A Survey on Computer Vision, Video Machine Learning, 3D Reconstruction, Video Analysis, and Visual Perception

Abstract

This survey paper explores the interdisciplinary realm of computer vision, video analysis, machine learning, 3D reconstruction, and visual perception, highlighting their transformative impact across various domains. The integration of these technologies fosters advancements in healthcare, environmental monitoring, surveillance, autonomous vehicles, and augmented reality. Key developments include the use of deep learning models for enhanced diagnostic accuracy, the potential of Neural Radiance Fields (NeRFs) in robotics, and the application of super-resolution processes in traffic scenarios. The survey underscores the importance of extensive datasets, such as the Oxford Multimotion Dataset, in advancing multimotion estimation algorithms. It also emphasizes the role of biologically inspired data augmentation methods in improving model robustness. Despite these advancements, challenges remain, including the need for more efficient real-time processing, improved depth estimation techniques, and the integration of multimodal data. The survey advocates for interdisciplinary collaboration to address these challenges and leverage the strengths of integrated approaches, ultimately enhancing machine perception and interaction capabilities. By aligning technological advancements with societal values, the field of computer vision is poised to make significant strides in understanding and interacting with the world.

1 Introduction

1.1 Interdisciplinary Nature and Core Technologies

Computer vision is fundamentally interdisciplinary, integrating insights from machine learning, cognitive science, engineering, and medicine to enhance visual understanding. The shift towards vision-language intelligence, as noted by Li et al., signifies a transition from single-modality processing to a multi-modal comprehension, broadening the field's interdisciplinary scope [1]. This evolution is particularly evident in the automated analysis and classification of fine art collections, where Kvak et al. emphasize the challenges of domain specificity and the intricacies of art movements, necessitating tailored interdisciplinary approaches [2].

In medical imaging, the introduction of transformers, discussed by Henry et al., showcases their ability to capture long-range dependencies and contextual information, overcoming the limitations of traditional convolutional neural networks (CNNs) and integrating advanced machine learning techniques in diagnostics [3]. Esteva et al. further highlight the expansive role of deep learning across various medical fields, including cardiology, pathology, dermatology, and ophthalmology, significantly improving diagnostic accuracy and efficiency [4].

Advancements in image segmentation are also notable, with Minaee et al. detailing the impact of deep learning models, such as fully convolutional networks and encoder-decoder architectures, on segmentation accuracy across diverse applications [5]. Additionally, Wang et al. propose biologically

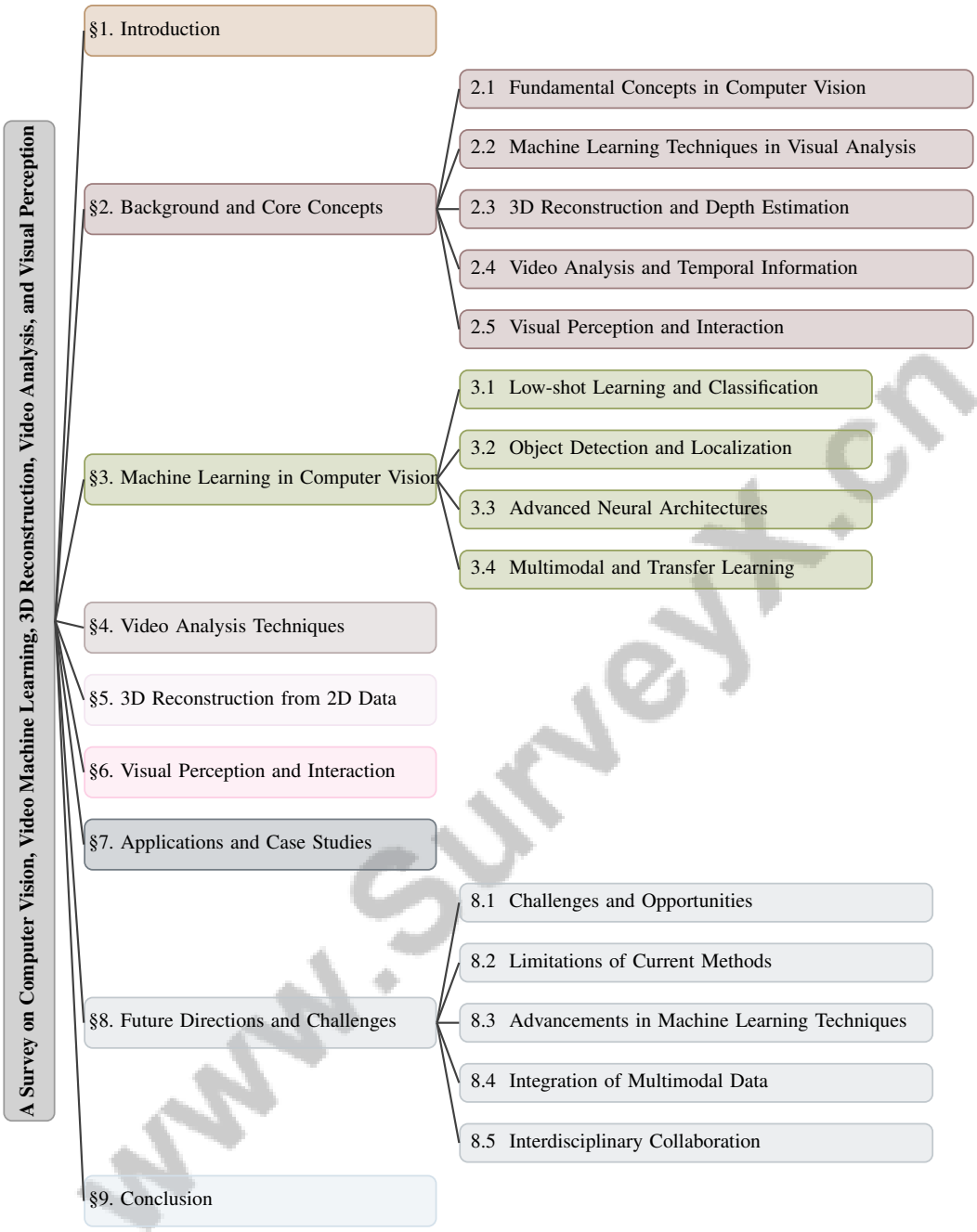

Figure 1: chapter structure

plausible augmentations for self-supervised learning, merging biological insights with machine learning to foster more robust models [6].

These interdisciplinary efforts, spanning numerous fields and technologies, drive the development of computer vision, enabling sophisticated visual recognition and analysis. The interplay between biological and artificial vision systems, as Nixon et al. discuss, addresses existing knowledge gaps and promotes advancements in both human and computer vision [7]. Such collaborations are essential for achieving new milestones in visual perception and interaction.

1.2 Significance and Applications

The transformative impact of computer vision, video analysis, machine learning, 3D reconstruction, and visual perception is evident across various fields, driving advancements and innovations. In healthcare, these technologies have significantly improved medical imaging and diagnostics, enhancing disease detection and treatment. The integration of deep learning models, particularly transformers, into medical image analysis has shown promise in classification, segmentation, registration, and reconstruction, marking a shift towards more effective medical solutions. During the COVID-19 pandemic, computer vision techniques were pivotal in disease diagnosis, prevention, and treatment, underscoring their critical role in healthcare [8].

In robotics, Neural Radiance Fields (NeRFs) have emerged as a vital tool for rendering 3D environments, facilitating scene understanding, navigation, manipulation, and interaction. These advancements demonstrate the potential of NeRFs to enhance robotic capabilities in complex environments. Furthermore, computer vision tools have proven significant in the early detection of Autism Spectrum Disorder (ASD), offering quicker and more accessible assessments [9].

In environmental monitoring and smart agriculture, computer vision and machine learning enhance the accuracy and efficiency of monitoring plant traits, aiding in crop monitoring and disease detection [10]. The evaluation of global perceptual similarity metrics is crucial for various applications, particularly in analyzing social media information flow, as highlighted by Vallez et al. [11]. Attention mechanisms have also significantly improved performance in visual tasks, including image classification, object detection, and video understanding, further emphasizing their importance [12].

The development of vision-language (VL) intelligence, as reviewed by Li et al., reflects the evolution from task-specific approaches to vision-language pre-training (VLP) methods using large-scale weakly-labeled data, showcasing the expanding capabilities of computer vision technologies [1]. Additionally, Minaee et al. provide a thorough review of deep learning-based image segmentation methods, addressing effectiveness and challenges, further underscoring the significance of these technologies in advancing image analysis [5].

These applications illustrate the broad impact of computer vision and related technologies across various domains, driving innovations that enhance both practical applications and theoretical understanding. The interdisciplinary nature of these technologies, as discussed by Nixon et al., highlights the ongoing advancements and collaborations necessary for achieving new milestones in visual perception and interaction [7].

1.3 Structure of the Survey

This survey is structured to provide a comprehensive exploration of the interdisciplinary field of computer vision, video analysis, machine learning, 3D reconstruction, and visual perception. The organization is informed by benchmark surveys and studies in the field, each offering distinct perspectives on complex topics such as dataset documentation, image perceptual similarity metrics, scene graph generation, data selection methods, and small object recognition challenges. These references collectively highlight key insights into dataset creation practices, visual information evaluation, and methodologies for enhancing understanding and reasoning in visual tasks, guiding the development of a coherent framework for addressing multifaceted issues in machine learning and computer vision [13, 14, 15, 11, 16].

The survey begins with an **Introduction** that discusses the interdisciplinary nature and core technologies, followed by an examination of the significance and diverse applications of these technologies. This section also outlines the overall structure of the survey.

Section 2: Background and Core Concepts delves into fundamental concepts and technologies, offering an overview of the interaction between components that facilitate advanced visual understanding. Discussions encompass foundational concepts in computer vision, advanced machine learning techniques for visual analysis, methodologies for 3D reconstruction, comprehensive approaches to video analysis, and modeling visual perception, highlighting applications in autonomous navigation, security surveillance, medical diagnostics, and automated artwork classification [2, 17, 18].

Section 3: Machine Learning in Computer Vision focuses on the role of machine learning in enhancing computer vision capabilities. This section covers low-shot learning, object detection,

advanced neural architectures, and multimodal and transfer learning techniques, paralleling structured approaches in surveys by Liu et al. [19] and Esteva et al. [4].

Section 4: Video Analysis Techniques examines methodologies in video analysis, emphasizing temporal information and addressing challenges in object detection, anomaly detection, and real-time analysis, similar to the organization seen in Wang et al.'s survey [20].

Section 5: 3D Reconstruction from 2D Data discusses the creation of 3D models from 2D images or video frames, including techniques, applications, and challenges, reflecting the structured categorization approach used by O'Connor et al. [21].

Section 6: Visual Perception and Interaction explores how visual perception is modeled and implemented in computer vision systems, focusing on interaction and perception enhancement, mirroring the detailed organization found in surveys by Frey et al. [22].

Section 7: Applications and Case Studies offers a comprehensive exploration of real-world applications and case studies illustrating the transformative effects of advanced technologies across diverse sectors. These include significant advancements in healthcare through medical imaging and predictive analytics, enhanced environmental monitoring via robust background subtraction techniques, improved surveillance systems utilizing computer vision for object detection, the development of autonomous vehicles leveraging machine learning for navigation, and the integration of augmented reality in various applications. This section highlights the current state of these technologies while discussing ongoing challenges and future directions [23, 18, 24].

