; Yang, J. Temporal–Spatial Mapping for Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 748–759. [CrossRef]
79. Chakraborty, S.; Mondal, R.; Singh, P.K.; Sarkar, R.; Bhattacharjee, D. Transfer Learning with Fine Tuning for Human Action Recognition from Still Images. *Multimed. Tools Appl.* **2021**, *80*, 20547–20578. [CrossRef]
80. Guha, R.; Khan, A.H.; Singh, P.K.; Sarkar, R.; Bhattacharjee, D. CGA: A New Feature Selection Model for Visual Human Action Recognition. *Neural Comput. Appl.* **2021**, *33*, 5267–5286. [CrossRef]
81. Forch, V.; Hamker, F.H. Recurrent Spatial Attention for Facial Emotion Recognition. In Proceedings of the Workshop Localize IT, Chemnitz Linux-Tage, Chemnitz, Germany, 16–17 March 2019.
82. Schröder, E.; Braun, S.; Mählisch, M.; Vitay, J.; Hamker, F. Feature Map Transformation for Multi-Sensor Fusion in Object Detection Networks for Autonomous Driving. In Proceedings of the Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Las Vegas, NV, USA, 25–26 April 2019; pp. 118–131.
83. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
84. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
85. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
86. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014. [CrossRef]
87. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158. [CrossRef] [PubMed]
88. Lundgren, A.V.A.; dos Santos, M.A.O.; Bezerra, B.L.D.; Bastos-Filho, C.J.A. Systematic Review of Computer Vision Semantic Analysis in Socially Assistive Robotics. *AI* **2022**, *3*, 229–249. [CrossRef]
89. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the State-of-the-Art Technologies of Semantic Segmentation Based on Deep Learning. *Neurocomputing* **2022**, *493*, 626–646. [CrossRef]
90. Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; Huang, T.S. Revisiting Dilated Convolution: A Simple Approach for Weakly-and Semi-Supervised Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7268–7277.
91. Zhang, M.; Zhou, Y.; Zhao, J.; Man, Y.; Liu, B.; Yao, R. A Survey of Semi-and Weakly Supervised Semantic Segmentation of Images. *Artif. Intell. Rev.* **2020**, *53*, 4259–4288. [CrossRef]
92. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
93. Arbelaez, P.; Hariharan, B.; Gu, C.; Gupta, S.; Bourdev, L.; Malik, J. Semantic Segmentation Using Regions and Parts. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012. [CrossRef]
94. Tighe, J.; Lazebnik, S. Finding Things: Image Parsing with Regions and Per-Exemplar Detectors. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013. [CrossRef]
95. He, Y.; Chiu, W.C.; Keuper, M.; Fritz, M. STD2P: RGBD Semantic Segmentation Using Spatio-Temporal Data-Driven Pooling. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]

96. Adamek, T.; O'Connor, N.E.; Murphy, N. Region-Based Segmentation of Images Using Syntactic Visual Features. In Proceedings of the WIAMIS 2005—6th International Workshop on Image Analysis for Multimedia Interactive Services, Montreux, Switzerland, 13–15 April 2005 .
97. Ji, J.; Lu, X.; Luo, M.; Yin, M.; Miao, Q.; Liu, X. Parallel Fully Convolutional Network for Semantic Segmentation. *IEEE Access Pract. Innov. Open Solut.* **2021**, *9*, 673–682. [CrossRef]
98. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [CrossRef]
99. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015. [CrossRef]
100. Dai, J.; He, K.; Sun, J. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015. [CrossRef]
101. Chen, L.; Wu, W.; Fu, C.; Han, X.; Zhang, Y. Weakly Supervised Semantic Segmentation with Boundary Exploration. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 347–362.
102. Kwak, S.; Hong, S.; Han, B. Weakly Supervised Semantic Segmentation Using Superpixel Pooling Network. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
103. Ouassit, Y.; Ardchir, S.; Yassine El Ghoumari, M.; Azouazi, M. A Brief Survey on Weakly Supervised Semantic Segmentation. *Int. J. Online Biomed. Eng.* **2022**, *18*, 83–113. [CrossRef]
104. Schmitt, M.; Prexl, J.; Ebel, P.; Liebel, L.; Zhu, X.X. Weakly Supervised Semantic Segmentation of Satellite Images for Land Cover Mapping–Challenges and Opportunities. *arXiv* **2020**, arXiv:2002.08254.
