ARQUITETURAS DE APRENDIZADO PROFUNDO PARA SISTEMAS MULTI-VIEW: uma revisão

Aline Gabrig[1], Eduardo Bontorin[2], Tiago Matitz[3], Ricardo Massao Kagami[4] e Renato Sellaro Dorighello[5]

RESUMO

Sistemas de visão multi-câmera apresentam desafios específicos relacionados a oclusões, sincronização de vistas e associação inter-câmera, aspectos que detectores monoculares não tratam de forma nativa. Este trabalho tem por objetivo identificar quais arquiteturas de aprendizado profundo são utilizadas no processamento de imagens provenientes de múltiplas câmeras, por meio de revisão bibliográfica da literatura. A análise organiza os modelos em três eixos principais: detectores convolucionais de dois e um estágio, detectores baseados em Transformers e arquiteturas multi-câmera com estratégias de fusão tardia ou processamento *end-to-end*. Os resultados indicam que mecanismos de auto-atenção favorecem a modelagem de dependências entre diferentes vistas, enquanto abordagens sem *Non-Maximum Suppression* (NMS) — técnica de pós-processamento para eliminação de caixas delimitadoras redundantes — reduzem limitações associadas à supressão de detecções válidas em cenários com campos de visão sobrepostos. Observa-se ainda que pipelines de fusão tardia, como *a Multi-View Convolutional Neural Networ*k (MVCNN) e o sistema de rastreamento multi-câmera de Zheng et al., oferecem maior compatibilidade com infraestruturas existentes, ao passo que arquiteturas *end-to-end*, como o MV-HFMD e o HMDMV, reduzem a propagação de erros, embora exijam maior capacidade computacional e conjuntos de dados multi-câmera anotados. Conclui-se que a escolha arquitetural deve considerar os requisitos de latência, acurácia, infraestrutura disponível e disponibilidade de dados do cenário aplicado.

Palavras-chave: Visão Multi-Câmera; Aprendizado Profundo; Detecção de Objetos; Classificação Multi-Vista.

DEEP-LEARNING ARCHITECTURES FOR MULTI-VIEW SYSTEMS: review

ABSTRACT

Multi-camera vision systems present specific challenges related to occlusions, view synchronization, and inter-camera association — aspects that monocular detectors do

[1] Centro Universitário Senai Paraná. – Engenharia Mecatrônica — aline.gabrig@outlook.com
[2] Centro Universitário Senai Paraná. – Engenharia Mecatrônica — eduardo.bontorin@outlook.com
[3] Centro Universitário Senai Paraná. – Engenharia Mecatrônica — tiagomatitz@gmail.com
[4] Centro Universitário Senai Paraná – Orientador — ricardo.zanotto@sistemafiep.org.br
[5] Centro Universitário Senai Paraná. – Corientador — renato.dorighello@sistemafiep.org.br (Autor correspondente)

not natively address. This work aims to identify which deep learning architectures are employed in the processing of images from multiple cameras, through a bibliographic review of the specialized literature. The analysis organizes the models along three main axes: two-stage and single-stage convolutional detectors, Transformer-based detectors, and multi-camera architectures with late fusion strategies or end-to-end processing. The results indicate that self-attention mechanisms favor the modeling of dependencies across different views, while Non-Maximum Suppression (NMS)-free approaches — a post-processing technique for eliminating redundant bounding boxes — reduce limitations associated with the suppression of valid detections in scenarios with overlapping fields of view. It is further observed that late fusion pipelines, such as the Multi-View Convolutional Neural Network (MVCNN) and Zheng et al.'s multi-camera tracking system, offer greater compatibility with existing infrastructures, whereas end-to-end architectures, such as MV-HFMD and HMDMV, reduce error propagation, albeit requiring greater computational capacity and annotated multi-camera datasets. It is concluded that the architectural choice must consider the latency, accuracy, available infrastructure, and data availability requirements of the target scenario.

Key words: Multi-Camera Vision; Deep Learning; Object Detection; Multi-View Classification.

1 INTRODUÇÃO

A visão computacional é uma área da inteligência artificial dedicada a sistemas capazes de extrair e interpretar informações visuais a partir de imagens, de modo análogo ao sistema visual humano (Zou *et al.*, 2023). Combina processamento de sinais, aprendizado de máquina e geometria projetiva para identificar o que está em uma cena, onde se localiza e como os elementos se relacionam (Munsif *et al.*, 2025), viabilizando o reconhecimento automatizado de padrões sem intervenção humana (Cao *et al.*, 2025).