Section 8: Future Directions and Challenges identifies current limitations and challenges, discussing potential future research directions and technological advancements, informed by the forward-looking structure seen in Fan et al.'s work [25].

The survey concludes with a **Conclusion** that reinforces the significance of integrating computer vision, video analysis, machine learning, 3D reconstruction, and visual perception, drawing on the comprehensive methodologies of Chang et al. [13] and Tan et al. [26].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Fundamental Concepts in Computer Vision

Computer vision, a multidisciplinary field, aims to enable machines to interpret visual data through foundational concepts. Central to this is object identification, which involves detecting and classifying instances from various categories within images, crucial for applications like autonomous vehicles where reliable perception is vital [27]. Visual object tracking, which maintains target tracking in video sequences despite occlusion and background clutter, remains a significant challenge [9].

Vision-language tasks necessitate structured AI models that handle multimodal data, enhancing machines' abilities to process complex scenes for tasks like image captioning [1]. Mapping color images to scene graphs underscores the importance of structured data representation [28]. Semantic segmentation and image analysis improve emergency response by facilitating detailed scene understanding in disaster scenarios [11]. Traditional segmentation methods often fall short, but deep learning models significantly advance semantic and instance segmentation tasks [5].

Attention mechanisms enhance visual task performance by focusing on salient regions, mimicking human attention [12]. Robust scene recognition techniques are needed due to the failure of models to generalize across different camera data [29]. Unsupervised learning approaches address identifying consistent interest points under transformations [30].

The multimotion estimation problem aims to estimate every motion in a scene, including camera motion, by segmenting these motions [31]. These foundational concepts drive sophisticated algorithm development, fostering advancements across applications. Continuous exploration of these concepts is essential for overcoming challenges and enhancing computer vision systems, bridging the gap between human perception and algorithms.

2.2 Machine Learning Techniques in Visual Analysis

Machine learning significantly advances visual analysis, improving task accuracy and efficiency. Convolutional neural networks (CNNs) are foundational, especially in object detection and classification. Kvak et al. demonstrate CNNs' versatility by using them for classifying visual artworks [2]. García et al. highlight super-resolution techniques that enhance image quality and improve small object detection [27]. Modeling human-like perception remains challenging, necessitating sophisticated models [7].

In motion analysis, Judd et al. highlight the challenge of estimating multiple motions, emphasizing the need for advanced models [31]. Biologically inspired augmentations, such as the CMS method, enhance visual representation learning by mimicking human processing [6].

Machine learning's influence is evident in diverse fields, including art classification and computer vision, where techniques like residual neural networks facilitate automated analysis [15, 2]. Continuous refinement of these techniques is essential for expanding model capabilities and ensuring robust performance in complex environments.

2.3 3D Reconstruction and Depth Estimation

3D reconstruction and depth estimation transform 2D data into 3D models, essential for applications like autonomous navigation. Mittal et al. classify Neural Radiance Fields (NeRFs), crucial for rendering 3D environments [32]. Torrente et al. propose an extended Hough transform method to enhance 3D representations by localizing feature curves [33].

Integrating multispectral and hyperspectral data enhances model accuracy, as highlighted by Sun et al. [34]. Rodriguez et al. discuss using polarization information for depth estimation, providing insights into material properties [35]. Xu et al. explore knowledge transfer techniques to bridge 2D and 3D representations [36].

Outdoor depth estimation faces challenges due to lighting conditions and scene complexities. Vyas et al. propose methods to enhance accuracy in outdoor environments [37]. Advancements in neural network architectures improve material classification tasks integral to 3D reconstruction [38].

The field is characterized by diverse techniques overcoming 2D to 3D translation challenges. Innovations in feature extraction, multispectral data integration, and neural architectures enhance 3D reconstruction systems, facilitating applications in agriculture, monitoring, and heritage [39, 40, 32, 34].

2.4 Video Analysis and Temporal Information

Temporal information is crucial in video analysis, capturing dynamic patterns essential for scene understanding. Temporal dynamics facilitate tasks like background subtraction, where environmental changes impact detection accuracy [24]. Detecting moving objects is complicated by dynamic backgrounds, requiring advanced methodologies [41].

The CNN-RNN Video Classifier processes features with recurrent networks to analyze temporal patterns, crucial for content classification [42]. This is vital in classifying children's videos for inappropriate content [17]. Visual tracking relies on temporal coherence for accurate tracking, despite occlusions [43]. Jethava et al. address inference challenges in video data exhibiting temporal coherence [44].

Temporal information is key in video editing analysis, providing insights into narrative flow [22]. NeRF limitations in handling lighting and material properties emphasize challenges in integrating spatial and temporal data [45].

Advancing video analysis systems requires addressing temporal modeling, computational efficiency, and data integration challenges. This includes scaling analytics for large deployments, optimizing inference, and developing lightweight models for real-time processing [46, 42, 47, 48].

2.5 Visual Perception and Interaction

Visual perception is critical in computer vision systems, influencing machine interaction with environments. Ye et al. highlight the impact of cultural backgrounds on perception, necessitating models that

accommodate diverse frameworks [49]. In multimodal learning, attention-related processes enhance interaction capabilities [50].

Visual perception integration in VQA systems demonstrates the complexity of analyzing multimodal images, requiring sophisticated understanding [51]. Pang et al. propose methods to enhance illusory contour perception, aligning network behavior with human perception [52]. Kancharala et al. provide insights into brain responses to visual stimuli, aiding model development [53].

Agarwal et al. highlight biases in models like CLIP, emphasizing fair perception across applications [54]. The absence of a unified framework for evaluating explainability presents challenges in visual perception modeling [55]. Balayn et al. emphasize systematic approaches in modeling perception, focusing on process relationships [56].

Recent studies underscore visual perception's role in facilitating effective interactions between systems and environments. Advancements in attention mechanisms and diverse datasets enhance system accuracy and applicability across domains, including navigation and healthcare [12, 49, 18]. By integrating cultural, physiological, and cognitive insights, researchers develop robust models that emulate human perception, enhancing interaction capabilities.

3 Machine Learning in Computer Vision

3.1 Low-shot Learning and Classification

Low-shot learning addresses the challenge of training models with minimal labeled data, crucial in domains like defect classification in additive manufacturing, where Liu et al. demonstrated the effectiveness of combining CNNs with transfer learning to reduce labeling needs [57]. The Kornia library by Riba et al. enhances model training in low-shot scenarios through differentiable image processing functions [58]. In visual tracking, Gao et al.'s cascaded approach utilizes subnetworks for target localization and classification, improving accuracy with limited data [43].

As illustrated in Figure 2, this figure depicts the hierarchical categorization of low-shot learning and classification techniques, applications, and challenges. Key methods include CNN-DC for defect classification, Kornia for differentiable vision tasks, and CSN for visual tracking. Applications span from Explainable Boosting Machines (EBMs), which enhance interpretability in low-shot learning by providing clear model predictions [59], to Rithish et al.'s MISR, which improves road safety monitoring through real-time road irregularity detection [60]. Challenges and innovations are also addressed, including attention mechanisms that improve learning by focusing on salient features, as discussed by Guo et al. [12], and the lottery ticket hypothesis explored by Chen et al., which shows that smaller subnetworks from pre-trained models maintain task performance [61].

Furthermore, Savinov et al.'s quad-networks optimize interest point ranking for repeatability, benefiting low-shot learning [30]. Ben-Yosef et al.'s FLIMIM model enhances generalization by focusing on local image details [28], while García et al.'s SRCNN improves small object detection by enhancing image resolution [27].

Li et al. emphasize the importance of language-aligned visual representations and large-scale weakly-labeled datasets for zero or few-shot learning [1]. These methodologies highlight hybrid models, iterative approaches, and multimodal data integration as critical for enhancing low-shot learning, improving predictive performance and model calibration [62, 15, 54, 16].

3.2 Object Detection and Localization

Object detection and localization, crucial for applications like autonomous driving and surveillance, involve identifying objects and determining their spatial positions within images. Deep learning, especially CNNs, has revolutionized object detection by providing robust feature extraction and classification frameworks. The integration of CNNs with RNNs enhances video content classification by analyzing temporal patterns, improving detection and localization in dynamic scenes [42].

Faster R-CNN and YOLO are leading architectures in this domain. Faster R-CNN is noted for its high accuracy, achieving significant detection rates in camera trap images, while YOLO provides a real-time framework balancing accuracy and speed, with YOLOv11 advancing object detection, instance segmentation, and pose estimation [64]. Oriented bounding boxes within YOLO architectures

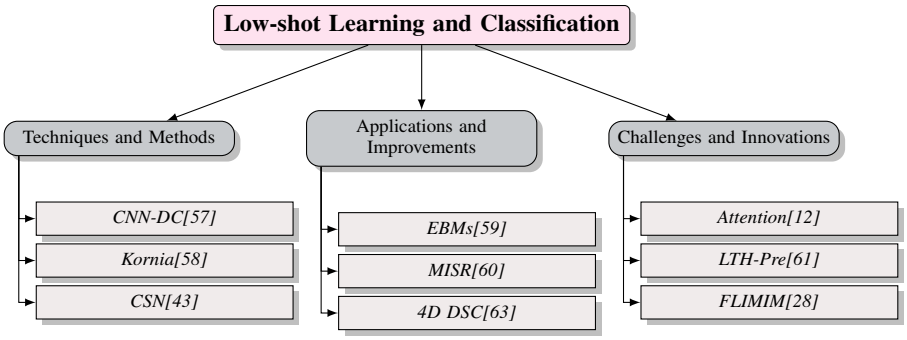

Figure 2: This figure illustrates the hierarchical categorization of low-shot learning and classification techniques, applications, and challenges. Key methods include CNN-DC for defect classification, Kornia for differentiable vision tasks, and CSN for visual tracking. Applications span from EBMs enhancing interpretability to MISR improving road safety monitoring. Challenges and innovations are addressed by attention mechanisms, LTH-Pre for efficient model use, and FLIMIM for detailed image interpretation.

improve detection accuracy in complex environments [65]. Despite these advancements, traditional template matching methods face challenges with illumination, viewpoint, and occlusion changes, requiring continuous innovation for more adaptive models [66].

3.3 Advanced Neural Architectures

Advancements in neural architectures have significantly enhanced computer vision capabilities, enabling complex tasks such as classification and detection. Neural Radiance Fields (NeRFs) in robotics, as surveyed by Wang et al., highlight systematic approaches to leverage neural architectures effectively [45]. Ferreri et al.'s Tran-Adapt method uses self-supervised learning to enhance RGB and depth modality translation, boosting neural network representational power [29]. Chen et al. explore the lottery ticket hypothesis, demonstrating that subnetworks within pre-trained models can match full model performance, reducing computational overhead [61].

Su et al. introduce LBP-inspired CNN modules, combining LBP's efficiency with CNNs' power, enhancing architectural efficiency and classification accuracy [67]. ResNet50V2, noted by Kvak et al., uses residual connections to improve training and feature extraction [2]. Ben-Yosef et al.'s FLIMIM emphasizes complex spatial relations, often overlooked by state-of-the-art models, enhancing image interpretation [28]. These advancements reflect a significant evolution in neural architectures, driven by computational strategies, diverse datasets, and domain-specific adaptations, including ethical considerations and cognitive mechanisms [68, 15, 42].