105. Gama, P.H.T.; Oliveira, H.; dos Santos, J.A.; Cesar, R.M., Jr. An overview on Meta-learning approaches for Few-shot Weakly-supervised Segmentation. *Comput. Graph.* **2023**, *113*, 77–88. [CrossRef]
106. Wang, J.; Ma, Y.; Zhang, L.; Gao, R.X.; Wu, D. Deep learning for smart manufacturing: Methods and applications. *J. Manuf. Syst.* **2018**, *48*, 144–156. [CrossRef]
107. Zhang, D.; Han, J.; Cheng, G.; Yang, M.H. Weakly Supervised Object Localization and Detection: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5866–5885. [CrossRef] [PubMed]
108. Azzi, Y.; Moussaoui, A.; Kechadi, M.T. Semantic Segmentation of Medical Images with Deep Learning: Overview. *Med. Technol. J.* **2020**, *4*, 568–575. [CrossRef]
109. Singh, R.; Rani, R. Semantic Segmentation using Deep Convolutional Neural Network: A Review. *Ssrn Electron. J.* **2020**, *1*, 1–8. [CrossRef]
110. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Semantic Segmentation With Context Encoding and Multi-Path Decoding. *IEEE Trans. Image Process.* **2020**, *29*, 3520–3533. [CrossRef]
111. Miled, M.; Messaoud, M.A.B.; Bouzid, A. Lip reading of words with lip segmentation and deep learning. *Multimed. Tools Appl.* **2023**, *82*, 551–571. [CrossRef]
112. Gianey, H.K.; Khandelwal, P.; Goel, P.; Maheshwari, R.; Galhotra, B.; Singh, D.P. Lip Reading Framework using Deep Learning and Machine Learning. In *Advances in Data Science and Analytics: Concepts and Paradigms*; Scrivener Publishing LLC: Beverly, MA, USA, 2023; pp. 67–87.
113. Wu, Y.; Wang, D.H.; Lu, X.T.; Yang, F.; Yao, M.; Dong, W.S.; Shi, J.B.; Li, G.Q. Efficient Visual Recognition: A Survey on Recent Advances and Brain-Inspired Methodologies. *Mach. Intell. Res.* **2022**, *19*, 366–411. [CrossRef]
114. Santosh, K.; Hegadi, R. Recent Trends in Image Processing and Pattern Recognition. In Proceedings of the Second International Conference, RTIP2R 2018, Solapur, India, 21–22 December 2018. Revised Selected Papers, Part I; Communications in Computer and Information Science ; Springer-Nature: Singapore, 2019.
115. Liu, H.; Yin, J.; Luo, X.; Zhang, S. Foreword to the Special Issue on Recent Advances on Pattern Recognition and Artificial Intelligence. *Neural Comput. Appl.* **2018**, *29*, 1–2. [CrossRef]
116. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features from Tiny Images*; Technical Report; University of Toronto: Toronto, ON, Canada, 2009.
117. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
118. Wang, L.; Huynh, D.Q.; Koniusz, P. A Comparative Review of Recent Kinect-Based Action Recognition Algorithms. *IEEE Trans. Image Process.* **2019**, *29*, 15–28. [CrossRef] [PubMed]
119. Al-Faris, M.; Chiverton, J.; Ndzi, D.; Ahmed, A.I. A Review on Computer Vision-Based Methods for Human Action Recognition. *J. Imaging* **2020**, *6*, 46. [CrossRef]
120. Yu, H.; Yang, Z.; Tan, L.; Wang, Y.; Sun, W.; Sun, M.; Tang, Y. Methods and datasets on semantic segmentation: A review. *Neurocomputing* **2018**, *304*, 82–103. [CrossRef]
121. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A review of semantic segmentation using deep neural networks. *Int. J. Multimed. Inf. Retr.* **2017**, *7*, 87–93. [CrossRef]
122. Hao, S.; Zhou, Y.; Guo, Y. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing* **2020**, *406*, 302–321. [CrossRef]

123. Sanjaya, Y.C.; Gunawan, A.A.S.; Irwansyah, E. Semantic Segmentation for Aerial Images: A Literature Review. *Eng. Math. Comput. Sci. (Emacs) J.* **2020**, *2*, 133–139. [CrossRef]
124. Chen, Y.L.; Cai, Y.R.; Cheng, M.Y. Vision-Based Robotic Object Grasping—A Deep Reinforcement Learning Approach. *Machines* **2023**, *11*, 275. [CrossRef]
125. Yadav, S.P.; Nagar, R.; Shah, S.V. Learning Vision-based Robotic Manipulation Tasks Sequentially in Offline Reinforcement Learning Settings. *arXiv* **2023**, arXiv:2301.13450.