Entre suas tarefas, a detecção de objetos identifica a classe semântica e a localização espacial de um objeto por meio de uma caixa delimitadora (Cao *et al.*, 2025; Zou *et al.*, 2023), servindo de base para segmentação, rastreamento e compreensão de cenas, com aplicações em direção autônoma (Grigorescu *et al.*, 2020), vigilância (Munsif *et al.*, 2025), robótica (Wang *et al.*, 2026) e inspeção industrial (Bhavanasi *et al.*, 2025).

No contexto industrial, a visão de máquina surgiu para superar limitações da inspeção visual humana, como fadiga e variações em processos de alta velocidade (Hütten *et al.*, 2024), sendo integrada à Indústria 4.0 por meio de sensores, robôs e sistemas de rastreabilidade (Javaid *et al.*, 2022). As primeiras aplicações empregavam técnicas de processamento de imagens, como limiarização e detecção de bordas, eficazes em ambientes controlados (Bhatt *et al.*, 2021), mas com generalização limitada.

Diante dessas limitações de generalização, o aprendizado profundo emergiu como alternativa, permitindo que modelos aprendessem características diretamente dos dados, com redes neurais convolucionais (CNNs — *convolutional neural networks*) ocupando posição central em classificação, detecção e segmentação (Abbas *et al.*, 2025; Diwan; Anirudh; Tembhurne, 2023; Nasir *et al.*, 2025), enquanto *Vision Transformers* vêm sendo explorados em inspeção de semicondutores e monitoramento de máquinas de Controle Numérico Computadorizado (CNC) (Hütten *et al.*, 2024; Radhakrishnan *et al.*, 2025; Akele; Gurau, 2026).

Muitos sistemas industriais ainda operam com imagens monoculares, configuração de menor custo, porém limitada em tarefas que exigem percepção espacial e análise tridimensional (Dong; Sheng, 2019). Abordagens de câmera única enfrentam oclusões, variações de ponto de vista e campo de visão restrito (Perez Soler *et al.*, 2025), além de ambiguidade de profundidade (Kamali *et al.*, 2026). Esses fatores motivaram o desenvolvimento de sistemas *multi-câmera*, que integram perspectivas complementares e apoiam tarefas de reconstrução tridimensional e inspeção de superfícies (Kamali *et al.*, 2026).

Embora existam revisões sobre detecção de objetos com aprendizado profundo (Zou *et al.*, 2023; Cao *et al.*, 2025) e sobre classificação *multi-view* (Berahmand *et al.*, 2025; Yu *et al.*, 2025), os trabalhos disponíveis cobrem o campo de forma ampla, sem recortar o cenário de câmeras calibradas nem tratar a detecção como componente do *pipeline*. Esta pesquisa busca preencher essa lacuna ao identificar quais arquiteturas de aprendizado profundo têm sido utilizadas para a

classificação cross-view em sistemas multi-câmera com vistas estruturadas e quais estratégias de fusão se mostram adequadas para integrar essas perspectivas em aplicações industriais, articulando detecção, fusão e classificação.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os fundamentos teóricos necessários à compreensão das arquiteturas de aprendizado profundo para sistemas *multi-câmera*. Inicia-se com a trajetória histórica da detecção de objetos, da atribuição manual de características ao aprendizado profundo, seguida pela descrição dos principais componentes arquiteturais — *backbones*, detectores de dois e um estágio, e detectores baseados em *Transformers*. Em sequência, são apresentados os conjuntos de dados e métricas de avaliação relevantes, encerrando com a fundamentação sobre arquiteturas multi-câmera e estratégias de fusão de vistas.

2.1 Atribuição de características manuais ao aprendizado profundo

No início dos anos 2000, o campo da detecção de objetos foi dominado por métodos baseados em características desenhadas manualmente (*handcrafted features*) combinadas com a técnica de janelas deslizantes, em que um descritor fixo era deslocado sobre toda a imagem em múltiplas escalas para identificar regiões contendo objetos (Zou *et al.*, 2023).

Esses detectores apresentavam limitações fundamentais: a capacidade representacional dos descritores manuais era insuficiente para capturar a variabilidade visual de cenas reais, e a baixa eficiência computacional dos sistemas da época tornava-os particularmente frágeis em cenários com oclusão e variação de aparência (Cao *et al.*, 2025).