3.4 Multimodal and Transfer Learning

Multimodal data integration and transfer learning are pivotal in advancing computer vision, allowing models to utilize diverse data sources and pre-trained models for enhanced performance. Miao et al. exemplify this with a Bayesian exploration method integrating probabilistic model ensembles and pre-trained models [62]. He et al.'s Tracking-by-Animation framework integrates tracking and reconstruction for improved multimodal data handling [69]. Inoshita et al. propose a gating network to optimally fuse trained models from different domains, enhancing object detection [70]. Ye et al. advocate for multilingual models to represent diverse perceptual patterns [49].

Mutahar et al.'s TreeICE framework uses Non-negative Concept Activation Vectors (NCAVs) for concept extraction and decision trees for explanations, integrating interpretability into multimodal learning [71]. Kästner et al. present an AR-based human assistance system combining deep neural networks with augmented reality, guiding users through complex tasks [72].

These methodologies underscore the transformative impact of multimodal and transfer learning in computer vision. By integrating diverse data sources and leveraging advanced pre-trained models like CLIP, researchers can develop robust systems addressing complex visual tasks across domains. This approach enhances model performance through self-supervised learning and multilingual datasets,

acknowledging cultural and contextual nuances in human perception, leading to improved outcomes in applications like medical imaging and automated art analysis [15, 73, 54, 49, 2].

4 Video Analysis Techniques

Category	Feature	Method
Object Detection and Tracking	Pattern Recognition	Q-Nets[30], FLIMIM[28]
	Focus and Efficiency Improvement	CVASD[9]
	Resolution and Detail Enhancement	SRCNN[27]
Anomaly Detection in Video Sequences	Task-Oriented Learning	SSMTL-AD[74], TBA[69]
	Collaborative Video Processing	CCA[46]
	Robustness and Vulnerability	ARA[75], AVIA[76], O-YOLO[65]
	Model Transparency	EBM[59]
Real-time Video Analysis	Adaptive Model Strategies	TMF-GN[70], SDL[77]
	Efficient Processing Techniques	CV[47], COVA[48], CIP[78]

Table 1: This table provides a comprehensive overview of various techniques employed in video analysis, categorized into object detection and tracking, anomaly detection in video sequences, and real-time video analysis. Each category highlights specific features and methods, showcasing advancements in pattern recognition, task-oriented learning, and adaptive model strategies. The table serves as a valuable resource for understanding the current state of video analysis methodologies and their applications in computer vision.

The analysis of video data plays a pivotal role in computer vision, involving techniques that extract significant insights from dynamic visual content. This section delves into foundational video analysis techniques crucial for applications such as surveillance and autonomous systems. Object detection and tracking are particularly noteworthy, enabling the identification and continuous monitoring of entities within video sequences. Table 4 presents a detailed summary of the methodologies utilized in video analysis, emphasizing their relevance across different domains and applications. The following subsection explores the complexities associated with these tasks, emphasizing their importance and challenges in practical implementations.

4.1 Object Detection and Tracking

Method Name	Techniques Employed	Practical Applications	Challenges Addressed
Q-Nets[30]	Neural Network	Image Detection	Human Labeling
SRCNN[27]	Super-resolution Processes	Traffic Surveillance Cameras	Resolution Loss
FLIMIM[28]	Bottom-up Processing	Autonomous Navigation	Dynamic Backgrounds
CVASD[9]	Video Analysis	Infant Behavior Tracking	Specialized Training

Table 2: This table presents a comparative analysis of various object detection and tracking methods, detailing the techniques employed, their practical applications, and the specific challenges they address. The methods include Q-Nets, SRCNN, FLIMIM, and CVASD, each contributing to advancements in areas such as image detection, traffic surveillance, autonomous navigation, and infant behavior tracking.

Object detection and tracking are vital tasks in computer vision, with applications spanning surveillance, autonomous navigation, and wildlife monitoring. These tasks involve identifying objects within video frames and consistently tracking their identities across successive frames, facing challenges like occlusions, dynamic backgrounds, and fluctuating lighting conditions. Techniques such as background subtraction and saliency detection are employed to separate moving objects from their backgrounds, ensuring accuracy and robustness in real-world scenarios [79, 24, 80, 42]. Recent advancements in neural architectures and deep learning have significantly improved detection accuracy and speed, enabling real-time applications.

As illustrated in Figure 3, the hierarchical categorization of object detection and tracking highlights key techniques, practical applications, and recent advancements in the field. This figure provides an overview of methods such as background subtraction, saliency detection, and attention mechanisms, alongside practical applications in small vehicle detection and infant behavior tracking, and recent advancements including neural architectures and spatio-temporal correlations. Additionally, Table 2 provides a detailed examination of key object detection and tracking methods, highlighting the techniques used, their practical applications, and the challenges they address in the context of computer vision.

Attention mechanisms enhance object detection and tracking by focusing on relevant temporal information, thereby improving task efficiency [12]. The integration of sensory evidence with cognitive expectations further boosts tracking accuracy, as cognitive models enhance the perception of moving objects in dynamic scenes [81]. Additionally, Savinov et al. explore unsupervised learning approaches to identify consistent features across varying conditions, crucial for robust object tracking [30].

In practical applications, detecting small vehicles in traffic videos is complicated by background clutter and resolution loss, as noted by García et al. [27]. Techniques addressing these challenges are vital for enhancing detection performance in complex environments. The Oxford Multimotion Dataset, introduced by Judd et al., provides a comprehensive collection of multimotion data, aiding in the development of algorithms capable of handling both static and dynamic scenes [31].

Ben-Yosef et al. focus on recognizing and interpreting local regions in images by identifying semantic features and their spatial relations, enhancing object detection capabilities [28]. Furthermore, Hashemi et al. demonstrate the use of computer vision tools for detecting and tracking behavioral markers in video recordings of infants, showcasing the applicability of these techniques in assessing visual attention and motor patterns [9]. This application highlights the broader impact of object detection and tracking technologies across diverse fields, including healthcare.

Advancements in object detection and tracking techniques are crucial for progressing computer vision applications. By employing sophisticated methodologies and improving computational efficiency, researchers can enhance the accuracy and scalability of video analysis systems, particularly in large-scale deployments with numerous camera feeds. Leveraging spatio-temporal correlations, utilizing neural compression techniques for efficient processing of longer videos, and implementing automated model specialization for edge devices contribute to reducing computational costs while maintaining or improving inference accuracy, making real-time video analytics feasible in applications such as smart surveillance and intelligent monitoring [47, 46, 82, 42, 48].

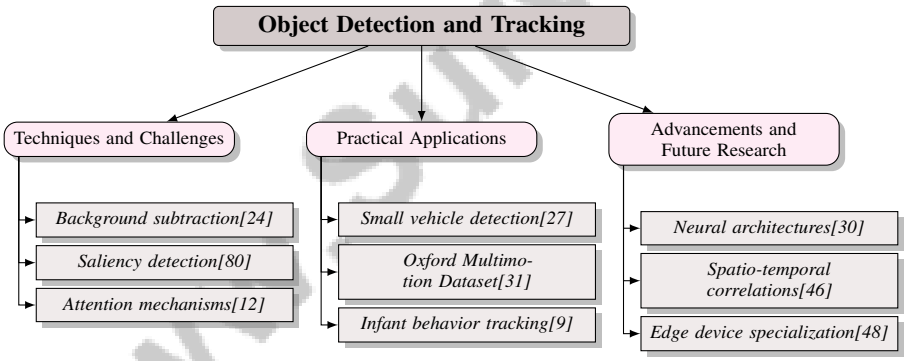

Figure 3: This figure illustrates the hierarchical categorization of object detection and tracking, highlighting key techniques, practical applications, and recent advancements. It provides an overview of the methods such as background subtraction, saliency detection, and attention mechanisms, practical applications in small vehicle detection and infant behavior tracking, and recent advancements including neural architectures and spatio-temporal correlations.

4.2 Anomaly Detection in Video Sequences

Anomaly detection in video sequences is a fundamental task in computer vision, aimed at identifying events that deviate from expected patterns, such as unusual activities in surveillance footage or unexpected movements in autonomous driving scenarios. A primary challenge is the limited availability of diverse anomalous events for training, necessitating models trained predominantly on normal data, which may restrict their generalization to unseen anomalies [74]. Table 3 presents a comprehensive comparison of different methods utilized in anomaly detection in video sequences, detailing their approaches, addressed challenges, and performance metrics.

Innovative methodologies have emerged to tackle these challenges. The ARA method discussed by Li et al. effectively highlights vulnerabilities in video object segmentation (VOS) models, crucial for improving anomaly detection capabilities [75]. Newson et al.'s work on video inpainting emphasizes

Method Name	Methodological Approaches	Challenges Addressed	Performance Evaluation
SSMTL-AD[74]	Self-supervised Learning	Limited Anomalous Data	Auc Metrics
ARA[75]	Adversarial Region Attack	Adversarial Attacks Vulnerability	Region Similarity J
AVIA[76]	Video Inpainting	Occluded Content	Visual Quality
TBA[69]	Tracking-by-Animation	Occlusions, Appearances	Multi-Object Tracking
CCA[46]	Video Inpainting	Limited Anomalous Event	Recall, Precision
O-YOLO[65]	Video Inpainting	Limited Anomalous Event	Mean Average Precision
EBM[59]	Explainable Boosting Machines	Limited Anomalous Event	Accuracy, Precision, Recall

Table 3: Table illustrating various methodologies employed in anomaly detection within video sequences, highlighting their distinct methodological approaches, the specific challenges they address, and their respective performance evaluation metrics. The table provides a comparative overview of methods such as Self-supervised Learning, Adversarial Region Attack, and Explainable Boosting Machines, among others, to enhance understanding of their contributions to the field.

the importance of detecting and filling occluded regions in video frames without manual tracking, essential for maintaining scene continuity during anomalies [76].

Integrating object tracking and reconstruction processes, as proposed in the Tracking-by-Animation (TBA) framework by He et al., facilitates efficient end-to-end learning without labeled data, underscoring the potential of combining multiple video analysis tasks to enhance anomaly detection [69]. Jain et al. advocate for analyzing video feeds collectively rather than independently, utilizing correlations to reduce processing loads and enhance detection accuracy, particularly beneficial in large-scale surveillance systems [46].

Experiments by Judd et al. with the Oxford Multimotion Dataset highlight challenges posed by occlusion and complexity, demonstrating varying success rates in accurately estimating motions critical for detecting anomalies in dynamic environments [31]. Additionally, Li et al. evaluate environmental conditions, emphasizing improved detection accuracy for small objects and effective functioning in diverse conditions, crucial for identifying subtle anomalies [65].

The use of Explainable Boosting Machines (EBMs) by Schug et al. underscores the importance of interpretability in anomaly detection, where models are evaluated using accuracy, precision, and recall metrics, providing insights into performance with limited data [59]. Ferreri et al. provide a comprehensive framework for evaluating multi-modal scene recognition models, contributing valuable insights into the domain adaptation problem relevant to anomaly detection [29].