126. Vuletić, J.; Polić, M.; Orsag, M. Robotic Strawberry Flower Treatment Based on Deep-Learning Vision. In *Human-Friendly Robotics 2022*; Borja, P., Della Santina, C., Peternel, L., Torta, E., Eds.; Springer International Publishing: Cham, Switzerland, 2023; Volume 26, pp. 189–204. [CrossRef]
127. Brogan, D.P.; DiFilippo, N.M.; Jouaneh, M.K. Deep Learning Computer Vision for Robotic Disassembly and Servicing Applications. *Array* **2021**, *12*, 100094. [CrossRef]
128. Keerthikeshwar, M.; Anto, S. Deep Learning for Robot Vision. In *Intelligent Manufacturing and Energy Sustainability*; Reddy, A., Marla, D., Favorskaya, M.N., Satapathy, S.C., Eds.; Springer: Singapore, 2021; Volume 213, pp. 357–365. [CrossRef]
129. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
130. Sun, W.; Wang, R. Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [CrossRef]
131. Browne, M.; Ghidary, S.S. Convolutional Neural Networks for Image Processing: An Application in Robot Vision. In *AI 2003: Advances in Artificial Intelligence*; Goos, G., Hartmanis, J., Van Leeuwen, J., Gedeon, T.D., Fung, L.C.C., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; Volume 2903, pp. 641–652. [CrossRef]
132. Ruiz-del-Solar, J.; Loncomilla, P.; Soto, N. A Survey on Deep Learning Methods for Robot Vision. *arXiv* **2018**, arXiv:1803.10862.
133. Bernstein, A.V.; Burnaev, E.V.; Kachan, O.N. Reinforcement Learning for Computer Vision and Robot Navigation. In *Machine Learning and Data Mining in Pattern Recognition*; Perner, P., Ed.; Springer International Publishing: Cham, Switzerland, 2018; Volume 10935, pp. 258–272. [CrossRef]
134. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
135. Laraba, S.; Tilmanne, J.; Dutoit, T. Leveraging Pre-trained CNN Models for Skeleton-Based Action Recognition. In *Computer Vision Systems*; Tzovaras, D., Giakoumis, D., Vincze, M., Argyros, A., Eds.; Springer International Publishing: Cham, Switzerland, 2019; Volume 11754, pp. 612–626. [CrossRef]
136. Zhang, J. Multi-Source Remote Sensing Data Fusion: Status and Trends. *Int. J. Image Data Fusion* **2010**, *1*, 5–24. [CrossRef]
137. Rajagopalan, S.S.; Morency, L.P.; Baltrušaitis, T.; Goecke, R. Extending Long Short-Term Memory for Multi-View Structured Learning. In *Computer Vision – ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; Volume 9911, pp. 338–353. [CrossRef]
138. Li, T.; Hua, M.; Wu, X.U. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2. 5). *IEEE Access* **2020**, *8*, 26933–26940. [CrossRef]
139. Kollias, D.; Zafeiriou, S. A Multi-component CNN-RNN Approach for Dimensional Emotion Recognition in-the-Wild. *arXiv* **2019**, arXiv:1805.01452.
