Journal of Perception and Control
https://journals.zycentre.com/jpc

Article

Advances in Computer Vision & Image Processing: Image Enhancement, 3D Reconstruction, Motion Analysis, and Deep Learning-based Vision

Sarah Johnson[1], David Lee[2]*
1. Department of Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
2. Department of Electrical and Electronic Engineering, Imperial College London, London, United Kingdom

ABSTRACT
This paper reviews recent advances in computer vision and image processing, focusing on image enhancement, 3D reconstruction, motion analysis, and deep learning-driven systems. It explores classical and deep learning-based enhancement techniques in low-light, medical, and satellite imaging. 3D reconstruction methods, from stereo vision to deep learning, are assessed in applications like virtual reality and robotics. Motion analysis evaluates object detection and tracking under dynamic conditions. The impact of deep learning on object detection, segmentation, and recognition is emphasized, along with its integration into autonomous systems, intelligent transportation, and surveillance. Finally, challenges such as environmental robustness, computational constraints, and data scarcity are discussed, with future research directions outlined.

Keywords: Computer Vision; Image Processing; Image Enhancement; 3D Reconstruction; Motion Analysis; Deep Learning-based Vision; Perception Systems; Control Applications

*CORRESPONDING AUTHOR:
David Lee, Department of Electrical and Electronic Engineering, Imperial College London, Email: d.lee@imperial.ac.uk

ARTICLE INFO
Received: 29 July 2025 | Revised: 1 August 2025 | Accepted: 2 August 2025| Published Online: 4 August 2025
CITATION
Sarah Johnson, David Lee. 2025. Advances in Computer Vision & Image Processing: Image Enhancement, 3D Reconstruction, Motion Analysis, and Deep Learning-based Vision. Journal of Perception and Control, 1(1): 51-68.

1. Introduction

Computer vision and image processing have evolved from niche academic disciplines to cornerstone technologies that underpin modern intelligent systems, enabling machines to interpret, analyze, and interact with the visual world. These fields bridge the gap between raw sensory data and meaningful semantic understanding, empowering applications across industries as diverse as healthcare, manufacturing, entertainment, and aerospace. Over the past decade, the convergence of high-performance computing, large-scale annotated datasets, and breakthroughs in deep learning has catalyzed unprecedented advancements, pushing the boundaries of what machines can "see" and comprehend.

1.1 Background

The origins of computer vision can be traced back to the 1960s, with early efforts focused on simple tasks such as character recognition and edge detection. However, progress was slow due to limited computational power and the lack of robust algorithms for handling the complexity of real-world visual data. The 1990s and 2000s witnessed significant strides in traditional computer vision, with the development of techniques like SIFT (Scale-Invariant Feature Transform) for feature matching, Viola-Jones algorithm for face detection, and advances in probabilistic modeling for scene understanding. These methods, while groundbreaking, were often task-specific and struggled with variations in lighting, viewpoint, and occlusion.

The paradigm shift came in 2012 with the introduction of AlexNet, a deep convolutional neural network (CNN) that achieved a revolutionary performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). This marked the beginning of the deep learning era in computer vision, where data-driven approaches began to outperform handcrafted features across a wide range of tasks. Today, deep learning models dominate benchmarks in object detection, image segmentation, and visual recognition, enabling capabilities such as real-time object tracking in video streams, precise 3D reconstruction from single images, and semantic understanding of complex scenes.

Image enhancement, 3D reconstruction, motion analysis, and deep learning-based vision form the core pillars of modern computer vision systems. Image enhancement lays the foundation by improving image quality, making subsequent processing more reliable. 3D reconstruction provides a spatial understanding of the environment, crucial for tasks like navigation and manipulation. Motion analysis enables the interpretation of dynamic events, essential for applications involving moving objects. Deep learning, as an enabling technology, has enhanced the performance of each of these pillars, enabling solutions to previously intractable problems.

1.2 Significance in Perception and Control

Perception and control systems rely on accurate and timely interpretation of sensory data to make informed decisions and execute actions. In this context, computer vision serves as a primary sensory modality, offering rich, high-dimensional information about the environment. For instance, in autonomous robotics, vision systems must perceive obstacles, recognize objects, and understand the spatial layout of the scene to plan safe and efficient paths. Similarly, in industrial automation, computer vision is used for quality control, where it inspects products for defects with a precision that surpasses human capabilities.

The integration of computer vision into perception and control loops is particularly critical in safety-critical applications. In autonomous vehicles, for example, the ability to detect pedestrians, lane markings, and other vehicles in real-time, even under adverse weather conditions, directly impacts passenger safety. Image enhancement ensures that vision systems remain robust in low-light or foggy conditions, while 3D reconstruction provides depth information necessary for estimating distances to other objects. Motion analysis tracks the movement of surrounding entities, allowing the vehicle to predict their trajectories and adjust its speed or direction

accordingly.

Beyond robotics and transportation, computer vision plays a vital role in healthcare, where it aids in medical imaging analysis. Enhanced medical images enable more accurate diagnosis of diseases, such as detecting tumors in MRI scans. 3D reconstruction of anatomical structures from 2D medical images helps surgeons plan complex procedures, while motion analysis of cardiac or joint movements provides insights into physiological functions.

1.3 Structure of the Paper

This paper is structured to provide a comprehensive and coherent overview of the field, guiding readers through the foundational techniques, advanced methodologies, and practical applications of computer vision. Section 2 delves into image enhancement, comparing traditional methods such as histogram equalization and filtering with deep learning-based approaches like CNNs and GANs. Section 3 focuses on 3D reconstruction, exploring both traditional techniques (stereo vision, SfM) and deep learning innovations (deep stereo matching, single-image 3D reconstruction). Section 4 examines motion analysis, covering motion detection, object tracking, and motion estimation, with a discussion of both classical algorithms and deep learning advancements. Section 5 explores deep learning-based vision in detail, highlighting its applications in object detection, image segmentation, and visual recognition. Section 6 discusses the integration of these techniques into perception and control systems, with case studies in robotics, autonomous vehicles, and surveillance. Section 7 identifies key challenges facing the field and outlines promising future directions. Finally, Section 8 concludes the paper by summarizing the main findings and emphasizing the importance of continued research in advancing computer vision technologies.