Collectively, these methodologies advance the field of anomaly detection in video sequences by incorporating innovative learning frameworks, leveraging pre-trained models, and exploring novel data modalities. The ongoing development of techniques for object extraction and performance evaluation is crucial for enhancing the robustness and reliability of computer vision systems, particularly in detecting anomalies across various applications. Addressing challenges related to data complexity and ensuring behavioral consistency in intelligent services is emphasized by recent studies, highlighting the need for rigorous evaluation frameworks and the identification of evolution risks in AI-driven solutions [40, 83].

4.3 Real-time Video Analysis

Real-time video analysis is a critical domain in computer vision, facing challenges such as efficiently processing vast amounts of data, maintaining high accuracy, and operating under constrained computational resources. Neural compression frameworks like Compressed Vision enable direct processing of long video sequences without decompression, significantly reducing memory requirements and enhancing training efficiency [47].

The COVA framework exemplifies improvements in real-time performance by optimizing the balance between accuracy and resource usage, achieving an average accuracy increase of 21

Dynamic model fusion techniques, as proposed by Inoshita et al., provide adaptability by adjusting the contributions of different models in real-time, essential for maintaining performance in rapidly changing environments [70]. The Structured Differential Learning (SDL) approach explored by Connell et al. combines offline video data with real-time in-car tests to enhance video analysis performance, highlighting the significance of integrating offline and online learning paradigms [77].

Advancements in object detection technology, such as YOLOv11, yield significant improvements in accuracy and efficiency while reducing parameter counts, making it suitable for various real-time applications [64]. PhyCV algorithms demonstrate real-time operational capabilities on edge devices, achieving efficient performance in tasks like edge detection and low-light enhancement, vital for deploying video analysis systems in resource-constrained environments [84].

Future research should focus on enhancing real-time rendering capabilities, addressing scalability issues, and exploring novel applications of Neural Radiance Fields (NeRFs) in augmented and virtual reality [32]. The method proposed by Vanhoorick et al. effectively reveals occlusions and performs various video understanding tasks without architectural changes, demonstrating potential improvements in real-time video analysis [63].

Addressing the challenges of real-time video analysis involves leveraging neural compression, dynamic model fusion, advanced object detection frameworks, and efficient data processing techniques. These advancements utilize spatio-temporal correlations among camera feeds and integrate deep learning techniques, facilitating the development of more efficient, robust, and scalable real-time video analysis systems capable of handling large-scale camera deployments without a proportional increase in computational costs [46, 42, 47].

Feature	Object Detection and Tracking	Anomaly Detection in Video Sequences	Real-time Video Analysis
Challenges Addressed	Occlusions, Dynamic Backgrounds	Limited Anomalous Data	Computational Constraints
Techniques Used	Background Subtraction, Saliency	Unsupervised Learning, Inpainting	Neural Compression, Fusion
Applications	Surveillance, Navigation	Surveillance, Autonomous Driving	Embedded Systems, Edge Devices

Table 4: This table provides a comprehensive comparison of methodologies employed in video analysis, focusing on object detection and tracking, anomaly detection in video sequences, and real-time video analysis. It highlights the challenges addressed, techniques used, and applications for each method, offering insights into their effectiveness and applicability in various domains.

5 3D Reconstruction from 2D Data

5.1 Techniques and Methodologies

Converting two-dimensional data into three-dimensional models involves sophisticated techniques addressing challenges in depth perception and spatial representation. Stereo vision, a fundamental approach, estimates depth by analyzing image disparities from different viewpoints, enhanced by optical convolution methods integrated with Convolutional Neural Networks (CNNs) for improved outdoor depth estimation [85, 39, 37, 21]. Structure from Motion (SfM) reconstructs 3D structures by analyzing object motion across image sequences, benefiting from innovative convolutional kernels that enhance spatial accuracy. Depth estimation techniques simulate parallax effects, enriching dynamic 3D models from static images. Frameworks like PyD-Net with Mask R-CNN advance the ability to represent spatial relationships and occlusions [86, 37, 63].

This is illustrated in Figure 4, which depicts key techniques in 3D reconstruction, focusing on stereo vision and structure from motion, event-based camera methods, and the integration of AI and synthetic data. The figure highlights the role of stereo vision and structure from motion in depth estimation, the use of event-based cameras for continuous 3D reconstruction, and the enhancement of modeling through AI and synthetic data.

Event-based cameras offer a novel approach to 3D reconstruction, demonstrated by the EvAC3D method, which processes continuous event streams to refine 3D meshes [87]. This method showcases continuous reconstruction potential through Apparent Contour Events. Advanced methodologies, including feature point projection onto optimal planes and using generalized Hough transforms for curve approximation, are crucial for recognizing 3D feature curves. Polarimetric applications enhance image segmentation, surface normal estimation, and pose estimation through polarization information, benefiting robotics, underwater navigation, and augmented reality, despite challenges in processing algorithms and standards [35, 2, 32, 65].

Artificial intelligence integration in 3D modeling boosts accuracy and efficiency. AI-generated synthetic images enrich training datasets, enhancing 3D reconstruction techniques. Fractal geometry in generating synthetic video datasets improves 3D modeling algorithms, enabling automatic creation of diverse video clips, thus enhancing neural model pre-training for action recognition and other tasks

[46, 76, 47, 88]. The Myriad X architecture supports real-time performance for complex AI tasks, suitable for scenarios requiring high computational power with low energy consumption, such as satellite pose estimation [89, 90, 91].

These advancements significantly enhance 3D reconstruction, enabling intricate 3D model creation from 2D data. Transferring pretrained 2D model architectures to 3D point-cloud understanding shows potential for improving classification performance and training efficiency, achieving up to 10% accuracy gains in few-shot scenarios and accelerating training by over 11 times. Integrating spectral information into 3D models enriches physical property representation, facilitating applications in smart agriculture and environmental monitoring. These innovations bolster 3D reconstruction accuracy and efficiency, expanding applicability across various fields [34, 40, 36]. By leveraging advancements in stereo vision, structure from motion, AI integration, and hardware innovations, researchers continue to enhance realistic and dynamic three-dimensional representations.

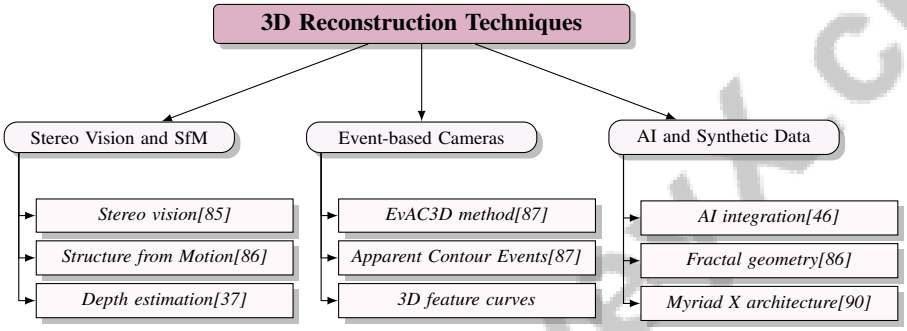

Figure 4: This figure illustrates key techniques in 3D reconstruction, focusing on stereo vision and structure from motion, event-based camera methods, and the integration of AI and synthetic data. It highlights the role of stereo vision and structure from motion in depth estimation, the use of event-based cameras for continuous 3D reconstruction, and the enhancement of modeling through AI and synthetic data.

5.2 Applications and Case Studies

3D reconstruction technologies have diverse applications, impacting scientific research and practical implementations. In geometric analysis, Torrente et al.'s method effectively identifies feature curves, offering improvements in robustness and flexibility over traditional techniques, crucial for precise analysis in computer-aided design and manufacturing [33]. Wu et al. demonstrated the feasibility of 3D reconstruction from public webcams, enabling large-scale environmental monitoring and urban planning by adapting existing methods for convincing results on camera poses and scene structures [92].

In object recognition, Thai et al. introduced a dataset with over 1.3 million images from 55 ShapeNet-Core.v2 categories, invaluable for training and evaluating 3D reconstruction algorithms in complex environments [93]. Event-based cameras, evaluated by Wang et al. using the MOEC-3D dataset, demonstrate high-fidelity 3D mesh reconstruction from continuous event streams, promising for dynamic environments like robotics and autonomous navigation [87].

These case studies highlight 3D reconstruction's diverse applications, enhancing industrial precision and enabling innovative environmental monitoring and autonomous systems solutions. As methodologies advance, the field's scope and impact are expected to broaden significantly, paving the way for applications in augmented reality, robotics, and environmental monitoring. Developments in Neural Radiance Fields (NeRFs) have spurred over 1,000 related preprints, enhancing complex 3D environment modeling. Breakthroughs in predicting 3D shapes from single images and integrating spectral information expand technology potential, with implications for smart agriculture, geological exploration, and cultural heritage preservation [32, 93, 34].

6 Visual Perception and Interaction

6.1 Modeling Visual Perception in Computer Vision

Modeling visual perception in computer vision aims to emulate human perceptual processes, enhancing machines' ability to interpret visual data. Ye et al. emphasize the significance of incorporating cultural and linguistic diversity in visual perception models, highlighting language as a cultural proxy influencing perception [49]. Pang et al. augment deep feedforward convolutional networks with predictive coding and feedback mechanisms, enhancing interpretative capabilities through recurrent dynamics [52]. Athar et al. demonstrate the efficacy of Differentiable Soft Masked Attention (DSMA) in improving segmentation accuracy by learning adaptive attention masks [94].

Montagnini et al. propose a hierarchical Bayesian framework for visual motion processing, enhancing tracking performance using predictive cues [81]. Singh et al. underscore the utility of segmentation masks in creating training data, particularly for animal detection in man-made environments [95]. The Kornia library, integrated into neural networks as described by Riba et al., enhances gradient computation efficiency in image processing [58]. In healthcare, Jaspers et al. provide a diverse dataset improving model training for surgical applications, thus enhancing model generalization [73].

Rithish et al. focus on visual perception in mobile imaging systems, improving road safety detection through real-time data [60]. Hashemi et al. automate the identification of Autism Spectrum Disorder (ASD) behavioral markers, showcasing the impact of visual perception modeling in non-invasive diagnostics [9].

These studies collectively illustrate the multifaceted approach to modeling visual perception, integrating cultural, cognitive, and technological insights to develop comprehensive models that enhance human-computer interaction across various applications. This approach addresses cultural perception variations and incorporates interactive visualizations that aid in model error identification and correction, ultimately improving model performance and understanding in diverse contexts [96, 54, 49].

Figure 5 illustrates the hierarchical categorization of advances in modeling visual perception in computer vision, emphasizing cultural diversity, network mechanisms, and specific applications. This visual representation further contextualizes the discussed methodologies, highlighting the interconnectedness of theoretical frameworks and practical implementations in the field.