Entre os principais algoritmos dessa época, destacam-se o detector Viola-Jones (Viola; Jones, 2004), que foi o primeiro a realizar detecção facial em tempo real, apoiando-se em imagens integrais, seleção de características via AdaBoost e

cascatas de classificadores. Em 2005, o descritor HOG (*Histogram of Oriented Gradients*) introduziu uma representação baseada em histogramas de gradientes orientados normalizados localmente, tornando-se referência para detecção de pedestres (Dalal; Triggs, 2005). O ápice dos detectores de caracterização manual foi o DPM (*Deformable Part-based Model*), vencedor das competições PASCAL VOC de 2007 a 2009, que decompunha objetos em partes deformáveis (Felzenszwalb *et al.*, 2010).

Em 2012, Krizhevsky, Sutskever e Hinton (2012) propuseram a AlexNet, inaugurando as pesquisas em detectores de aprendizado profundo. Sendo a primeira rede neural profunda a incorporar camadas convolucionais em larga escala, com função de ativação *Rectified Linear Unit* (ReLU) e regularização por *dropout*, a AlexNet demonstrou que CNNs podiam aprender representações hierárquicas de bordas e texturas nas camadas iniciais até partes de objetos e categorias semânticas nas camadas profundas. Essa capacidade ofereceu uma alternativa decisiva aos descritores manuais, respondida em 2014 pelo R-CNN (*Region-based Convolutional Neural Network*) (Girshick *et al.*, 2014), que inaugurou os detectores modernos baseados em aprendizado profundo.

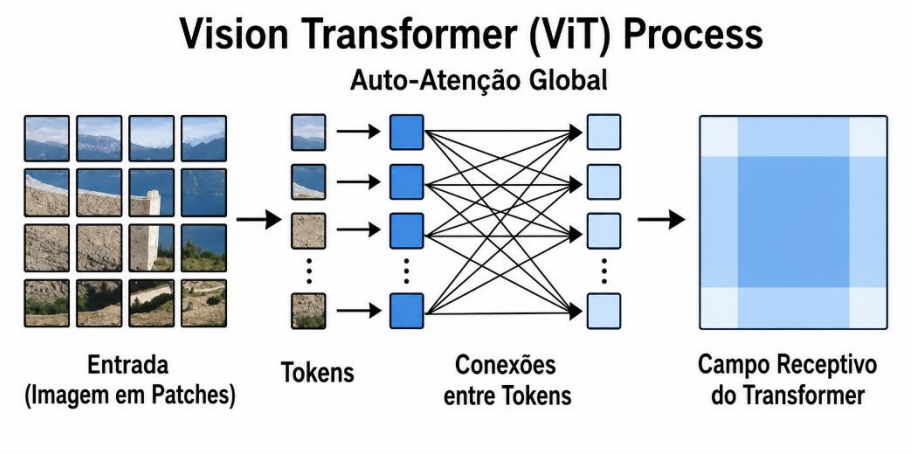
Mais recentemente, os *Transformers* introduzidos por Vaswani *et al.* (2017), originalmente desenvolvidos para processamento de linguagem natural, foram adaptados para visão computacional, profundamente afetando todo o campo. O *Vision Transformer* (ViT) divide a imagem em quadros não sobrepostos projetados em *embeddings* de *tokens* (Dosovitskiy *et al.*, 2021).

Diferente das CNNs, os *Transformers* utilizam mecanismos de auto-atenção que permitem modelar relações globais entre todas as regiões da imagem desde a primeira camada, oferecendo um campo receptivo de escala global conforme ilustrado na Figura 1.

Atualmente, as dez melhores abordagens de detecção no conjunto de dados COCO (*Common Objects in Context*) (Lin *et al.*, 2014) são baseadas em *Transformers*, e a diferença de desempenho entre *Transformers* e CNNs vem se

ampliando gradualmente (Cao *et al.*, 2025).

Figura 1 — Vision Transformer (ViT) Process



Fonte: Adaptado de Dosovitskiy et al. (2021).