140. Rožanec, J.M.; Zajec, P.; Theodoropoulos, S.; Koehorst, E.; Fortuna, B.; Mladenić, D. Synthetic Data Augmentation Using GAN for Improved Automated Visual Inspection. *Ifac-Papersonline* **2023**, *56*, 11094–11099. [CrossRef]
141. Tasdelen, A.; Sen, B. A Hybrid CNN-LSTM Model for Pre-miRNA Classification. *Sci. Rep.* **2021**, *11*, 14125. [CrossRef] [PubMed]
142. Zieba, M.; Wang, L. Training Triplet Networks with GAN. *arXiv* **2017**, arXiv:1704.02227.
143. Sergiyenko, O.Y.; Tyrsa, V.V. 3D Optical Machine Vision Sensors with Intelligent Data Management for Robotic Swarm Navigation Improvement. *IEEE Sens. J.* **2020**, *21*, 11262–11274. [CrossRef]
144. Jiang, H.; Peng, L.; Wang, X. Machine Vision and Big Data-Driven Sports Athletes Action Training Intervention Model. *Sci. Program.* **2021**, *2021*, 9956710. [CrossRef]
145. Elfiky, N. Application of Analytics in Machine Vision Using Big Data. *Asian J. Appl. Sci.* **2019**, *7*, 376–385. [CrossRef]
146. Popov, S.B. The Big Data Methodology in Computer Vision Systems. In Proceedings of the International Conference Information Technology and Nanotechnology (ITNT-2015), Samara, Russia, 29 June–1 July 2015; Volume 1490, pp. 420–425.
147. Tuor, T.; Wang, S.; Ko, B.J.; Liu, C.; Leung, K.K. Overcoming Noisy and Irrelevant Data in Federated Learning. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 5020–5027.
148. Zhang, H.; Bosch, J.; Olsson, H.H. Real-Time End-to-End Federated Learning: An Automotive Case Study. In Proceedings of the 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 12–16 July 2021; pp. 459–468.
149. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R. Advances and Open Problems in Federated Learning. In *Foundations and Trends® in Machine Learning*; Now Publishers Inc.: Boston, MA, USA, 2021; Volume 14, pp. 1–210.
150. Caldas, S.; Duddu, S.M.K.; Wu, P.; Li, T.; Konečný, J.; McMahan, H.B.; Smith, V.; Talwalkar, A. LEAF: A Benchmark for Federated Settings. *arXiv* **2019**, arXiv:1812.01097.

151. Tyagi, S.; Rajput, I.S.; Pandey, R. Federated Learning: Applications, Security Hazards and Defense Measures. In Proceedings of the 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), Dehradun, India, 17–18 March 2023; pp. 477–482.
152. Federated Learning: Collaborative Machine Learning without Centralized Training Data. 2017. Available online: https://blog.research.google/2017/04/federated-learning-collaborative.html (accessed on 9 March 2023).
153. Kant, S.; da Silva, J.M.B.; Fodor, G.; Göransson, B.; Bengtsson, M.; Fischione, C. Federated Learning Using Three-Operator ADMM. *IEEE J. Sel. Top. Signal Process.* **2022**, *17*, 205–221. [CrossRef]
154. Tao, J.; Gao, Z.; Guo, Z. Training Vision Transformers in Federated Learning with Limited Edge-Device Resources. *Electronics* **2022**, *11*, 2638. [CrossRef]
155. Guo, T.; Guo, S.; Wang, J. pFedPrompt: Learning Personalized Prompt for Vision-Language Models in Federated Learning. In Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 30 April–4 May 2023; pp. 1364–1374. [CrossRef]
156. Hubel, D.H.; Wiesel, T.N. Receptive Fields of Single Neurones in the Cat's Striate Cortex. *J. Physiol.* **1959**, *148*, 574. [CrossRef]
157. Hubel, D.H.; Wiesel, T.N. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. *J. Physiol.* **1962**, *160*, 106. [CrossRef]
158. Kandel, E.R. An Introduction to the Work of David Hubel and Torsten Wiesel. *J. Physiol.* **2009**, *587*, 2733. [CrossRef] [PubMed]
159. Wurtz, R.H. Recounting the Impact of Hubel and Wiesel. *J. Physiol.* **2009**, *587*, 2817–2823. [CrossRef] [PubMed]
160. Shimonomura, K.; Kushima, T.; Yagi, T. Binocular Robot Vision Emulating Disparity Computation in the Primary Visual Cortex. *Neural Netw.* **2008**, *21*, 331–340. [CrossRef] [PubMed]
161. Shimonomura, K.; Yagi, T. Neuromorphic Vergence Eye Movement Control of Binocular Robot Vision. In Proceedings of the 2010 IEEE International Conference on Robotics and Biomimetics, Tianjin, China, 14–18 December 2010; pp. 1774–1779.