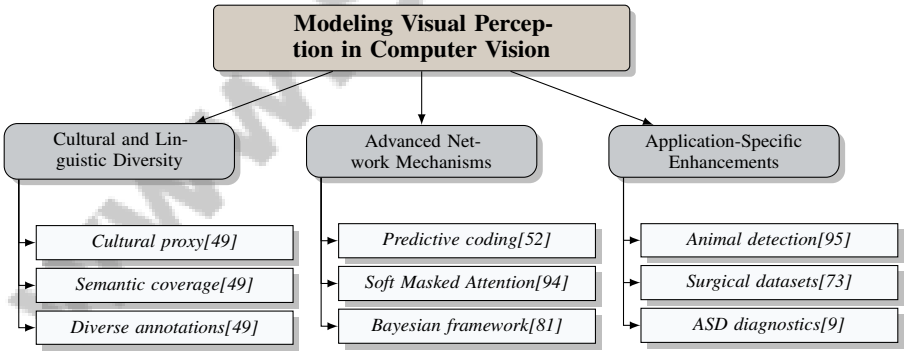

Figure 5: This figure illustrates the hierarchical categorization of advances in modeling visual perception in computer vision, emphasizing cultural diversity, network mechanisms, and specific applications.

6.2 Techniques for Enhancing Machine Perception

Enhancing machine perception in computer vision focuses on improving accuracy, efficiency, and robustness. Integrating synthetic data into training pipelines significantly increases dataset diversity and volume, enhancing machine learning models' generalization capabilities [97]. This approach is particularly valuable when acquiring large amounts of labeled real-world data is challenging. Song et al. emphasize the role of interactive visualizations in identifying and labeling prediction errors, facilitating model performance enhancement through intuitive error analysis interfaces [96].

Attention mechanisms, emulating the human visual system's selective focus, enhance task performance by concentrating computational resources on salient visual data features. These mechanisms improve accuracy and efficiency in applications like semantic segmentation and video understanding, leading to robust and interpretable models [68, 50, 12, 98]. Biologically inspired augmentations, as proposed by Wang et al., align model processing with human visual strategies, improving robustness and efficiency. Advanced architectures like CLIP and transformer-based models capture global context and long-range dependencies, reducing reliance on task-specific training data and facilitating flexible image classification through natural language. Vision-language intelligence integration enriches these models, enhancing generalization in zero or few-shot learning scenarios [99, 54, 28, 1].

The integration of synthetic data, interactive visualizations, attention mechanisms, and biologically inspired augmentations represents a comprehensive approach to advancing machine perception capabilities. These techniques collectively contribute to developing sophisticated computer vision systems capable of diverse tasks—such as autonomous vehicle navigation, security monitoring, and medical image diagnostics—often achieving or surpassing human-level accuracy, demonstrating their versatility and automation potential in niche tasks [99, 54, 18].

6.3 Interaction with the Environment

Visual perception is crucial for computer vision systems to effectively interact with their environment. Advanced algorithms emulating human perceptual processes enable machines to interpret and respond to dynamic surroundings, essential for applications like autonomous vehicles, robotics, and augmented reality (AR). These applications require rapid scene analysis and accurate object recognition for navigation, user assistance, and interactive experiences. Recent advancements in deep learning, particularly neural networks, have improved object detection and depth estimation, facilitating intuitive user-environment interactions [18, 37, 72, 65].

In autonomous systems, perceiving and navigating complex environments is essential. Vanhoorick et al. highlight the importance of robust 4D visual representations for maintaining object persistence despite occlusions, crucial for accurate environmental interaction in dynamic settings [63]. These representations enable systems to predict and respond to environmental changes, enhancing navigation and decision-making capabilities.

Neural Radiance Fields (NeRFs) in robotic applications, as surveyed by Wang et al., demonstrate their potential in rendering realistic 3D environments for interaction and manipulation tasks [45]. In AR, deep neural networks integrated with real-time data processing enable seamless physical world interaction. Kästner et al.'s AR-based human assistance system exemplifies how visual perception can enhance user interaction by providing contextual guidance and feedback [72].

Visual perception modeling also extends to detecting and interpreting human activities, as explored by Hashemi et al. in autism spectrum disorder (ASD) assessment [9]. By analyzing behavioral markers through visual data, computer vision systems can meaningfully interact with human users, providing insights and assistance in healthcare.

These advancements underscore the importance of visual perception in facilitating environmental interaction. By mimicking human perceptual processes and integrating them with advanced machine learning models, computer vision systems have significantly improved their interaction capabilities. This enhancement supports a wide range of applications, from autonomous vehicle navigation and medical image analysis to interactive image search and augmented reality systems, empowering users to better identify and correct model errors, ultimately leading to increased efficiency and effectiveness across various domains [96, 54, 72, 18].

In recent years, the field of computer vision has witnessed significant advancements, leading to its diverse applications across multiple domains. As illustrated in Figure 6, the figure encapsulates various case studies that demonstrate the integration of computer vision in areas such as healthcare, where it aids in diagnostic processes; environmental monitoring, which utilizes visual data for ecological assessments; surveillance systems that enhance security measures; autonomous vehicles that rely on real-time visual processing for navigation; and augmented reality applications that enrich user experiences. Each category not only underscores the technological innovations but also highlights the substantial contributions made to their respective fields, thereby reinforcing the transformative potential of computer vision in contemporary society.

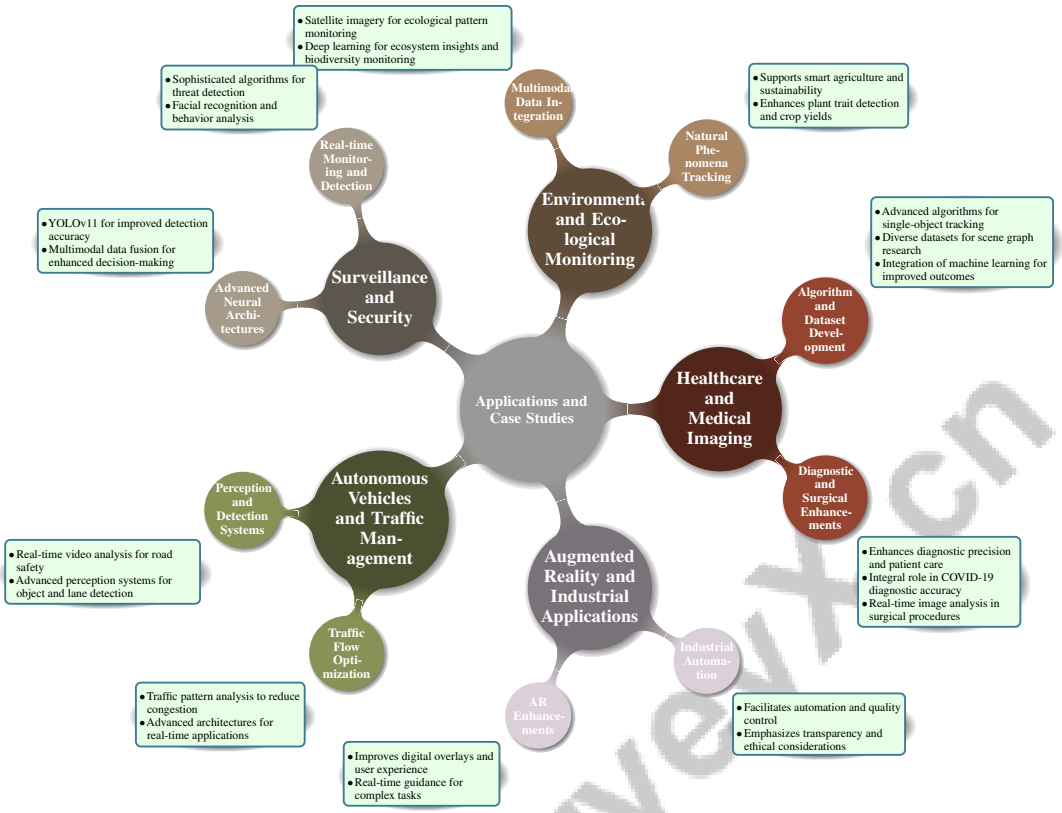

Figure 6: This figure illustrates the diverse applications and case studies of computer vision across various fields, including healthcare, environmental monitoring, surveillance, autonomous vehicles, and augmented reality. Each category highlights specific advancements and contributions to their respective domains.

7 Applications and Case Studies

7.1 Healthcare and Medical Imaging

Computer vision has become integral to healthcare and medical imaging, enhancing diagnostic precision and patient care. The COVID-19 pandemic underscored its importance, with Ulhaq et al. noting its role in improving diagnostic accuracy through various imaging techniques [100]. Advanced algorithms for single-object tracking, as detailed by Han et al., and diverse datasets for scene graph research, emphasized by Chang et al., are crucial for dynamic medical image analysis [101, 13].

As illustrated in Figure 7, the role of computer vision in healthcare encompasses several key areas, including diagnostic enhancements, surgical aids, and the ongoing development of algorithms and datasets. Beyond diagnostics, computer vision aids surgical procedures via real-time image analysis, integrating visual data with textual and sensor inputs to enhance patient assessments and personalize treatment plans. This integration of advanced machine learning techniques improves diagnostic accuracy and predictive analytics, ultimately benefiting patient outcomes [18, 23, 50, 4, 51]. The ongoing development of robust algorithms and diverse datasets is essential for advancing these capabilities [102, 3].

7.2 Environmental and Ecological Monitoring

In environmental and ecological monitoring, computer vision has emerged as a powerful tool for tracking natural phenomena, enhancing data collection accuracy. These technologies support smart agriculture and sustainability by improving plant trait detection and optimizing crop yields [10]. Vallez et al. highlight the use of perceptual similarity metrics in analyzing social media information flow, underscoring computer vision's role in monitoring environmental changes [11].

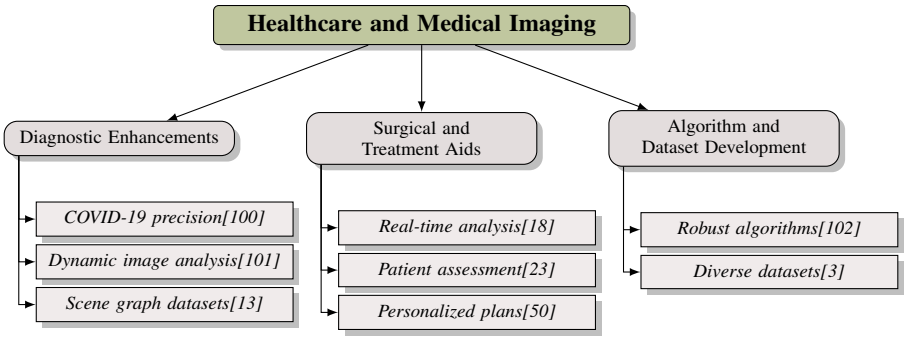

Figure 7: This figure illustrates the role of computer vision in healthcare, highlighting diagnostic enhancements, surgical aids, and the development of algorithms and datasets.

The integration of multimodal data, such as satellite imagery, enhances monitoring of ecological patterns like deforestation and climate change. Techniques like deep learning object detection provide insights into ecosystem health and sustainability [13, 40, 26]. These advancements enable precise biodiversity monitoring and contribute to conservation efforts [12].

7.3 Surveillance and Security

Computer vision in surveillance and security systems enhances real-time monitoring and threat detection. Sophisticated algorithms improve object detection, including facial recognition and behavior analysis, vital for public safety. Models like CCTVCV achieve high accuracy in detecting surveillance cameras, addressing privacy concerns [18, 103, 104].