2.2 Backbones: Extratores de Características

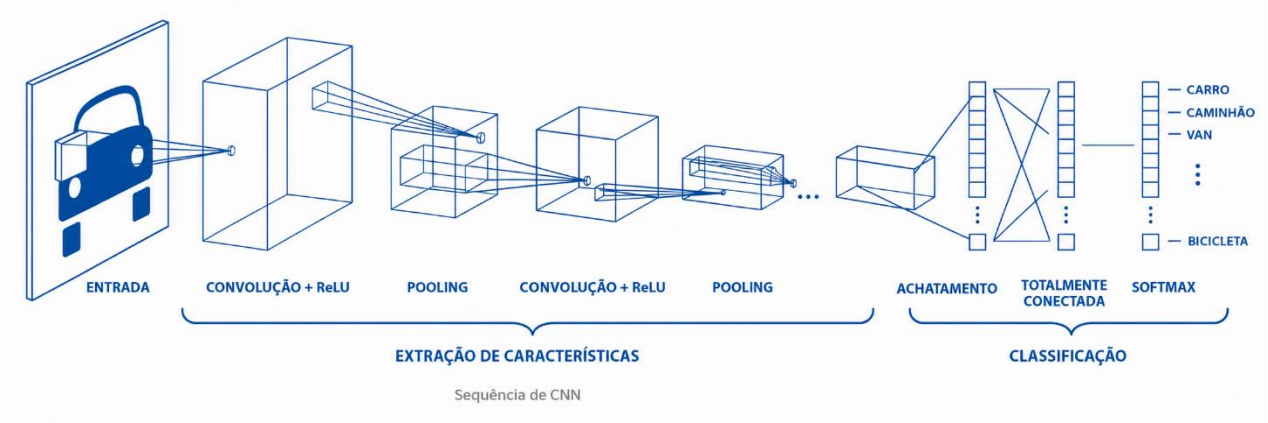
Os *backbones* são sub-redes responsáveis por extrair representações hierárquicas de características (*features*) das imagens, constituindo o componente fundamental sobre o qual os detectores modernos são construídos (Cao *et al.*, 2025; Zou *et al.*, 2023). Esses extratores derivam das CNNs, que processam a imagem por meio de camadas convolucionais sucessivas em vez de operar diretamente sobre os pixels brutos (Krizhevsky; Sutskever; Hinton, 2012).

Cada camada aplica filtros (*kernels*) que percorrem toda a imagem para detectar padrões locais, seguidos de ativações ReLU, que introduzem não-linearidade (Krizhevsky; Sutskever; Hinton, 2012), e de operações de *pooling*, que reduzem a resolução espacial e conferem invariância parcial a translações (Zaidi *et al.*, 2022).O empilhamento sucessivo dessas operações constrói a hierarquia de características citada, das bordas e texturas nas camadas iniciais às estruturas semânticas nas camadas profundas (Krizhevsky; Sutskever; Hinton, 2012; Zou *et al.*, 2023).

A Figura 2 ilustra essa sequência: a etapa de extração de características, composta por convoluções, ativações ReLU e *pooling*, antecede o achatamento e as camadas totalmente conectadas, responsáveis pela classificação. Nos detectores, o

backbone corresponde precisamente a essa porção convolucional de extração, sobre a qual são acoplados os módulos de detecção (Cao *et al.*, 2025).

Figura 2 — Sequência de uma CNN



Fonte: Adaptado de Demir (2022).

A partir da AlexNet (Krizhevsky; Sutskever; Hinton, 2012), essas arquiteturas evoluíram em profundidade, eficiência e capacidade de modelagem global. O VGGNet (Simonyan; Zisserman, 2014) consolidou filtros 3×3 empilhados como padrão; o MobileNet (Howard *et al.*, 2017) e o EfficientNet (Tan; Le, 2019) priorizaram eficiência computacional por meio de convoluções separáveis e escalonamento composto, respectivamente. Com os *Transformers*, o *Swin Transformer* (Liu *et al.*, 2021) introduziu atenção em janelas deslocadas com complexidade linear, enquanto o YOLO11 (Ultralytics, 2024) convergiu os dois paradigmas com o módulo C2PSA (*Cross-Stage Partial with Self-Attention*). A Tabela 1 sintetiza essas contribuições.