162. Gochin, P.M.; Lubin, J.M. A Hierarchical Machine Vision System Based on a Model of the Primate Visual System. In Proceedings of the 5th IEEE International Symposium on Intelligent Control 1990, Philadelphia, PA, USA, 5–7 September 1990; pp. 61–65.
163. Zeevi, Y.Y. Adaptive Machine Vision: What Can Be Learned from Biological Systems. In *Intelligent Robots and Computer Vision VIII: Algorithms and Techniques*; SPIE: Philadelphia, PA, USA, 1990; Volume 1192, pp. 560–568.
164. Milner, B.; Squire, L.R.; Kandel, E.R. Cognitive Neuroscience and the Study of Memory. *Neuron* **1998**, *20*, 445–468. [CrossRef] [PubMed]
165. Allison, T. Electrophysiological Studies of Human Face Perception. I: Potentials Generated in Occipitotemporal Cortex by Face and Non-face Stimuli. *Cereb. Cortex* **1999**, *9*, 415–430. [CrossRef] [PubMed]
166. Dawson, G.; Webb, S.J.; McPartland, J. Understanding the Nature of Face Processing Impairment in Autism: Insights From Behavioral and Electrophysiological Studies. *Dev. Neuropsychol.* **2005**, *27*, 403–424. [CrossRef]
167. Di Nuovo, A.; Conti, D.; Trubia, G.; Buono, S.; Di Nuovo, S. Deep Learning Systems for Estimating Visual Attention in Robot-Assisted Therapy of Children with Autism and Intellectual Disability. *Robotics* **2018**, *7*, 25. [CrossRef]
168. El Arbaoui, F.E.Z.; El Hari, K.; Saidi, R. A Survey on the Application of the Internet of Things in the Diagnosis of Autism Spectrum Disorder. In *Advanced Technologies for Humanity*; Saidi, R., El Bhiri, B., Maleh, Y., Mosallam, A., Essaaidi, M., Eds.; Lecture Notes on Data Engineering and Communications Technologies; Springer: Cham, Switzerland, 2022; pp. 29–41. [CrossRef]
169. Javed, H.; Park, C.H. Behavior-Based Risk Detection of Autism Spectrum Disorder Through Child-Robot Interaction. In Proceedings of the Hri'20: Companion of the 2020 Acm/Ieee International Conference on Human-Robot Interaction, New York, NY, USA, 23–26 March 2020; pp. 275–277. [CrossRef]
170. Kollias, K.F.; Syriopoulou-Delli, C.K.; Sarigiannidis, P.; Fragulis, G.F. The Contribution of Machine Learning and Eye-tracking Technology in Autism Spectrum Disorder Research: A Review Study. In Proceedings of the 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST), Thessaloniki, Greece, 5–7 July 2021; pp. 1–4.
171. Kollias, K.F.; Syriopoulou-Delli, C.K.; Sarigiannidis, P.; Fragulis, G.F. The Contribution of Machine Learning and Eye-Tracking Technology in Autism Spectrum Disorder Research: A Systematic Review. *Electronics* **2021**, *10*, 2982. [CrossRef]
172. Kollias, K.F.; Syriopoulou-Delli, C.K.; Sarigiannidis, P.; Fragulis, G.F. Autism Detection in High-Functioning Adults with the Application of Eye-Tracking Technology and Machine Learning. In Proceedings of the 2022 11th International Conference on Modern Circuits and Systems Technologies (MOCAST), Bremen, Germany, 8–10 June 2022; pp. 1–4.