Deep learning models, particularly CNNs, facilitate robust feature extraction and classification in surveillance. Attention mechanisms enhance these tasks by focusing on salient features [12]. Computer vision also enables behavioral pattern analysis, identifying anomalies that may indicate security threats [74]. The fusion of multimodal data, including audio and thermal imaging, enhances decision-making in surveillance systems [97, 83].

Advanced neural architectures, like YOLOv11, offer improvements in detection accuracy and processing speed, suitable for real-time applications [64]. These advancements contribute to innovative surveillance solutions that enhance safety and address privacy issues [15, 82].

7.4 Autonomous Vehicles and Traffic Management

Computer vision technologies in autonomous vehicles and traffic management systems enhance road safety and operational efficiency. Real-time video analysis identifies road irregularities and traffic violations, while deep learning models improve detection in complex environments [60, 27]. Autonomous vehicles rely on advanced perception systems for object detection, lane detection, and obstacle avoidance.

Robust detection systems are crucial for functioning in diverse conditions, as emphasized by Li et al. [65]. Techniques like oriented bounding boxes enhance detection accuracy in complex environments [65]. In traffic management, computer vision facilitates traffic pattern analysis, optimizing flow and reducing congestion [47].

Advanced architectures, including YOLOv11, improve detection accuracy and efficiency, supporting real-time applications [64]. These technologies enhance road safety and urban mobility by identifying irregularities and optimizing traffic flow [105, 18].

7.5 Augmented Reality and Industrial Applications

Computer vision advancements have significantly enhanced augmented reality (AR) and industrial applications, improving digital overlays and efficiency. In AR, accurate object detection and recognition enhance user experience, as demonstrated by Li et al. [65]. In industrial settings, computer vision facilitates automation and quality control, ensuring adherence to quality standards [38, 57].

AR applications enhance worker efficiency and safety by providing real-time guidance for complex tasks. By integrating advanced computer vision techniques, these applications improve task performance and reduce errors [32, 65]. This approach emphasizes transparency and ethical considerations in developing machine learning technologies, ensuring reliability and societal impact [56, 15].

8 Future Directions and Challenges

The trajectory of computer vision is shaped by challenges and opportunities, notably in data limitations and technological integration. This section delves into these critical aspects, highlighting potential areas for future research and innovation.

8.1 Challenges and Opportunities

Computer vision faces notable challenges in video analysis, real-time processing, and multimodal data integration. Efficiently handling large video datasets while maintaining classification accuracy remains a key issue. Techniques like HSMD offer near real-time processing but require optimization for scalability across diverse applications [41]. Additionally, adapting adversarial attack methods for comprehensive video object segmentation presents research opportunities [75].

Depth estimation for outdoor monocular images complicates lightweight model development for real-time applications, emphasizing the need for improved datasets [37]. Integrating depth estimation with other tasks could enhance capabilities in dynamic environments, prompting further research into model optimization and dataset expansion [65].

Data noise and class imbalance in disaster response models pose challenges but also offer chances to improve robustness and accuracy [106]. The diversity of datasets in surgical computer vision suggests models like SurgeNet could extend to surgical phase recognition [73].

Scalability issues arise from increased computational complexity, highlighting the need for efficient algorithms to manage complex data structures without compromising performance [44]. Addressing biases from human-annotated data is essential for developing fair computer vision systems [107].

Execution speed challenges in machine learning algorithms present opportunities for performance optimization across programming environments [108]. Traditional CNNs face robustness issues against image corruption, suggesting opportunities to enhance classification accuracy through approaches like KerCNN [109].

Integrating AR into industrial workflows encounters challenges such as reliance on outdated technologies like fiducial markers [72]. Future research should focus on incorporating diverse sensory inputs and developing complex motion processing models [81].

Generating scene graphs without bounding box labels presents challenges and opportunities for methodological improvements [110]. Recent advancements, like Explainable Boosting Machines (EBMs), enhance interpretability but often overlook efficiency challenges, particularly for real-time performance and edge device deployment [59, 89]. Proposed methods emphasize reduced memory usage and computational complexity while maintaining competitive performance [111].

Challenges in dynamic urban settings present opportunities for improving detection algorithm robustness [60]. Future research should validate models in clinical environments, enhance data quality, and explore new computer vision applications for COVID-19 [8].

The availability of large annotated datasets constrains the current landscape, and transformers require substantial computational resources [3]. Future research should improve NeRF's adaptability and generalization and explore multimodal interactions in robotics [45]. Long-term occlusions necessitate enhancements in tracking capabilities [63].

Reliance on high-quality event data in methods like EvAC3D underscores the need for improved noise reduction and pose estimation techniques [87]. Scalability issues due to small, well-labeled datasets limit model generalization [1]. Future research could focus on expanding datasets to encompass diverse scenarios and refining algorithms based on benchmarks [31]. Dataset diversity and reproducibility limitations highlight the need for detailed model specifications to ensure robust research outcomes [5]. The biologically plausible data augmentation approach offers a promising avenue for enhancing representation robustness [6].

8.2 Limitations of Current Methods

Current computer vision methodologies face limitations across diverse applications. A major concern is the reliance on synthetic environments for model training, which may not capture real-world complexities, leading to performance discrepancies [112]. Extensive human effort for dataset creation, particularly in outdoor monocular depth estimation, restricts model generalizability [37].

In defect classification within additive manufacturing, class imbalance and dependency on initial labeled samples impact model robustness [57]. Overfitting due to varying image counts per category affects result generalization in material classification [38].

The Open World Recognition (ORE) framework encounters confusion in cluttered scenes or with occluded objects, highlighting the need for robust detection algorithms [113]. Existing benchmarks inadequately address individual animal re-identification, crucial for accurate population estimation in ecological studies [114].

Video inpainting techniques depend on manual tracking or segmentation, necessitating automated solutions for complex scenes [76]. The trade-off between interpretability and performance in concept-based explanations presents a challenge in achieving fidelity while maintaining transparency [71].

The Kornia library provides differentiable image processing functions but may lack optimization for specific tasks compared to dedicated libraries [58]. Existing benchmarks inadequately address execution speed across programming environments, emphasizing the need for comprehensive evaluation frameworks [108].

In dynamic environments, research often fails to address complex material interactions, presenting challenges for real-world applications [45]. The reliance on specific rendering conditions in 3D reconstruction may not generalize to all scenarios, indicating a need for adaptable methodologies [93].

Limitations of the Memory-efficient Instance Segmentation System (MISS) include challenges in generalizing to other datasets lacking rich prior knowledge [111]. Biases in training data and difficulties in generalizing across diverse populations remain significant barriers, necessitating extensive validation in clinical settings [4].

These limitations underscore the need for research and innovation to bolster robustness, scalability, and applicability of computer vision technologies across various applications. Effectively addressing challenges associated with maintenance, transparency, and evolution risks of intelligent services is essential for advancing the field and creating reliable solutions for real-world applications [15, 26, 40, 83].

8.3 Advancements in Machine Learning Techniques

Recent machine learning advancements have significantly enhanced computer vision systems, addressing complex challenges across applications. Integrating Local Binary Patterns (LBP) into CNN architectures illustrates how merging traditional image processing with deep learning improves performance, paving the way for robust models [67]. This integration is vital for developing models capable of handling diverse visual tasks involving intricate textures and patterns.

Incorporating contextual features into classification processes offers promising avenues for future research, extending applications to contemporary curatorial practices [2]. This approach enhances models' interpretative capabilities, enabling better understanding and classification of complex visual data.

Exploring hybrid models that leverage both transformers and CNNs represents a promising direction, as highlighted by Henry et al. By integrating advanced architectures like CLIP, ALIGN, and transformer-based models, researchers can create adaptable and efficient models that excel across various tasks, enhancing accuracy and reducing reliance on task-specific training data [68, 115, 54, 2]. Incorporating self-supervised learning techniques and improving data annotation processes can further refine these models.

Attention mechanisms require more efficient approaches and exploration of hybrid models integrating different types of attention [12]. These advancements are crucial for enhancing the focus and interpretative capabilities of systems in tasks requiring detailed analysis of complex scenes. Figure 8

illustrates the primary advancements in machine learning techniques, highlighting integration methods, attention mechanisms, and optimization strategies. It emphasizes the combination of traditional and modern approaches, efficient focus enhancement, and model optimization for diverse applications.

Applying the lottery ticket hypothesis to neural networks presents opportunities for optimizing pruning techniques and extending their application [61]. This research underscores the potential for creating efficient models by leveraging critical network components.

Future research should prioritize incorporating temporal dynamics into frameworks, indicating potential advancements in techniques for dynamic and sequential data [6]. This focus is vital for enhancing model adaptability in real-time applications.

Developing real-time models and exploring weakly-supervised and unsupervised learning techniques are essential for advancing interpretability and efficiency of architectures [5]. These advancements can lead to robust and scalable models across domains.

These advancements signify substantial progress in addressing challenges faced by computer vision systems. By emphasizing efficiency, scalability, and adaptability, researchers can develop robust models enhancing capabilities across applications. Future research should optimize techniques by investigating diverse datasets and improving real-time applications, refining robustness and efficiency through enhanced documentation practices that consider ethical implications. Employing advanced evaluation frameworks for performance metrics will deepen understanding of model behavior across scenarios [116, 15, 54, 40].

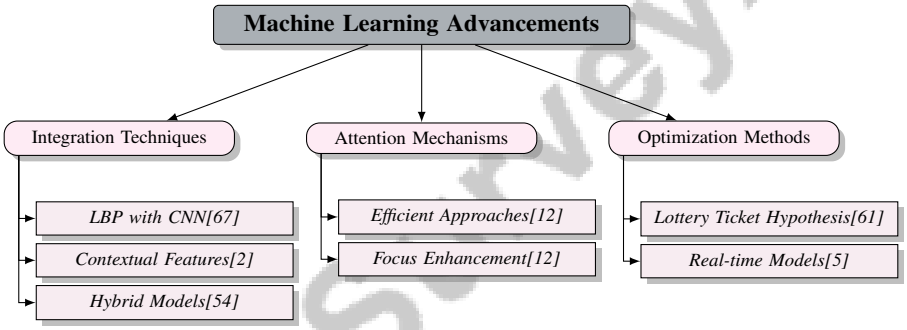

Figure 8: This figure illustrates the primary advancements in machine learning techniques, highlighting integration methods, attention mechanisms, and optimization strategies. It emphasizes the combination of traditional and modern approaches, efficient focus enhancement, and model optimization for diverse applications.

8.4 Integration of Multimodal Data

Integrating multimodal data in computer vision advances understanding and analysis of complex visual environments by synthesizing diverse data sources, including visual, textual, and sensory information. This approach is transformative in applications like visual question answering (VQA), where combining visual and textual data enhances system comprehension and performance [51]. Synthesizing multimodal data is crucial for developing robust frameworks capable of analyzing and interpreting complex scenes.

Future research is expected to refine transformer model efficiency and scalability, explore applications in complex visual tasks, and integrate additional modalities to improve robustness across contexts. These efforts are essential for advancing systems' capabilities in handling varied data types. Integrating polarization data as a multimodal approach offers potential for enhancing visual understanding by providing additional information layers [35].

Developing robust models that generalize across datasets requires improving data augmentation techniques and exploring multimodal integration for enhanced diagnostic capabilities [117]. This direction is crucial for creating systems that adapt to dynamic environments, such as in open-world object detection, where future research could enhance robustness against occlusions and explore methods for unknown identification and class learning [113].