2. Image Enhancement

Image enhancement is a preprocessing step that aims to improve the visual quality of images by reducing noise, enhancing contrast, sharpening details, and correcting for artifacts introduced during image acquisition. The goal is to produce an image that is more suitable for human interpretation or for subsequent computer vision tasks such as object detection and segmentation. Image enhancement is particularly important in scenarios where image quality is compromised, such as low-light photography, satellite imaging with atmospheric distortion, or medical imaging with low signal-to-noise ratios.

2.1 Traditional Image Enhancement Methods

Traditional image enhancement methods are based on mathematical transformations and statistical analysis of image pixels. These methods are often computationally efficient and easy to implement, making them suitable for real-time applications. However, they may lack the flexibility to handle complex image degradations or adapt to varying scenarios.

2.1.1 Histogram Equalization

Histogram equalization is a histogram-based technique that adjusts the intensity distribution of an image to enhance contrast. The underlying principle is to spread out the most frequent intensity values, thereby increasing the dynamic range of the image. This is achieved by transforming the intensity values such that the cumulative distribution function (CDF) of the enhanced image is approximately uniform.

Global histogram equalization applies a single transformation to the entire image, which can be effective for images with a narrow intensity range. However, it may over-enhance noise in homogeneous regions and lead to unnatural-looking results. To address these limitations, local histogram equalization (LHE) processes the image in small, overlapping regions (tiles). For each tile, a histogram is computed and equalized, allowing for better preservation of local details. Adaptive histogram equalization (AHE) is a variant of LHE that limits the contrast enhancement in each tile to avoid noise amplification, using a clip limit to cap the histogram bins. Contrast

Limited Adaptive Histogram Equalization (CLAHE) is a widely used implementation of AHE that has been successful in medical imaging, particularly in enhancing X-ray and MRI images [1].

Despite their advantages, histogram-based methods have limitations. They do not account for spatial correlations between pixels, which can lead to over-enhancement of noise in smooth regions. Additionally, they may not perform well in images with complex intensity distributions, such as those with multiple illumination sources.

2.1.2 Filtering Techniques

Filtering is a fundamental technique in image enhancement, used primarily for noise reduction and edge preservation. Filters operate by convolving the image with a kernel, where the kernel values determine the transformation applied to each pixel and its neighbors.

Linear filters, such as Gaussian filters, are used for smoothing images by averaging pixel values in a local neighborhood. The Gaussian kernel weights pixels according to a Gaussian function, with closer pixels contributing more to the average. This results in a blurring effect that reduces high-frequency noise but can also smooth out fine details and edges.

Non-linear filters are more effective than linear filters in preserving edges while reducing noise. Median filters replace each pixel with the median value of its neighborhood, making them highly effective in removing impulse noise (salt-and-pepper noise) without significantly blurring edges. However, they may introduce artifacts in images with fine textures.

Bilateral filters are another class of non-linear filters that smooth the image while preserving edges. They consider both the spatial distance and the intensity difference between pixels, applying a larger weight to pixels that are close in both space and intensity. This allows the filter to smooth regions with similar intensities while maintaining sharp edges between regions of different intensities [2]. Bilateral filtering is widely used in applications such as portrait photography, where it smooths skin tones while preserving facial features.

Anisotropic diffusion is a filtering technique that adapts to image edges by smoothing along edges rather than across them. It works by iteratively diffusing pixel values, with the diffusion rate controlled by a function of the image gradient. In regions with high gradients (edges), the diffusion rate is reduced, preserving the edge; in smooth regions, the diffusion rate is higher, reducing noise. This technique is effective for noise reduction while maintaining edge integrity but can be computationally expensive due to its iterative nature.

2.2 Deep Learning-based Image Enhancement

Deep learning has revolutionized image enhancement by enabling models to learn complex mappings from degraded images to high-quality images. These models are trained on large datasets of paired degraded and clean images, allowing them to capture intricate patterns and adapt to a wide range of image degradations.

2.2.1 Convolutional Neural Networks (CNNs) for Image Denoising

CNNs have emerged as powerful tools for image denoising, outperforming traditional methods in scenarios with high noise levels or complex noise patterns. CNN-based denoising models learn to map noisy images to clean images by leveraging the hierarchical feature learning capabilities of convolutional layers.

DnCNN (Denoising Convolutional Neural Network) is a pioneering model that uses deep CNNs for image denoising. It consists of multiple convolutional layers with ReLU activation functions, followed by a final convolutional layer that outputs the residual (the difference between the noisy image and the clean image). By learning the residual, DnCNN focuses on modeling the noise, which simplifies the learning process [3]. The model is trained on synthetically noisy images generated by adding Gaussian noise to clean images, enabling it to

generalize to different noise levels.

Other CNN-based denoising models include FFDNet (Fast and Flexible Denoising Network), which handles varying noise levels by incorporating noise level maps as input, and CBDNet (Content-Blind Denoising Network), which is designed to denoise images without prior knowledge of the noise type. These models have shown superior performance in denoising real-world images, such as those affected by sensor noise in low-light conditions.

2.2.2 Generative Adversarial Networks (GANs) for Image Enhancement

Generative Adversarial Networks (GANs) have proven highly effective in image enhancement tasks, particularly in generating visually realistic results. A GAN consists of two networks: a generator that produces enhanced images from degraded inputs, and a discriminator that distinguishes between real enhanced images and those generated by the generator. Through adversarial training, the generator learns to produce images that are indistinguishable from real high-quality images, while the discriminator becomes increasingly adept at detecting fakes.

ESRGAN (Enhanced Super-Resolution Generative Adversarial Network) is an extension of the SRGAN model, designed for both super-resolution and image enhancement. It uses a residual-in-residual dense block (RRDB) architecture to capture rich feature information, and a relativistic discriminator that learns to distinguish between the relative quality of generated and real images. ESRGAN produces images with sharper details and more natural textures compared to traditional super-resolution methods [4].

Pix2Pix is another GAN-based model that performs image-to-image translation, which can be applied to tasks such as enhancing low-light images, converting grayscale images to color, and removing rain or snow from images. It uses a U-Net architecture for the generator and a PatchGAN discriminator that evaluates image patches rather than the entire image, enabling it to capture local details.

While GANs produce visually appealing results, they suffer from challenges such as training instability, mode collapse (where the generator produces a limited range of outputs), and difficulty in quantifying performance due to the lack of a clear objective function. Recent advancements, such as StyleGAN and CycleGAN, have addressed some of these issues by introducing style-based generators and cycle consistency losses, respectively.