Tabela 1 – Comparativo de *backbones*

Backbone	Ano	Paradigma	Principal contribuição
VggNet	2014	CNN	Filtros 3×3 empilhados, estabelecendo padrão de profundidade uniforme
MobileNet	2017	CNN	Convoluções separáveis em profundidade para dispositivos móveis
EfficientNet	2019	CNN	Escalonamento composto de profundidade, largura e resolução

Swin Transformer	2021	*Transformer*	Janelas deslocadas com complexidade linear e representação hierárquica multi-escala
YOLO11 (*backbone*)	2024	Híbrido (CNN + atenção)	Módulo C2PSA integrando auto-atenção leve a arquitetura convolucional

Fonte: Adaptado de Cao *et al.* (2025) e Zou *et al.* (2023).

2.3 Detectores de Dois Estágios

Os detectores de dois estágios seguem um paradigma hierárquico do grosseiro ao fino (*coarse-to-fine*), a primeira etapa identifica regiões candidatas de interesse, com foco em alta revocação, enquanto a segunda etapa realiza classificação e regressão refinada das caixas, enfatizando a capacidade discriminativa (Cao *et al.*, 2025).

Embora alcancem alta acurácia, esses detectores têm uso limitado em aplicações de tempo real devido à sua menor velocidade e maior complexidade computacional. A Tabela 2 sintetiza a evolução arquitetural dos principais modelos dessa família, da proposta original de regiões R-CNN ao refinamento progressivo em múltiplos estágios Cascade R-CNN.

Tabela 2 – Comparativo de detectores de dois estágios

Detector	Ano	Principal contribuição
R-CNN	2014	Primeira aplicação de CNNs à detecção; uso de *selective* search para propostas de região
SPPNet	2014	Camada de *Spatial Pyramid Pooling*, eliminando a necessidade de redimensionar regiões e acelerando a inferência
Fast R-CNN	2015	Treinamento conjunto de classificação e regressão; introdução do *RoI Pooling*
Faster R-CNN	2015	*Region Proposal Network* (RPN) integrada à rede principal; *pipeline end-to-end*
Cascade R-CNN	2018	Refinamento progressivo em múltiplos estágios com limiares de *IoU* crescentes

Fonte: Adaptado de Cao *et al.* (2025) e Zou *et al.* (2023).

2.4 Detectores de Um Estágio e a Revolução do YOLO

Os detectores de um estágio prescindem da etapa de geração de propostas, prevendo diretamente as coordenadas de caixas delimitadoras e as probabilidades

de classe a partir de mapas de características multi-escala em uma única passagem pela rede, paradigma estabelecido por trabalhos seminais como o YOLO (*You Only Look Once*) (Redmon *et al.*, 2016) e o SSD (*Single Shot MultiBox Detector*) (Liu *et al.*, 2016). Essa simplificação arquitetural reduz a complexidade computacional, viabilizando o processamento em tempo real (Zou *et al.*, 2023).

A série YOLO consolidou o paradigma de um estágio por meio de iterações sucessivas que estabeleceram um padrão arquitetural sistemático em três componentes: *backbone*, *neck* e *head* (Cao *et al.*, 2025), cujas principais versões e contribuições são sintetizadas na Tabela 3.

Tabela 3 – Comparativo entre versões YOLO

Versão	Ano	Principal contribuição
Yolo v1	2016	Pipeline unificado *end-to-end*; divisão da imagem em grade com predição simultânea de caixas e classes
YOLOv2 / YOLO9000	2017	*Anchor boxes*, *batch normalization* e treinamento conjunto em múltiplos datasets
YOLOv8	2023	Cabeça desacoplada (*decoupled head*), arquitetura *anchor-free* e atribuição de rótulos alinhada à tarefa
YOLO11	2024	Módulo C2PSA (*Cross-Stage Partial with Self-Attention*) inspirado em Transformers e uso de convoluções separáveis em profundidade

Fonte: Adaptado de Cao *et al.* (2025) e Zaidi *et al.* (2022).