Exploring shape bias integration with performance metrics, like out-of-distribution generalization and adversarial robustness, presents avenues for improving model performance [118]. Developing ethical guidelines for dataset curation and exploring alternative data sourcing methods are critical for ensuring fair outcomes [119].

Integrating multimodal data represents a significant opportunity for advancing computer vision by providing richer insights into visual scenes. Optimizing architectures and leveraging diverse data sources, researchers can develop robust, adaptable, and ethically responsible models capable of performing across applications. Future research could enhance algorithms' capabilities to manage complex movements and improve patch comparisons through additional features [76], and expand Kornia's functionality and optimize performance [58].

8.5 Interdisciplinary Collaboration

Interdisciplinary collaboration is crucial for advancing computer vision, integrating diverse perspectives and expertise to address complex challenges. This is evident in polarization imaging, where combining insights from various domains enhances capabilities [35]. Integrating diverse methodologies drives innovation and improves outcomes in specialized areas.

Developing robust AI systems relies on trust through user authentication and data integrity, benefiting from interdisciplinary input [120]. Involving experts from cybersecurity and human-computer interaction can lead to more secure systems.

In healthcare, integrating AI into applications requires interdisciplinary collaboration to address technological and ethical considerations [4]. This ensures developed technologies are technically sound and ethically aligned with standards.

Integrating Local Binary Patterns (LBP) into neural networks highlights interdisciplinary efforts in adapting methods for irregular data types [67]. Drawing expertise from mathematics and computer science, researchers can develop versatile models handling diverse datasets.

Interdisciplinary collaboration enhances understanding of collaborative processes, integrating computer vision techniques with methods for evaluating team dynamics [121]. This facilitates comprehensive models applicable to a broader range of scenarios.

During the COVID-19 pandemic, collaboration between researchers and healthcare professionals ensures practical application of technologies for disease control [8]. This collaboration translates research innovations into solutions addressing public health challenges.

Exploration of geometric transformations in image processing benefits from interdisciplinary research, enhancing algorithm robustness [122]. Involving experts from geometry and computer science, researchers develop resilient algorithms handling complex transformations.

Integrating different architectures into the GAN framework exemplifies interdisciplinary collaboration's potential in advancing computer vision [110]. Combining insights from various approaches, researchers create powerful models pushing image generation boundaries.

Interdisciplinary collaboration is vital for advancing computer vision, integrating diverse expertise to foster innovation and tackle complex challenges. By combining knowledge from machine learning, social computing, and cognitive science, researchers address data curation ethics, team dynamics, and practical applications in healthcare and autonomous systems [18, 15, 121]. This collaboration enables development of robust, versatile, and ethically responsible technologies impacting society.

9 Conclusion

This survey underscores the pivotal role of integrating computer vision, video analysis, machine learning, 3D reconstruction, and visual perception in driving innovation across multiple domains. These technologies enhance our interaction with visual data, enabling advancements in fields such as medical diagnostics, where deep learning models improve accuracy and efficiency. The development of extensive datasets is essential for refining multimotion estimation algorithms, providing the necessary challenges and ground-truth data for effective evaluation.

In robotics, Neural Radiance Fields (NeRFs) offer significant potential for advancing 3D rendering and modeling capabilities, though further research is required to overcome current limitations. The application of super-resolution processes in traffic scenarios highlights the importance of sophisticated image processing techniques in improving object detection reliability. Moreover, biologically inspired data augmentation methods provide promising alternatives to traditional approaches, enhancing model robustness by aligning with biological principles.

The survey emphasizes the importance of interdisciplinary collaboration and the pursuit of innovative methodologies to address existing challenges. By leveraging integrated approaches, the field of computer vision is poised to enhance machine perception and interaction capabilities, aligning technological advancements with societal needs and values.

References

[1] Feng Li, Hao Zhang, Yi-Fan Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, PengChuan Zhang, and Lei Zhang. Vision-language intelligence: Tasks, representation learning, and large models, 2022.

[2] Daniel Kvak. Leveraging computer vision application in visual arts: A case study on the use of residual neural network to classify and analyze baroque paintings, 2022.

[3] Emerald U. Henry, Onyeka Emebob, and Conrad Asotie Omonhinmin. Vision transformers in medical imaging: A review, 2022.

[4] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021.

[5] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey, 2020.

[6] Binxu Wang, David Mayo, Arturo Deza, Andrei Barbu, and Colin Conwell. On the use of cortical magnification and saccades as biological proxies for data augmentation, 2021.

[7] Mark Nixon and Alberto Aguado. *Feature extraction and image processing for computer vision*. Academic press, 2019.

[8] Anwaar Ulhaq, Asim Khan, Douglas Gomes, and Manoranjan Paul. Computer vision for covid-19 control: A survey, 2020.

[9] Jordan Hashemi, Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, and Guillermo Sapiro. Computer vision tools for the non-invasive assessment of autism-related behavioral markers, 2012.

[10] Mostafa Dehghani, Alexey Gritsenko, Anurag Arnab, Matthias Minderer, and Yi Tay. Scenic: A jax library for computer vision research and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21393–21398, 2022.

[11] Cyril Vallez, Andrei Kucharavy, and Ljiljana Dolamic. Needle in a haystack, fast: Benchmarking image perceptual similarity metrics at scale, 2022.

[12] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022.

[13] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application, 2022.

[14] Lukas Tuggener, Ismail Elezi, Jürgen Schmidhuber, Marcello Pelillo, and Thilo Stadelmann. Deepscores – a dataset for segmentation, detection and classification of tiny objects, 2018.

[15] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. Do datasets have politics? disciplinary values in computer vision dataset development, 2021.

[16] Mohsen Joneidi, Alireza Zaeemzadeh, Nazanin Rahnavard, and Mubarak Shah. Iterative projection and matching: Finding structure-preserving representatives and its application to computer vision, 2018.

[17] Syed Hammad Ahmed, Shengnan Hu, and Gita Sukthankar. The potential of vision-language models for content moderation of children's videos, 2023.

[18] Alan F. Smeaton. Computer vision for supporting image search, 2021.

[19] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective, 2021.

[20] Wei Wang, Weiqing Min, Tianhao Li, Xiaoxiao Dong, Haisheng Li, and Shuqiang Jiang. A review on vision-based analysis for automatic dietary assessment, 2022.

[21] Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. Deep learning vs. traditional computer vision. In *Advances in computer vision: proceedings of the 2019 computer vision conference (CVC), volume 1 1*, pages 128–144. Springer, 2020.

[22] Nathan Frey, Peggy Chi, Weilong Yang, and Irfan Essa. Automatic non-linear video editing transfer, 2021.

[23] Junfeng Gao, Yong Yang, Pan Lin, and Dong Sun Park. Computer vision in healthcare applications. *Journal of healthcare engineering*, 2018:5157020, 2018.

[24] T. Bouwmans and B. Garcia-Garcia. Background subtraction in real applications: Challenges, current models and future directions, 2019.

[25] Ke Fan, Jiangning Zhang, Ran Yi, Jingyu Gong, Yabiao Wang, Yating Wang, Xin Tan, Chengjie Wang, and Lizhuang Ma. Textual decomposition then sub-motion-space scattering for open-vocabulary motion generation, 2024.

[26] Chenjiao Tan, Qian Cao, Yiwei Li, Jielu Zhang, Xiao Yang, Huaqin Zhao, Zihao Wu, Zhengliang Liu, Hao Yang, Nemin Wu, Tao Tang, Xinyue Ye, Lilong Chai, Ninghao Liu, Changying Li, Lan Mu, Tianming Liu, and Gengchen Mai. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications, 2023.

[27] Iván García, Rafael Marcos Luque, and Ezequiel López. Improved detection of small objects in road network sequences, 2021.

[28] Guy Ben-Yosef, Liav Assif, Daniel Harari, and Shimon Ullman. A model for full local image interpretation, 2021.

[29] Andrea Ferreri, Silvia Bucci, and Tatiana Tommasi. Multi-modal rgb-d scene recognition across domains, 2021.

[30] Nikolay Savinov, Akihito Seki, Lubor Ladicky, Torsten Sattler, and Marc Pollefeys. Quad-networks: unsupervised learning to rank for interest point detection, 2017.

[31] Kevin M. Judd and Jonathan D. Gammell. The oxford multimotion dataset: Multiple se(3) motions with ground truth, 2019.

[32] Ansh Mittal. Neural radiance fields: Past, present, and future, 2024.

[33] Maria-Laura Torrente, Silvia Biasotti, and Bianca Falcidieno. Recognition of feature curves on 3d shapes using an algebraic approach to hough transforms, 2017.

[34] Yajie Sun, Ali Zia, Vivien Rolland, Charissa Yu, and Jun Zhou. Spectral 3d computer vision – a review, 2023.

[35] Joaquin Rodriguez, Lew-Fock-Chong Lew-Yan-Voon, Renato Martins, and Olivier Morel. Pola4all: survey of polarimetric applications and an open-source toolkit to analyze polarization, 2023.

[36] Chenfeng Xu, Shijia Yang, Tomer Galanti, Bichen Wu, Xiangyu Yue, Bohan Zhai, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2point: 3d point-cloud understanding with 2d image pretrained models, 2022.

[37] Pulkit Vyas, Chirag Saxena, Anwesh Badapanda, and Anurag Goswami. Outdoor monocular depth estimation: A research review, 2022.

[38] Anca Sticlaru. Material classification using neural networks, 2017.

[39] Nati Ofir and Jean-Christophe Nebel. Classic versus deep learning approaches to address computer vision challenges, 2021.

[40] Sophie Crommelinck, Mila Koeva, Michael Ying Yang, and George Vosselman. Robust object extraction from remote sensing data, 2019.

[41] Pedro Machado, Andreas Oikonomou, Joao Filipe Ferreira, and T. M. McGinnity. Hsmd: An object motion detection algorithm using a hybrid spiking neural network architecture, 2021.

[42] Pradyumn Patil, Vishwajeet Pawar, Yashraj Pawar, and Shruti Pisal. Video content classification using deep learning, 2021.

[43] Peng Gao, Yipeng Ma, Ruyue Yuan, Liyi Xiao, and Fei Wang. Learning cascaded siamese networks for high performance visual tracking, 2019.

[44] Vinay Jethava. Extension of path probability method to approximate inference over time, 2009.

[45] Guangming Wang, Lei Pan, Songyou Peng, Shaohui Liu, Chenfeng Xu, Yanzi Miao, Wei Zhan, Masayoshi Tomizuka, Marc Pollefeys, and Hesheng Wang. Nerf in robotics: A survey, 2024.

[46] Samvit Jain, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, and Joseph E. Gonzalez. Scaling video analytics systems to large camera deployments, 2019.

[47] Olivia Wiles, Joao Carreira, Iain Barr, Andrew Zisserman, and Mateusz Malinowski. Compressed vision for efficient video understanding, 2022.

[48] Daniel Rivas, Francesc Guim, Jordà Polo, Pubudu M. Silva, Josep Ll. Berral, and David Carrera. Towards automatic model specialization for edge video analytics, 2021.

[49] Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, and Ranjay Krishna. Computer vision datasets and models exhibit cultural and linguistic diversity in perception, 2024.