2.2.3 Transformer-based Image Enhancement

Transformers, originally developed for natural language processing, have recently been applied to image enhancement with promising results. Unlike CNNs, which are limited by local receptive fields, transformers use self-attention mechanisms to capture long-range dependencies in images, making them effective for tasks that require global context.

IPT (Image Processing Transformer) is a transformer-based model that achieves state-of-the-art performance in various image enhancement tasks, including denoising, super-resolution, and deraining. It uses a multi-head self-attention mechanism to model relationships between pixels across the entire image, enabling it to handle complex image degradations. IPT is trained in a unified manner for multiple tasks, leveraging shared features to improve generalization.

Another example is the Swin Transformer, which divides the image into non-overlapping patches and applies self-attention within local windows, reducing computational complexity while maintaining the ability to capture long-range dependencies. Swin Transformer-based models have shown excellent performance in image denoising and super-resolution, particularly in preserving fine details.

Transformer-based methods offer advantages in handling global image structures but are often more computationally intensive than CNNs, making them challenging to deploy in real-time applications. Research is ongoing to develop lightweight transformer architectures that balance performance and efficiency.

3. 3D Reconstruction

3D reconstruction is the process of creating three-dimensional digital representations of physical objects or scenes from one or more two-dimensional images or other sensory data. This technology has applications in diverse fields, including virtual reality (VR), augmented reality (AR), robotics, medical imaging, and cultural heritage preservation. 3D reconstruction enables machines to perceive the spatial layout of the environment, which is essential for tasks such as navigation, manipulation, and interaction with physical objects.

3.1 Traditional 3D Reconstruction Techniques

Traditional 3D reconstruction techniques rely on geometric principles and photogrammetric methods to recover 3D structure from 2D images. These methods typically require multiple images of the scene taken from different viewpoints or additional information such as camera calibration parameters.

3.1.1 Stereo Vision

Stereo vision, inspired by human binocular vision, uses two or more cameras to capture images of a scene from different viewpoints. By calculating the disparity (the difference in position) of corresponding points in these images, the depth of the points can be inferred using triangulation.

The key steps in stereo vision are camera calibration, feature matching, disparity estimation, and depth calculation. Camera calibration determines the intrinsic parameters (focal length, principal point) and extrinsic parameters (position and orientation) of each camera, which are necessary for accurate 3D reconstruction. Feature matching involves identifying corresponding points in the stereo images, using features such as SIFT or SURF that are invariant to scale, rotation, and illumination changes.

Disparity estimation is the most challenging step in stereo vision, as it requires finding the correct match for each pixel in the left image within the right image. Global methods for disparity estimation, such as graph cuts and belief propagation, model the problem as an energy minimization task, considering both the similarity of pixel intensities and the smoothness of the disparity map. Local methods, such as block matching, compare small windows around each pixel to find the best match, offering faster computation but potentially less accurate results [5].

Stereo vision is widely used in robotics for obstacle detection and navigation, as well as in autonomous vehicles for depth perception. However, it has limitations, including sensitivity to image noise, occlusion, and textureless regions, where feature matching is difficult.

3.1.2 Structure from Motion (SfM)

Structure from Motion (SfM) reconstructs the 3D structure of a scene from a sequence of images taken from unknown viewpoints. Unlike stereo vision, SfM does not require calibrated cameras, making it more flexible for applications such as photogrammetry and virtual tourism.

The SfM pipeline consists of several stages: feature detection and matching, camera pose estimation, triangulation, and bundle adjustment. Feature detection and matching identify corresponding points across images, similar to stereo vision. Camera pose estimation determines the position and orientation of each camera relative to the scene using the matched features. Triangulation computes the 3D coordinates of the matched points using the estimated camera poses. Bundle adjustment is a global optimization step that refines both the camera poses and the 3D point coordinates to minimize the reprojection error (the difference between the observed and predicted positions of points in the images) [6].

SfM can handle unordered image collections, making it suitable for reconstructing large-scale scenes from internet photos, as demonstrated by projects like Photo Tourism [6]. However, it is prone to error accumulation, especially in large scenes, which can lead to drift in the reconstructed structure. Additionally, SfM performs poorly in textureless or

dynamically changing scenes, where feature matching is unreliable.

3.1.3 Multi-View Stereo (MVS)

Multi-View Stereo (MVS) extends stereo vision to multiple images, enabling more accurate and dense 3D reconstructions. MVS algorithms generate a dense point cloud by matching pixels across multiple viewpoints, leveraging the redundancy provided by additional images to handle occlusion and textureless regions.

MVS methods can be categorized into feature-based, patch-based, and depth-map fusion approaches. Feature-based methods use sparse feature matches to initialize the reconstruction, then propagate depth information to neighboring pixels. Patch-based methods compare image patches across multiple views to estimate depth for each pixel, using photometric consistency as a criterion. Depth-map fusion methods generate a depth map for each image using stereo matching, then fuse these depth maps into a consistent 3D model, resolving conflicts between overlapping depth maps.

MVS has been used to create detailed 3D models of historical sites, such as the ruins of Pompeii, and in industrial inspection for quality control. However, it is computationally intensive, requiring significant processing power for large datasets.

3.2 Deep Learning in 3D Reconstruction

Deep learning has transformed 3D reconstruction by enabling end-to-end learning of 3D structure from images, reducing the reliance on handcrafted features and geometric assumptions. These models leverage large datasets to learn complex mappings from 2D images to 3D representations.

3.2.1 Deep Stereo Matching

Deep learning-based stereo matching models have achieved state-of-the-art performance in disparity estimation by learning to extract discriminative features and model the dependencies between pixels. These models typically use CNNs to extract features from left and right images, then compute a cost volume that measures the similarity between features at different disparities. The cost volume is then processed to produce a disparity map.

PSMNet (Pyramid Stereo Matching Network) is a leading deep stereo matching model that uses a pyramid feature extraction network to capture multi-scale features. It constructs a cost volume using concatenated features from the left and right images, then applies 3D convolutional layers to regularize the cost volume and predict the disparity map [7]. PSMNet outperforms traditional methods in challenging scenarios such as textureless regions and occlusions, thanks to its ability to learn context-aware features.

Other deep stereo models, such as GANet (Gated Attention Network), incorporate attention mechanisms to focus on relevant features during cost volume aggregation, further improving performance. These models have been deployed in autonomous vehicles and robotics, where accurate depth perception is critical.