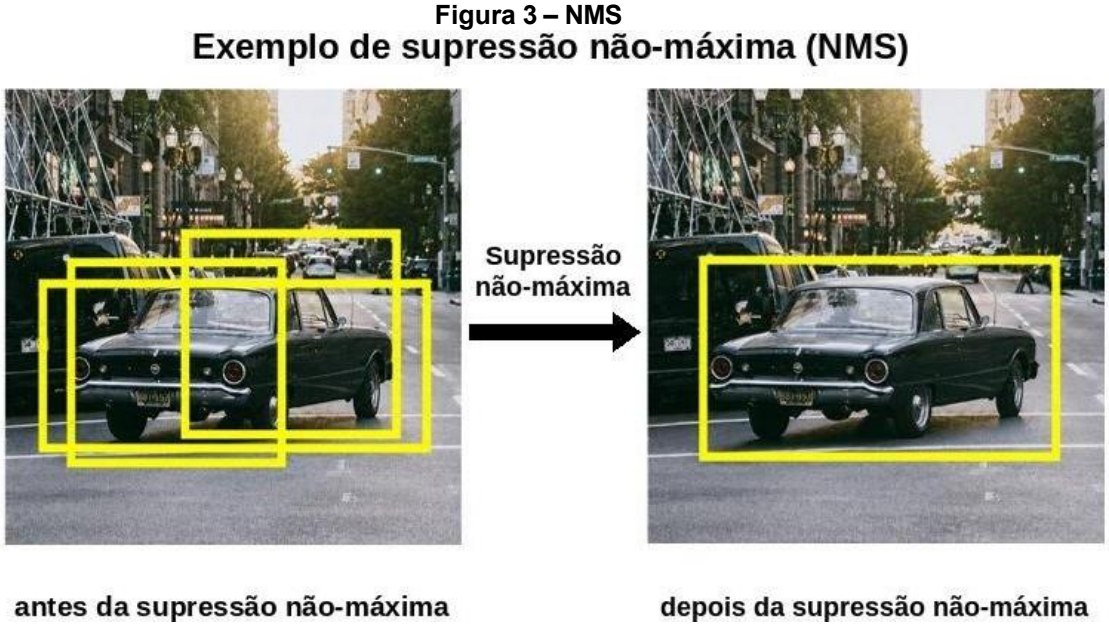
O paradigma inaugurado pelo YOLO tornou viável a detecção de objetos em tempo real utilizando *hardware* comercial comum (Zaidi *et al.*, 2022), consolidando-se como a arquitetura de detecção mais empregada em aplicações industriais e em cenários *multi-câmera*, devido à sua eficiência de inferência (Abbas *et al.*, 2025; Munsif *et al.*, 2025).

2.5 Detectores Baseados em Transformers

A *Non-Maximum Suppression* (NMS) é uma técnica de pós-processamento empregada para eliminar caixas redundantes entre detecções vizinhas com pontuações similares (Figura 3). Em geral, seleciona-se a caixa de maior pontuação

e removem-se as adjacentes cujo *Intersection over Union* (IoU) exceda um limiar predefinido (Diwan; Anirudh; Tembhurne, 2023). Embora tenha sido peça padrão dos detectores por mais de duas décadas, a NMS apresenta desvantagens: a caixa de maior pontuação pode não ser a melhor representante do objeto, pode suprimir objetos próximos em cenas densas, depende de limiares ajustados manualmente e não é diferenciável, o que impede um treinamento *end-to-end* completo (Lu *et al.*, 2025).

Figura 3 – NMS



Fonte: Adaptado de Diwan; Anirudh; Tembhurne (2023).

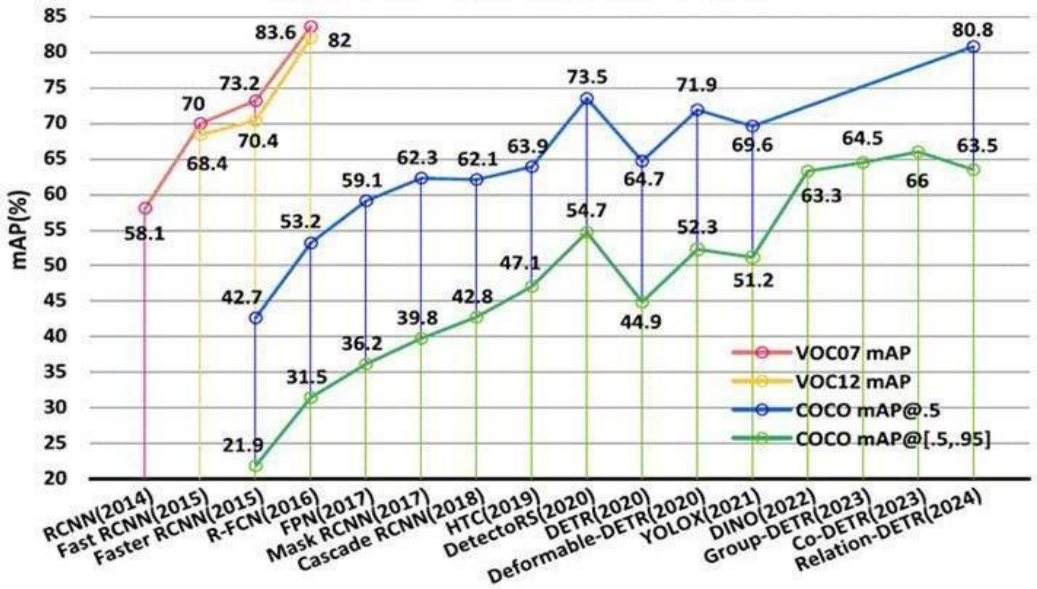
Em 2020, Carion *et al.* (2020) propuseram o DETR (*DEtection TRansformer*), que reformulou a detecção como um problema de predição direta de conjuntos (*set prediction*), divergindo das abordagens baseadas em CNN (Cao *et al.*, 2025). O modelo emprega correspondência bipartida entre as saídas do decodificador *Transformer* e os objetos de referência, dispensando caixas de âncora e a NMS.