[50] Babette Bühler. Multimodal machine learning for automated assessment of attention-related processes during learning, 2024.

[51] Xiaolan Chen, Ruoyu Chen, Pusheng Xu, Weiyi Zhang, Xianwen Shang, Mingguang He, and Danli Shi. Visual question answering in ophthalmology: A progressive and practical perspective, 2024.

[52] Zhaoyang Pang, Callum Biggs O'May, Bhavin Choksi, and Rufin VanRullen. Predictive coding feedback results in perceived illusory contours in a recurrent neural network, 2021.

[53] Vamshi K. Kancharala, Debanjali Bhattacharya, and Neelam Sinha. Spatial encoding of bold fmri time series for categorizing static images across visual datasets: A pilot study on human vision, 2023.

[54] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: Towards characterization of broader capabilities and downstream implications, 2021.

[55] Meike Nauta and Christin Seifert. The co-12 recipe for evaluating interpretable part-prototype image classifiers, 2023.

[56] Agathe Balayn, Bogdan Kulynych, and Seda Guerses. Exploring data pipelines through the process lens: a reference model forcomputer vision, 2021.

[57] Xiao Liu, Alessandra Mileo, and Alan F. Smeaton. Defect classification in additive manufacturing using cnn-based vision processing, 2023.

[58] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch, 2019.

[59] Daniel Schug, Sai Yerramreddy, Rich Caruana, Craig Greenberg, and Justyna P. Zwolak. Extending explainable boosting machines to scientific image data, 2023.

[60] Harish Rithish, Raghava Modhugu, Ranjith Reddy, Rohit Saluja, and C. V. Jawahar. Evaluating computer vision techniques for urban mobility on large-scale, unconstrained roads, 2021.

[61] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models, 2021.

[62] Yibo Miao, Yu Lei, Feng Zhou, and Zhijie Deng. Bayesian exploration of pre-trained models for low-shot image classification, 2024.

[63] Basile Van Hoorick, Purva Tendulkar, Didac Suris, Dennis Park, Simon Stent, and Carl Vondrick. Revealing occlusions with 4d neural fields, 2022.

[64] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024.

[65] Vladislav Li, Barbara Villarini, Jean-Christophe Nebel, Thomas Lagkas, Panagiotis Sarigiannidis, and Vasileios Argyriou. Evaluation of environmental conditions on object detection using oriented bounding boxes for ar applications, 2024.

[66] M. W. Spratling. Explaining away results in accurate and tolerant template matching, 2019.

[67] Zhuo Su, Matti Pietikäinen, and Li Liu. From local binary patterns to pixel difference networks for efficient visual representation learning, 2023.

[68] Mohit Vaishnav. Phd thesis: Exploring the role of (self-)attention in cognitive and computer vision architecture, 2023.

[69] Zhen He, Jian Li, Daxue Liu, Hangen He, and David Barber. Tracking by animation: Unsupervised learning of multi-object attentive trackers, 2019.

[70] Tetsuo Inoshita, Yuichi Nakatani, Katsuhiko Takahashi, Asuka Ishii, and Gaku Nakano. Trained model fusion for object detection using gating network, 2020.

[71] Gayda Mutahar and Tim Miller. Concept-based explanations using non-negative concept activation vectors and decision tree for cnn models, 2022.

[72] Linh Kästner, Leon Eversberg, Marina Mursa, and Jens Lambrecht. Integrative object and pose to task detection for an augmented-reality-based human assistance system using neural networks, 2020.

[73] Tim J. M. Jaspers, Ronald L. P. D. de Jong, Yasmina Al Khalil, Tijn Zeelenberg, Carolus H. J. Kusters, Yiping Li, Romy C. van Jaarsveld, Franciscus H. A. Bakker, Jelle P. Ruurda, Willem M. Brinkman, Peter H. N. De With, and Fons van der Sommen. Exploring the effect of dataset diversity in self-supervised learning for surgical computer vision, 2024.

[74] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning, 2021.

[75] Ping Li, Yu Zhang, Li Yuan, Jian Zhao, Xianghua Xu, and Xiaoqin Zhang. Adversarial attacks on video object segmentation with hard region discovery, 2023.

[76] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes, 2015.

[77] Jonathan Connell and Benjamin Herta. Structured differential learning for automatic threshold setting, 2018.

[78] Mark Buckler, Suren Jayasuriya, and Adrian Sampson. Reconfiguring the imaging pipeline for computer vision, 2017.

[79] Amgad Ahmed, Suhong Kim, Mohamed Elgharib, and Mohamed Hefeeda. User-assisted video reflection removal, 2020.

[80] Yupei Zhang and Kwok-Leung Chan. Saliency detection with moving camera via background model completion, 2021.

[81] Anna Montagnini, Laurent Perrinet, and Guillaume S Masson. Visual motion processing and human tracking behavior, 2016.

[82] Chongke Wu, Sicong Shao, Cihan Tunc, and Salim Hariri. Video anomaly detection using pre-trained deep convolutional neural nets and context mining, 2020.

[83] Alex Cummaudo, Rajesh Vasa, John Grundy, Mohamed Abdelrazek, and Andrew Cain. Losing confidence in quality: Unspoken evolution of computer vision services, 2019.

[84] Yiming Zhou, Callen MacPhee, Madhuri Suthar, and Bahram Jalali. Phycv: The first physics-inspired computer vision library, 2023.

[85] Ming Li and ChenHao Guo. Automatic annotation of visual deep neural networks, 2021.

[86] Allan Pinto, Manuel A. Córdova, Luis G. L. Decker, Jose L. Flores-Campana, Marcos R. Souza, Andreza A. dos Santos, Jhonatas S. Conceição, Henrique F. Gagliardi, Diogo C. Luvizon, Ricardo da S. Torres, and Helio Pedrini. Parallax motion effect generation through instance segmentation and depth estimation, 2020.

[87] Ziyun Wang, Kenneth Chaney, and Kostas Daniilidis. Evac3d: From event-based apparent contours to 3d models via continuous visual hulls, 2023.

[88] Davyd Svyezhentsev, George Retsinas, and Petros Maragos. Pre-training for action recognition with automatically generated fractal datasets, 2024.

[89] Lorenzo Papa, Paolo Russo, Irene Amerini, and Luping Zhou. A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking, 2024.

[90] Sai Vishwanath Venkatesh, Atra Akandeh, and Madhu Lokanath. Metapix: A data-centric ai development platform for efficient management and utilization of unstructured computer vision data, 2024.

[91] Vasileios Leon, Panagiotis Minaidis, George Lentaris, and Dimitrios Soudris. Accelerating ai and computer vision for satellite pose estimation on the intel myriad x embedded soc, 2024.

[92] Tianyu Wu, Konrad Schindler, and Cenek Albl. 3d reconstruction from public webcams, 2021.

[93] Anh Thai, Stefan Stojanov, Vijay Upadhya, and James M. Rehg. 3d reconstruction of novel object shapes from single images, 2021.

[94] Ali Athar, Jonathon Luiten, Alexander Hermans, Deva Ramanan, and Bastian Leibe. Differentiable soft-masked attention, 2022.

[95] Abhineet Singh, Marcin Pietrasik, Gabriell Natha, Nehla Ghouaiel, Ken Brizel, and Nilanjan Ray. Animal detection in man-made environments, 2020.

[96] Hayeong Song, Gonzalo Ramos, and Peter Bodik. Evaluating how interactive visualizations can assist in finding samples where and how computer vision models make mistakes, 2024.

[97] Reddy Mandati, Vladyslav Anderson, Po chen Chen, Ankush Agarwal, Tatjana Dokic, David Barnard, Michael Finn, Jesse Cromer, Andrew Mccauley, Clay Tutaj, Neha Dave, Bobby Besharati, Jamie Barnett, and Timothy Krall. Integrating artificial intelligence models and synthetic image data for enhanced asset inspection and defect identification, 2024.

[98] António Farinhas, André F. T. Martins, and Pedro M. Q. Aguiar. Multimodal continuous visual attention mechanisms, 2021.

[99] Gracile Astlin Pereira and Muhammad Hussain. A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships, 2024.

[100] Anwaar Ulhaq, Asim Khan, Douglas Gomes, and Manoranjan Paul. Computer vision for covid-19 control: a survey. *arXiv preprint arXiv:2004.09420*, 2020.

[101] Ruize Han, Wei Feng, Qing Guo, and Qinghua Hu. Single object tracking research: A survey, 2022.

[102] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. Transformers in medical image analysis: A review, 2022.

[103] Hannu Turtiainen, Andrei Costin, Tuomo Lahtinen, Lauri Sintonen, and Timo Hamalainen. Towards large-scale, automated, accurate detection of cctv camera objects using computer vision. applications and implications for privacy, safety, and cybersecurity. (preprint), 2021.

[104] Purnendu Prabhat, Himanshu Gupta, and Ajeet Kumar Vishwakarma. Face detection: Present state and research directions, 2024.

[105] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and trends® in computer graphics and vision*, 12(1–3):1–308, 2020.

[106] Marc Bosch, Christian Conroy, Benjamin Ortiz, and Philip Bogden. Improving emergency response during hurricane season using computer vision, 2020.

[107] Dat Tien Nguyen, Firoj Alam, Ferda Ofli, and Muhammad Imran. Automatic image filtering on social networks using deep learning and perceptual hashing during crises, 2017.

[108] Ahmed A. Elsayed and Waleed A. Yousef. Matlab vs. opencv: A comparative study of different machine learning algorithms, 2019.

[109] Noemi Montobbio, Laurent Bonnasse-Gahot, Giovanna Citti, and Alessandro Sarti. Kercnns: biologically inspired lateral connections for classification of corrupted images, 2019.

[110] Matthew Klawonn and Eric Heim. Generating triples with adversarial networks for scene graph construction, 2018.

[111] Chih-Chung Hsu and Chia-Ming Lee. Miss: Memory-efficient instance segmentation framework by visual inductive priors flow propagation, 2024.

[112] Brady Zhou, Philipp Krähenbühl, and Vladlen Koltun. Does computer vision matter for action? *Science Robotics*, 4(30):eaaw6661, 2019.

[113] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection, 2021.

[114] Stefan Schneider, Graham W. Taylor, and Stefan C. Kremer. Deep learning object detection methods for ecological camera trap data, 2018.

[115] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook, 2023.

[116] Xuechen Zhang, Samet Oymak, and Jiasi Chen. Post-hoc models for performance estimation of machine learning inference, 2021.

[117] Imran Ul Haq. An overview of deep learning in medical imaging, 2022.

[118] Tiago Oliveira, Tiago Marques, and Arlindo L. Oliveira. Connecting metrics for shape-texture knowledge in computer vision, 2023.

[119] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE, 2021.

[120] Christian Meske and Enrico Bunde. Transparency and trust in human-ai-interaction: The role of model-agnostic explanations in computer vision-based decision support, 2020.

[121] Zang Guo and Roghayeh Barmaki. Deep neural networks for collaborative learning analytics: Evaluating team collaborations using student gaze point prediction, 2020.

[122] Juan Gerardo Alcázar, Gema M. Diaz-Toca, and Carlos Hermosa. On the problem of detecting when two implicit plane algebraic curves are similar, 2015.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.