3.2.2 Single Image 3D Reconstruction

Reconstructing a 3D model from a single image is a challenging task, as it requires inferring the missing depth information from 2D cues such as perspective, shading, and texture. Deep learning models have made significant progress in this area by leveraging prior knowledge of object shapes learned from large datasets.

MeshRCNN extends Mask R-CNN to predict 3D meshes of objects from single RGB images. It uses a CNN to extract image features, then predicts a 3D mesh proposal for each detected object. A graph neural network refines the mesh by enforcing geometric constraints, resulting in detailed 3D shapes [8]. MeshRCNN is capable of reconstructing complex objects with varying topologies, making it suitable for applications such as virtual reality and product design.

Another approach to single image 3D reconstruction is to predict a depth map, which can be converted into a point cloud or a 3D mesh. Models like DORN (Deep Ordinal Regression Network)

predict depth using ordinal regression, treating depth estimation as a classification task where each class corresponds to a range of depth values. This approach is robust to ambiguous depth cues and has been used in autonomous driving for monocular depth estimation.

Despite recent advancements, single image 3D reconstruction remains limited by the ambiguity of 2D-to-3D mapping, especially for objects with symmetric shapes or complex geometries. Future research is focused on incorporating additional cues, such as semantic information or physical constraints, to improve reconstruction accuracy.

3.2.3 Volumetric and Implicit Representations

Deep learning has also enabled the use of volumetric and implicit representations for 3D reconstruction. Volumetric methods represent the 3D scene as a voxel grid, where each voxel indicates the presence or absence of matter. Models like 3D U-Net generate volumetric reconstructions by processing 2D images with a 3D convolutional network.

Implicit representations, such as signed distance functions (SDFs) or neural radiance fields (NeRFs), define the 3D shape as a continuous function. NeRF, for example, uses a neural network to map 3D coordinates and viewing directions to color and density, enabling high-quality 3D reconstructions from multiple images. NeRF has revolutionized novel view synthesis and 3D reconstruction, producing photorealistic results for small-scale scenes.

These representations are particularly useful for reconstructing complex, detailed shapes but can be computationally expensive, requiring significant memory and processing power.

4. Motion Analysis

Motion analysis involves the detection, tracking, and estimation of motion in image sequences, enabling the understanding of dynamic events and the behavior of moving objects. This field is critical for applications such as video surveillance, human-computer interaction, sports analysis, and autonomous driving, where the ability to interpret motion is essential for decision-making.

4.1 Motion Detection

Motion detection aims to identify regions in a video that correspond to moving objects, separating them from the static background. This is typically the first step in motion analysis, providing a focus for subsequent processing such as tracking and recognition.

4.1.1 Background Subtraction

Background subtraction is a widely used motion detection technique that models the background of the scene and subtracts it from each frame to detect foreground objects. The key challenge is to maintain an accurate background model that adapts to changes in lighting, weather, and other environmental factors.

Gaussian Mixture Models (GMMs) are commonly used for background modeling, where each pixel is represented by a mixture of Gaussian distributions. These distributions model the variations in pixel intensity over time, with the most persistent distributions corresponding to the background. For each new frame, pixels are classified as foreground if their intensity does not fit any of the background Gaussians [9]. GMMs can adapt to gradual changes in the background but may struggle with sudden changes or dynamic backgrounds (e.g., waving trees).

More recent background subtraction methods use deep learning to model the background, leveraging the ability of CNNs to capture complex patterns. For example, DeepBS models the background as a deep neural network that generates frames similar to the background, with foreground pixels identified as those that deviate significantly from the generated background. These models offer improved performance in challenging scenarios but are more computationally intensive.

4.1.2 Optical Flow-based Motion Detection

Optical flow estimates the motion of pixels between consecutive frames, providing a dense motion field that can be used to detect moving objects.

Pixels with significant flow magnitudes are classified as foreground, while those with little or no flow are considered background.

The Lucas-Kanade algorithm is a classic method for estimating sparse optical flow, computing the flow for a set of feature points by assuming that the flow is constant within a local window. Dense optical flow algorithms, such as Farneback, estimate flow for every pixel by fitting polynomial expansions to local image patches [11]. Optical flow-based motion detection is robust to changes in lighting but can be sensitive to noise and occlusion.

4.2 Object Tracking

Object tracking involves following the movement of one or more objects in a video sequence, maintaining their identities over time. It is essential for applications such as video surveillance (tracking suspects), human-computer interaction (tracking hand gestures), and sports analysis (tracking athletes).

4.2.1 Correlation Filters

Correlation filters are efficient tracking algorithms that learn a filter to match the appearance of the target object. The filter is trained to maximize the correlation between the target and its surroundings, enabling fast detection of the target in subsequent frames.

Kernelized Correlation Filters (KCF) extend correlation filters by using kernel functions to map the image features to a higher-dimensional space, allowing them to model non-linear relationships. KCF achieves real-time performance by leveraging the Fast Fourier Transform (FFT) for efficient filter training and application [10]. However, KCF may fail when the target undergoes significant appearance changes (e.g., rotation, scaling) or is occluded.

4.2.2 Deep Learning-based Tracking

Deep learning has significantly improved tracking performance by enabling models to learn robust features that are invariant to appearance changes and occlusion. Siamese networks are a popular architecture for tracking, where two identical subnetworks extract features from the target template and the search region, with the similarity between the features used to locate the target.

SiamRPN (Siamese Region Proposal Network) combines a Siamese network with a region proposal network (RPN) to generate bounding box proposals for the target. The RPN predicts the location and size of the target, enabling accurate tracking even when the target's appearance changes [10]. SiamRPN achieves state-of-the-art performance in challenging tracking benchmarks, outperforming traditional methods in scenarios involving occlusion, scale variation, and fast motion.

Transformer-based trackers, such as TransT, use self-attention mechanisms to model the relationships between the target and the surrounding context, further improving robustness. These trackers are capable of handling complex scenes with multiple objects and cluttered backgrounds.

4.2.3 Multi-Object Tracking (MOT)

Multi-Object Tracking (MOT) extends single-object tracking to multiple objects, requiring the algorithm to track each object while maintaining their identities. MOT is more challenging than single-object tracking due to occlusion, overlapping objects, and varying object counts.

MOT methods typically combine detection and tracking, using detection algorithms to localize objects in each frame and association algorithms to link detections across frames. DeepSORT (Deep Simple Online and Realtime Tracking) uses CNN features to represent object appearances, enabling accurate association even when objects are occluded. It combines the appearance features with motion information (using a Kalman filter) to predict object positions and resolve identity switches.