173. Kollias, K.F.; Maia Marques Torres E Silva, L.M.; Sarigiannidis, P.; Syriopoulou-Delli, C.K.; Fragulis, G.F. Implementation of Robots in Autism Spectrum Disorder Research: Diagnosis and Emotion Recognition and Expression. In Proceedings of the 2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST), Athens, Greece, 28–30 June 2023; pp. 1–4. [CrossRef]
174. Ramirez-Duque, A.A.; Frizera-Neto, A.; Bastos, T.F. Robot-Assisted Diagnosis for Children with Autism Spectrum Disorder Based on Automated Analysis of Nonverbal Cues. In Proceedings of the 2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob), Enschede, The Netherlands, 26–29 August 2018; pp. 456–461. [CrossRef]
175. Ramírez-Duque, A.A.; Frizera-Neto, A.; Bastos, T.F. Robot-Assisted Autism Spectrum Disorder Diagnostic Based on Artificial Reasoning. *J. Intell. Robot. Syst.* **2019**, *96*, 267–281. [CrossRef]
176. Riva, G.; Riva, E. CARERAID: Controlled Autonomous Robot for Early Detection and Rehabilitation of Autism and Intellectual Disability. *Cyberpsycho. Behav. Soc. Netw.* **2019**, *22*, 747–748. [CrossRef]
177. Romero-García, R.; Martínez-Tomás, R.; Pozo, P.; de la Paz, F.; Sarriá, E. Q-CHAT-NAO: A Robotic Approach to Autism Screening in Toddlers. *J. Biomed. Inform.* **2021**, *118*, 103797. [CrossRef]

178. Shelke, N.A.; Rao, S.; Verma, A.K.; Kasana, S.S. Autism Spectrum Disorder Detection Using AI and IoT. In Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing, Noida, India, 4–6 August 2022; pp. 213–219.
179. Shushma, G.; Jacob, I.J. Autism Spectrum Disorder Detection Using AI Algorithm. In Proceedings of the 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 23–25 February 2022; pp. 1–5.
180. Felleman, D.J.; Van Essen, D.C. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cereb. Cortex* **1991**, *1*, 1. [CrossRef] [PubMed]
181. Van Essen, D.C.; Drury, H.A.; Joshi, S.; Miller, M.I. Functional and Structural Mapping of Human Cerebral Cortex: Solutions Are in the Surfaces. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 788–795. [CrossRef] [PubMed]
182. Fox, M.D.; Snyder, A.Z.; Vincent, J.L.; Corbetta, M.; Van Essen, D.C.; Raichle, M.E. The Human Brain Is Intrinsically Organized into Dynamic, Anticorrelated Functional Networks. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 9673–9678. [CrossRef]
183. Glasser, M.F.; Coalson, T.S.; Robinson, E.C.; Hacker, C.D.; Harwell, J.; Yacoub, E.; Ugurbil, K.; Andersson, J.; Beckmann, C.F.; Jenkinson, M.; et al. A Multi-Modal Parcellation of Human Cerebral Cortex. *Nature* **2016**, *536*, 171–178. [CrossRef] [PubMed]
184. Tang, L.; Xiao, H.; Li, B. Can sam segment anything? when sam meets camouflaged object detection. *arXiv* **2023**, arXiv:2304.04709.
185. Wang, L.; Zhang, X.; Song, Z.; Bi, J.; Zhang, G.; Wei, H.; Tang, L.; Yang, L.; Li, J.; Jia, C.; et al. Multi-modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy. *IEEE Trans. Intell. Veh.* **2023**, *8*, 3781–3798. [CrossRef]
186. Rajasegaran, J.; Pavlakos, G.; Kanazawa, A.; Feichtenhofer, C.; Malik, J. On the Benefits of 3D Pose and Tracking for Human Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 18–22 June 2023; pp. 640–649.
187. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
188. Zhang, Y.; Guo, Q.; Du, Z.; Wu, A. Human Action Recognition for Dynamic Scenes of Emergency Rescue Based on Spatial-Temporal Fusion Network. *Electronics* **2023**, *12*, 538. [CrossRef]