A versão inicial do DETR apresentava convergência lenta e desempenho inferior em objetos pequenos, limitações corrigidas com o lançamento do *Deformable* DETR (Zhu *et al.*, 2021) com atenção deformável focada em regiões de interesse na imagem, e por uma família de variantes (*Conditional*-DETR, DAB-DETR, DN-DETR,

DINO, RT-DETR) que aceleraram a convergência e elevaram a acurácia (Cao *et al.*, 2025). O RT-DETR foi o primeiro detector baseado em *Transformer* em tempo real, competindo com detectores YOLO de escala comparável.

Essa mudança alcançou também os detectores convolucionais. Em 2024, o YOLOv10 tornou-se o primeiro da família YOLO a eliminar a NMS, adotando atribuições duais consistentes durante o treinamento (Wang *et al.*, 2026). O movimento é relevante para sistemas multi-câmera, nos quais a redução de latência por câmera impacta a viabilidade de processamento sincronizado em tempo real (Munsif *et al.*, 2025). A Figura 4 sintetiza a evolução da acurácia dos detectores em termos de mAP (%), considerando os *benchmarks* VOC07, VOC12, COCO mAP@0,5 e COCO mAP@[0,5:0,95], dos métodos tradicionais aos modelos baseados em *Transformers*, que hoje representam o estado da arte em precisão de detecção (Cao *et al.*, 2025).

Figura 4 – Melhora na acurácia em detecção de objetos



Fonte: Cao *et al.* (2025).

A trajetória descrita representa o estado da arte em sistemas monoculares. Contudo, a projeção bidimensional a partir de um único ponto de vista impõe limitações geométricas: objetos ocluídos, pontos cegos e ambiguidades de

profundidade permanecem irresolvíveis em uma única perspectiva (Yang; Chung; Jang, 2026). Superar essas limitações exige uma mudança de paradigma — em vez de aprimorar o detector monocular, adicionar perspectivas complementares ao sistema. Nesse contexto emergem as arquiteturas de *Multi-View Classification* (MVC), tema central desta revisão e apresentado na seção 4.

2.6 Conjuntos de Dados (*Datasets*)

Os conjuntos de dados constituem peça fundamental para o desenvolvimento de detectores, fornecendo material em larga escala para treinamento supervisionado, validação de hiperparâmetros e avaliação comparativa em condições padronizadas (Zou *et al.*, 2023).

Entre os principais *benchmarks* monoculares, destacam-se o PASCAL VOC (Everingham *et al.*, 2010) nas versões VOC2007 (~10 mil imagens, 20 categorias) e VOC2012 (~22 mil imagens, anotações ampliadas); o ILSVRC (Russakovsky *et al.*, 2015), com 1,2 milhão de imagens e 1.000 categorias para classificação em larga escala; e o MS-COCO (Lin *et al.*, 2014), com ~330 mil imagens naturais, 80 categorias e 1,5 milhão de instâncias anotadas em cenas complexas, tornando-se o padrão de fato para avaliação de detectores modernos (Cao *et al.*, 2025).