Recent MOT approaches use end-to-end deep learning models that jointly learn to detect and track objects, such as TrackR-CNN and QDTrack. These models optimize both detection and tracking performance in a unified framework, achieving state-of-the-art results on MOT benchmarks.

4.3 Motion Estimation

Motion estimation computes the motion vectors that describe the displacement of pixels or objects between consecutive frames. This information is used in video compression (to reduce redundancy), video stabilization, and activity recognition.

4.3.1 Optical Flow

Optical flow is a fundamental motion estimation technique that estimates the apparent motion of brightness patterns in the image. It is based on the assumption that the brightness of a pixel remains constant between frames (the brightness constancy assumption).

Lucas-Kanade is a sparse optical flow algorithm that estimates flow for a set of feature points by solving a system of linear equations derived from the brightness constancy assumption. It assumes that the flow is constant within a local window, making it efficient but sensitive to large motions.

Dense optical flow algorithms, such as Farneback, estimate flow for every pixel by fitting a polynomial to the image intensity function in a local window and computing the flow from the polynomial coefficients [11]. Dense optical flow provides a comprehensive view of motion in the scene but is more computationally expensive than sparse methods.

Deep learning-based optical flow models, such as FlowNet and PWC-Net, have achieved significant improvements in accuracy. FlowNet uses a CNN to directly predict optical flow from pairs of images, with separate encoders for each image and a decoder that combines the features to estimate flow [11]. PWC-Net improves efficiency by using pyramid warping and cost volume filtering, making it suitable for real-time applications.

4.3.2 Motion Estimation for Rigid and Non-Rigid Objects

Motion estimation can be categorized into rigid and non-rigid motion, depending on whether the object maintains a fixed shape during movement. Rigid motion estimation (e.g., tracking a moving car) is often modeled using affine transformations, which describe translation, rotation, scaling, and shearing.

Non-rigid motion estimation (e.g., tracking a person's face) is more complex, requiring models that can handle deformations. Active Appearance Models (AAMs) are used for non-rigid motion estimation, combining a shape model with an appearance model to track deformable objects. Deep learning approaches, such as CNNs with spatial transformers, have also been applied to non-rigid motion estimation, achieving state-of-the-art results in tasks like facial landmark tracking.

5. Deep Learning-based Vision

Deep learning has emerged as the dominant paradigm in computer vision, enabling breakthroughs in tasks that were previously considered intractable. By learning hierarchical representations from data, deep neural networks can capture complex visual patterns, leading to significant improvements in accuracy and robustness. This section focuses on the applications of deep learning in object detection, image segmentation, and visual recognition.

5.1 Object Detection

Object detection involves locating and classifying objects in an image, outputting bounding boxes and class labels for each detected object. It is a fundamental task in computer vision, with applications in autonomous driving (detecting vehicles, pedestrians), surveillance (detecting suspicious objects), and robotics (detecting graspable objects).

5.1.1 Two-Stage Detectors

Two-stage detectors first generate region proposals (potential object locations) and then classify these proposals. Faster R-CNN is a landmark two-stage detector that uses a Region Proposal Network (RPN) to generate proposals efficiently. The RPN shares convolutional features with the classification network, enabling end-to-end training. For each proposal, a RoI (Region of Interest) pooling layer extracts fixed-size features, which are then

classified by a fully connected network [13]. Faster R-CNN achieves high accuracy but has a relatively slow inference speed, limiting its use in real-time applications.

Cascade R-CNN improves upon Faster R-CNN by using a sequence of detectors trained with increasing IoU (Intersection over Union) thresholds. This cascaded approach refines the proposals iteratively, reducing false positives and improving detection accuracy for small or occluded objects.

5.1.2 One-Stage Detectors

One-stage detectors skip the region proposal step, directly predicting bounding boxes and class probabilities for each pixel or grid cell. This makes them faster than two-stage detectors, making them suitable for real-time applications.

YOLO (You Only Look Once) divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell. YOLOv5, the latest iteration, incorporates improvements such as cross-stage partial networks (CSP) for feature extraction, spatial pyramid pooling (SPP) for handling varying object sizes, and efficient NMS (Non-Maximum Suppression) for post-processing [12]. YOLOv5 balances speed and accuracy, making it popular in applications like autonomous driving and video surveillance.

SSD (Single Shot MultiBox Detector) uses multiple feature maps at different scales to detect objects of varying sizes, with each feature map responsible for predicting objects at a specific scale. SSD is faster than Faster R-CNN but may struggle with small objects.

5.1.3 Transformer-based Detectors

Transformer-based detectors, such as DETR (Detection Transformer), use self-attention mechanisms to model the relationships between objects and image features. DETR treats object detection as a set prediction problem, directly outputting a set of bounding boxes and class labels without relying on handcrafted components like NMS. It uses a CNN to extract image features, which are then processed by a transformer encoder-decoder architecture to predict the final detections. DETR achieves competitive performance with Faster R-CNN while offering a simpler, end-to-end framework.

5.2 Image Segmentation

Image segmentation divides an image into meaningful regions, assigning a label to each pixel. It is more granular than object detection, providing detailed information about the shape and structure of objects. Image segmentation has applications in medical imaging (segmenting tumors), autonomous driving (segmenting road, sky, and obstacles), and satellite imagery (segmenting land cover types).

5.2.1 Semantic Segmentation

Semantic segmentation classifies each pixel into a predefined category (e.g., “car,” “road,” “pedestrian”). U-Net is a widely used semantic segmentation architecture with an encoder-decoder structure. The encoder downsamples the image to capture high-level semantic features, while the decoder upsamples the features and combines them with skip connections from the encoder to recover fine-grained details [13]. U-Net has been highly successful in medical imaging, where accurate segmentation of small structures is critical.

DeepLab is another prominent semantic segmentation model that uses atrous convolution (dilated convolution) to increase the receptive field without reducing spatial resolution. It incorporates a spatial pyramid pooling module to capture multi-scale context, enabling robust segmentation of objects at different sizes. DeepLabv3+ extends this with an encoder-decoder structure, further improving performance on small objects.

5.2.2 Instance Segmentation

Instance segmentation distinguishes between different instances of the same class (e.g., separating two cars in an image). Mask R-CNN is a pioneering instance segmentation model that extends Faster R-CNN by adding a mask branch to predict a binary mask for each detected object [13]. The mask branch

uses a CNN to predict a pixel-wise mask within the bounding box proposed by the RPN, achieving high accuracy in both object detection and segmentation.

Recent instance segmentation models, such as Mask2Former, use transformers to model the relationships between object instances and image pixels. Mask2Former achieves state-of-the-art performance by leveraging the ability of transformers to capture long-range dependencies, making it effective for segmenting complex scenes with overlapping objects.

5.3 Visual Recognition

Visual recognition involves identifying and classifying objects, scenes, and activities in images and videos. It is a broad field that includes image classification, scene recognition, and video action recognition.

5.3.1 Image Classification

Image classification assigns a label to an entire image (e.g., "cat," "dog," "mountain"). Convolutional Neural Networks (CNNs) have revolutionized image classification, with architectures like AlexNet, VGG, and ResNet achieving successive improvements in accuracy.

ResNet (Residual Network) addresses the problem of vanishing gradients in deep networks by introducing residual blocks, which learn residual mappings instead of direct mappings. This allows training of extremely deep networks (up to 152 layers) without performance degradation [14]. ResNet won the ILSVRC 2015 competition, setting a new state-of-the-art in image classification.

More recent image classification models, such as EfficientNet, use neural architecture search to optimize network depth, width, and resolution, achieving higher accuracy with fewer parameters. EfficientNet scales these dimensions uniformly using a compound coefficient, balancing performance and efficiency.

5.3.2 Video Action Recognition

Video action recognition involves classifying the action being performed in a video (e.g., "running," "jumping," "driving"). It requires modeling both spatial and temporal information, as actions unfold over time.

I3D (Inflated 3D Network) extends 2D CNNs to 3D by inflating the filters and pooling kernels from pre-trained 2D models. This allows the network to capture spatiotemporal features, enabling effective action recognition [15]. I3D achieves state-of-the-art performance on benchmark datasets like Kinetics, demonstrating the effectiveness of 3D convolutions for video understanding.

SlowFast networks are another approach to video action recognition, using two pathways: a slow pathway that processes low-frame-rate video to capture spatial details, and a fast pathway that processes high-frame-rate video to capture temporal dynamics. The pathways are fused to combine spatial and temporal information, achieving high accuracy with efficient computation.

Transformer-based models, such as Video Swin Transformer, apply self-attention across both spatial and temporal dimensions, enabling them to capture long-range spatiotemporal dependencies. These models have set new state-of-the-art results on video action recognition benchmarks, highlighting the potential of transformers for video understanding.

6. Integration in Perception and Control Systems

The true power of computer vision techniques lies in their integration into perception and control systems, where they enable intelligent decision-making and action execution. This section explores how image enhancement, 3D reconstruction, motion analysis, and deep learning-based vision are combined to solve real-world problems in robotics, autonomous vehicles, and surveillance.

6.1 Robotics

Robots rely on perception systems to interact with their environment, and computer vision is a

key component of these systems. In mobile robotics, vision-based navigation requires the robot to perceive obstacles, recognize landmarks, and understand the spatial layout of the scene.

Image enhancement techniques are used to improve the quality of images captured by the robot's cameras, especially in challenging environments such as low-light warehouses or outdoor scenes with varying illumination. For example, denoising algorithms reduce noise in images captured by low-cost cameras, while contrast enhancement improves the visibility of obstacles in dimly lit areas.

3D reconstruction enables robots to build maps of their environment and localize themselves within these maps (Simultaneous Localization and Mapping, SLAM). Visual SLAM systems, such as ORB-SLAM, use SfM techniques to reconstruct the 3D structure of the scene while tracking the robot's pose. Deep learning-based 3D reconstruction methods, such as those using CNNs for depth estimation, are increasingly being integrated into SLAM systems to improve accuracy and robustness.

Motion analysis is crucial for robots interacting with dynamic environments, such as human-robot collaboration. Object tracking allows robots to follow moving objects (e.g., a human worker passing a tool), while motion estimation helps predict the future positions of these objects to avoid collisions. For example, in industrial settings, robots equipped with vision systems can track the movement of parts on a conveyor belt, adjusting their grasp positions accordingly.

Deep learning-based vision enables robots to recognize and manipulate objects with a high degree of autonomy. Object detection and segmentation algorithms allow robots to identify objects in cluttered scenes, while grasp planning networks predict optimal grasp points based on the object's shape and orientation. For instance, the Google Brain Robot team has developed robots that use deep learning to pick and place objects in unstructured environments, demonstrating the effectiveness of vision-based control.

6.2 Autonomous Vehicles

Autonomous vehicles (AVs) depend on a suite of sensors, including cameras, LiDAR, and radar, to perceive their environment. Computer vision plays a central role in processing camera data, providing rich information about the road, traffic participants, and traffic signals.

Image enhancement is critical for AVs to operate in diverse weather and lighting conditions. For example, in rain or fog, image dehazing algorithms improve visibility by removing atmospheric scattering, while low-light enhancement techniques brighten images captured at night. These enhancements ensure that subsequent vision tasks, such as object detection, remain reliable.

3D reconstruction from stereo cameras or LiDAR-camera fusion provides AVs with a 3D representation of the environment, essential for depth perception and collision avoidance. Stereo vision systems estimate depth using disparity, while LiDAR provides precise distance measurements that can be fused with camera data to improve accuracy. Deep learning-based depth estimation models, such as those using CNNs, complement these sensors by providing dense depth maps in regions where LiDAR data is sparse.

Motion analysis enables AVs to track the movement of other vehicles, pedestrians, and cyclists, predicting their trajectories to make informed driving decisions. Optical flow and object tracking algorithms estimate the speed and direction of moving objects, allowing the AV's control system to adjust speed, brake, or steer to avoid collisions. For example, if a pedestrian is detected crossing the road, the AV can predict their path and slow down or stop in time.

Deep learning-based vision is at the core of AV perception systems, with object detection models identifying traffic lights, stop signs, and lane markings. Semantic segmentation algorithms classify each pixel in the image into categories such as "road," "sidewalk," and "vehicle," providing a detailed understanding of the scene layout. These outputs are

fed into the AV's decision-making system, which uses them to plan trajectories and control the vehicle's actuators.

Companies like Tesla and Waymo have demonstrated the potential of vision-based autonomous driving, with their vehicles navigating complex urban environments using camera data processed by deep learning algorithms. However, challenges remain, such as handling edge cases (e.g., unusual weather conditions) and ensuring the safety of these systems.

6.3 Surveillance Systems

Surveillance systems use computer vision to monitor public spaces, detect threats, and ensure public safety. Modern surveillance systems are increasingly intelligent, leveraging advanced vision techniques to automate tasks that were previously performed by human operators.

Image enhancement is used to improve the quality of surveillance footage, which is often captured in low-light conditions or with low-resolution cameras. Super-resolution algorithms upscale images to reveal fine details, such as facial features or license plates, while denoising techniques reduce graininess in night vision footage.

Motion detection is a basic function of surveillance systems, triggering alerts when unusual movement is detected. Background subtraction algorithms are commonly used to detect foreground objects, with deep learning-based methods improving performance in dynamic backgrounds (e.g., busy streets with moving crowds).

Object tracking allows surveillance systems to follow the movement of individuals or vehicles across multiple cameras, providing a comprehensive view of their path. Multi-object tracking algorithms, such as DeepSORT, maintain the identities of tracked objects even when they are occluded by other objects or move out of the camera's field of view.

Deep learning-based visual recognition enables advanced surveillance capabilities, such as face recognition and anomaly detection. Face recognition systems can identify individuals from surveillance footage, helping law enforcement agencies locate suspects. Anomaly detection algorithms learn normal patterns of behavior in a scene (e.g., pedestrian movement in a park) and alert operators when deviations occur (e.g., a person running towards a restricted area).

The integration of computer vision with control systems in surveillance enables automated responses, such as pan-tilt-zoom (PTZ) cameras tracking a moving object or drones being dispatched to investigate an alert. For example, in smart cities, surveillance cameras connected to a central control system can detect traffic accidents, automatically alerting emergency services and adjusting traffic lights to clear the congestion.

However, the use of computer vision in surveillance raises privacy concerns, as it enables widespread monitoring of public spaces. Balancing security and privacy is a key challenge, requiring the development of ethical guidelines and technical safeguards (e.g., anonymization of footage) to protect individual rights.

7. Challenges and Future Directions

Despite the significant progress in computer vision, several challenges remain, limiting the performance and applicability of current systems. Addressing these challenges will be crucial for advancing the field and enabling new applications. This section outlines the key challenges and identifies promising future directions.

7.1 Challenges

7.1.1 Robustness to Environmental Variations

Computer vision systems often struggle with variations in lighting, weather, and occlusion, which can degrade performance. For example, image enhancement methods may fail to produce clear images in extreme low-light conditions, and object detectors may misclassify objects in heavy rain or snow. Occlusion, where part of an object is hidden by

another object, is another major challenge, particularly for object tracking and 3D reconstruction.

One reason for this fragility is that many deep learning models are trained on datasets that do not fully capture the diversity of real-world conditions. For example, datasets may be dominated by images captured in good lighting, leading to poor generalization to low-light scenarios. Additionally, models often learn spurious correlations (e.g., associating a particular color with a class) rather than robust features, making them sensitive to changes in the environment.

7.1.2 Computational Complexity

Many state-of-the-art computer vision models, particularly deep learning-based ones, are computationally intensive, requiring powerful GPUs for real-time inference. This limits their deployment on edge devices with limited resources, such as drones, smartphones, and IoT sensors. For example, a large CNN for object detection may require billions of operations per frame, making it impractical for a battery-powered robot.

Model compression techniques, such as pruning, quantization, and knowledge distillation, have been developed to reduce the computational footprint of deep learning models. However, these techniques often involve a trade-off between accuracy and efficiency, with significant compression leading to performance degradation.

7.1.3 Lack of Annotated Data

Deep learning models rely on large amounts of annotated data for training, but annotating data is time-consuming, expensive, and error-prone. This is particularly true for tasks requiring detailed annotations, such as 3D reconstruction (where each point in a point cloud must be labeled) and video action recognition (where annotations must be temporally aligned).

The problem is exacerbated for specialized domains, such as medical imaging, where annotated data is scarce due to privacy concerns and the need for expert knowledge. In such cases, models may overfit to the limited training data, performing poorly on new, unseen data.

7.1.4 Explainability and Trustworthiness

Deep learning models are often described as "black boxes," with their decisions being difficult to interpret. This lack of explainability is a significant barrier to their adoption in safety-critical applications, such as healthcare and autonomous driving, where users need to understand why a model made a particular decision.

Additionally, deep learning models can be vulnerable to adversarial attacks, where small perturbations to the input image cause the model to make incorrect predictions. For example, a stop sign modified with adversarial perturbations may be misclassified as a speed limit sign by an autonomous vehicle, leading to dangerous behavior. Ensuring the trustworthiness of computer vision systems is therefore a critical challenge.

7.2 Future Directions

7.2.1 Explainable AI in Computer Vision

Explainable AI (XAI) aims to develop models that provide clear, human-understandable explanations for their decisions. In computer vision, this could involve highlighting the regions of an image that influenced a model's prediction (e.g., the pixels in a stop sign that led to its classification).

Recent approaches to XAI in computer vision include attention maps, which show where a model focuses its attention, and counterfactual explanations, which describe how the input would need to change to alter the model's prediction. For example, a counterfactual explanation for a misclassified image might indicate that adding a certain feature (e.g., a red light) would cause the model to correctly classify it as a stop sign.

Integrating XAI into computer vision systems will not only increase trust but also facilitate debugging and improvement of models, making them more reliable in critical applications.

7.2.2 Multimodal Perception

Multimodal perception involves combining information from multiple sensors (e.g., cameras, LiDAR, radar, and microphones) to improve perception accuracy and robustness. Each sensor has its strengths: cameras provide rich visual information, LiDAR offers precise depth measurements, radar is robust to weather, and microphones capture audio cues (e.g., a car horn).

Fusing these modalities can compensate for the limitations of individual sensors. For example, fusing camera and LiDAR data in autonomous vehicles can improve depth perception in low-light conditions, where LiDAR remains reliable. Deep learning models for multimodal fusion, such as those using transformers to model cross-modal relationships, are being developed to effectively combine different types of data.

Multimodal perception will be crucial for enabling computer vision systems to operate in diverse and challenging environments, where no single sensor is sufficient.

7.2.3 Edge Computing

Edge computing involves deploying computer vision models on edge devices (e.g., smartphones, drones, IoT sensors) rather than in the cloud, reducing latency and bandwidth usage. This is essential for real-time applications such as autonomous robotics and smart surveillance, where timely processing is critical.

Developing lightweight deep learning models for edge computing is a key research direction. Techniques such as model pruning (removing redundant neurons), quantization (using lower-precision weights), and knowledge distillation (training a small model to mimic a large one) are being used to reduce model size and computational requirements.

Another approach is to design hardware-efficient architectures, such as MobileNet and EfficientNet, which are optimized for deployment on mobile devices. These models use depth-wise separable convolutions and other efficiency-enhancing techniques to reduce computation while maintaining performance.

7.2.4 4D Reconstruction

4D reconstruction extends 3D reconstruction by incorporating the time dimension, enabling the dynamic reconstruction of moving objects and scenes. This involves capturing not only the 3D structure of the scene but also how it changes over time, providing a comprehensive understanding of dynamic events.

4D reconstruction has applications in virtual reality (creating realistic avatars that mimic human movements), robotics (interacting with moving objects), and healthcare (tracking the motion of organs during surgery). Recent approaches to 4D reconstruction use deep learning to model the spatiotemporal dynamics of the scene, with some models predicting future states based on past observations.

Challenges in 4D reconstruction include handling large amounts of data (due to the time dimension) and accurately modeling non-rigid deformations. Future research will focus on developing efficient 4D representation learning methods and improving the accuracy of dynamic reconstructions.

7.2.5 Lifelong Learning

Lifelong learning (also known as continuous learning) enables computer vision systems to learn from new data over time without forgetting previously acquired knowledge. This is in contrast to traditional deep learning models, which often suffer from catastrophic forgetting when trained on new tasks.

Lifelong learning is essential for deploying computer vision systems in real-world environments, where the data distribution may change over time (e.g., new types of objects appearing in a surveillance scene). Techniques for lifelong learning in computer vision include memory replay (periodically retraining on old data), elastic weight consolidation (protecting important weights), and modular architectures (adding new modules for new tasks).

By enabling systems to adapt to new situations

while retaining existing knowledge, lifelong learning will make computer vision more flexible and practical for long-term deployments.

8. Conclusion

Computer vision and image processing have undergone a remarkable transformation in recent years, driven by advances in deep learning and the increasing availability of large datasets and computational resources. This paper has provided a comprehensive overview of the latest developments in image enhancement, 3D reconstruction, motion analysis, and deep learning-based vision, highlighting their integration into perception and control systems.

Image enhancement techniques, both traditional and deep learning-based, have improved the quality of visual data, laying the foundation for reliable subsequent processing. 3D reconstruction methods, ranging from stereo vision to deep learning-based single-image reconstruction, have enabled machines to perceive the spatial structure of the environment. Motion analysis algorithms, including those for detection, tracking, and estimation, have facilitated the understanding of dynamic events. Deep learning has been a unifying force, enhancing the performance of each of these areas and enabling solutions to previously challenging problems.

The integration of these techniques into robotics, autonomous vehicles, and surveillance systems has demonstrated their practical value, enabling intelligent behavior in diverse applications. However, significant challenges remain, including robustness to environmental variations, computational complexity, lack of annotated data, and the need for explainability.

Future research directions, such as explainable AI, multimodal perception, edge computing, 4D reconstruction, and lifelong learning, hold the promise of addressing these challenges and advancing the field further. By developing more robust, efficient, and adaptable computer vision systems, we can unlock new applications and improve existing ones, contributing to the development of intelligent systems that enhance our daily lives.

As computer vision continues to evolve, it will play an increasingly important role in bridging the gap between the physical and digital worlds, enabling machines to perceive, understand, and interact with their environment in ways that were once the stuff of science fiction. The continued collaboration between researchers, engineers, and domain experts will be crucial for realizing this vision and ensuring that computer vision technologies are developed responsibly and ethically.

References

[1] Gonzalez, R. C., & Woods, R. E. (2018). Digital image processing (4th ed.). Pearson.

[2] Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. Proceedings of the IEEE International Conference on Computer Vision, 839-846.

[3] Zhang, K., Zuo, W., Chen, Y., et al. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing, 26(7), 3142-3155.

[4] Wang, X., Yu, K., Wu, S., et al. (2018). ESRGAN: Enhanced super-resolution generative adversarial networks. European Conference on Computer Vision Workshops, 0-0.

[5] Hartley, R., & Zisserman, A. (2003). Multiple view geometry in computer vision (2nd ed.). Cambridge University Press.

[6] Snavely, N., Seitz, S. M., & Szeliski, R. (2006). Photo tourism: Exploring photo collections in 3D. ACM Transactions on Graphics (TOG), 25(3), 835-846.

[7] Chang, J. R., Chen, Y. S. (2018). Pyramid stereo matching network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5410-5418.

[8] Kronander, J., Varol, G., & Kragic, D. (2018). MeshRCNN: 3D mesh reconstruction via region-based CNN. arXiv preprint arXiv:1804.06516.

[9] Zivkovic, Z. (2004). Improved adaptive Gaussian mixture model for background subtraction. Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04), 2, 28-31.

[10]Li, B., Yan, J., Wu, W., et al. (2018). High-performance visual tracking with siamese region proposal network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8971-8980.

[11]Dosovitskiy, A., Fischer, P., Ilg, E., et al. (2015). Flownet: Learning optical flow with convolutional networks. Proceedings of the IEEE International Conference on Computer Vision, 2758-2766.

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

[13]He, K., Gkioxari, G., Dollár, P., et al. (2017). Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 2961-2969.

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

[15]Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6299-6308.

[16]Siegwart, R., Nourbakhsh, I. R., & Scaramuzza, D. (2018). Introduction to autonomous mobile robots (2nd ed.). MIT press.

[17]Urmson, C., Anhalt, J., Bagnell, D., et al. (2008). Autonomous driving in urban environments: Boss and the urban challenge. Journal of Field Robotics, 25(8-9), 425-466.

[18]Chen, C., Chen, C., Wang, Y., et al. (2020). Deep learning for visual surveillance: A survey. IEEE Transactions on Circuits and Systems for Video Technology, 30(11), 4046-4067.

[19]Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[20]LeCun, Y., Bottou, L., Bengio, Y., et al. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[21]Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[22]Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, 234-241.

[23]Vaswani, A., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30, 5998-6008.

[24]Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3354-3361.

[25]Chen, L. C., Papandreou, G., Kokkinos, I., et al. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834-848.