No contexto de sistemas *multi-câmera*, conjuntos específicos foram desenvolvidos com câmeras sincronizadas e anotações cruzadas entre vistas. O DukeMTMC (Ristani *et al.*, 2016) reúne 8 câmeras externas com 85 minutos de vídeo a 60 fps e 2.834 identidades para reidentificação de pedestres. O WILDTrack (Chavdarova *et al.*, 2018) conta com 7 câmeras em espaço público, focado em rastreamento em ambiente não controlado.

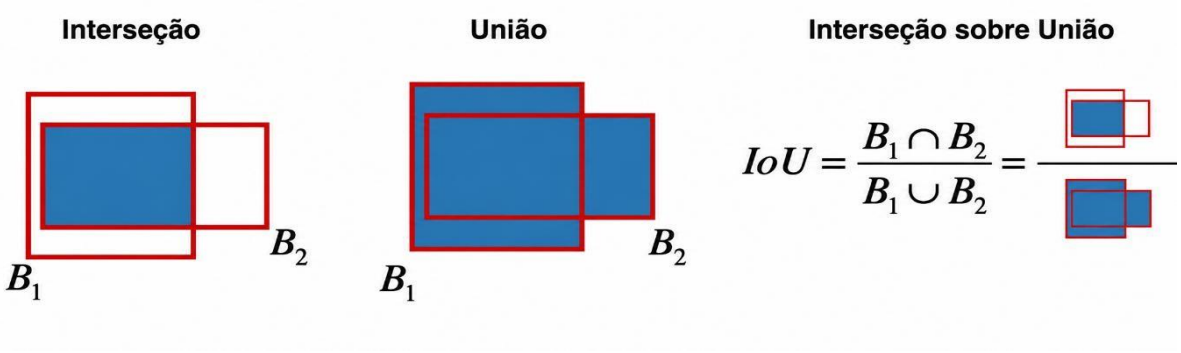
O MEVID (Davila *et al.*, 2023) abrange vídeos *multi-view* de longa duração para reidentificação de pessoas. O NTU RGB+D (Liu *et al.*, 2020) dispõe de até 106 pontos de vista com imagens RGB e de profundidade, voltado ao reconhecimento de 120 classes de ações humanas. O MultiviewX (Hou; Zheng; Gould, 2020) é um conjunto sintético com 6 câmeras calibradas para detecção de pedestres em cenários com alto grau de oclusão. Esses conjuntos viabilizam o treinamento e a avaliação de tarefas

como rastreamento e reidentificação inter-câmera (Amosa *et al.*, 2023).

2.6 Métricas de avaliação

A avaliação de detectores apoia-se em métricas que capturam a qualidade da classificação e a precisão da localização espacial. A *Intersection over Union (IoU)* exemplificada na Figura 5, quantifica a sobreposição entre a caixa predita e a de referência pela razão entre as áreas de interseção e de união; uma detecção é considerada correta quando o *IoU* excede um limiar pré-definido, sendo 0,5 o valor mais adotado (Padilla; Netto; Da Silva, 2020).

Figura 5 – IoU



Fonte: Diwan; Anirudh; Tembhurne (2023).

Dessa classificação binária derivam os conceitos de *true positive* (TP), *false positive* (FP) e *false negative* (FN), que sustentam o cálculo de precisão e *recall,* mostradas na Tabela 4 (Padilla; Netto; Da Silva, 2020).

Tabela 4 – Matriz de confusão

		VALOR PREDITO	
		SIM	**NÃO**
REAL	**SIM**	Verdadeiro positivo (TP)	Falso negativo (FN)
	NÃO	Falso positivo (FN)	Verdadeiro negativo (TN)

Fonte: Adaptado de Zaidi *et al.* (2022).

O F1-*score* corresponde à média harmônica entre precisão e *recall*, oferecendo um valor único quando há interesse em equilibrar as duas medidas, sobretudo em cenários com classes desbalanceadas (Padilla; Netto; Da Silva, 2020). Já a *Average Precision* (AP) sintetiza essa relação pela área sob a curva precisão *versus recall*, podendo ser calculada por interpolação em 11 pontos, como proposto pelo PASCAL VOC (Everingham *et al.*, 2010), ou pela interpolação sobre todos os pontos, variante adotada por desafios subsequentes (Padilla; Netto; Da Silva, 2020).

A precisão, recall e o F1 Score podem ser calculados conforme Zaidi et al. (2022):

$$Precisão = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

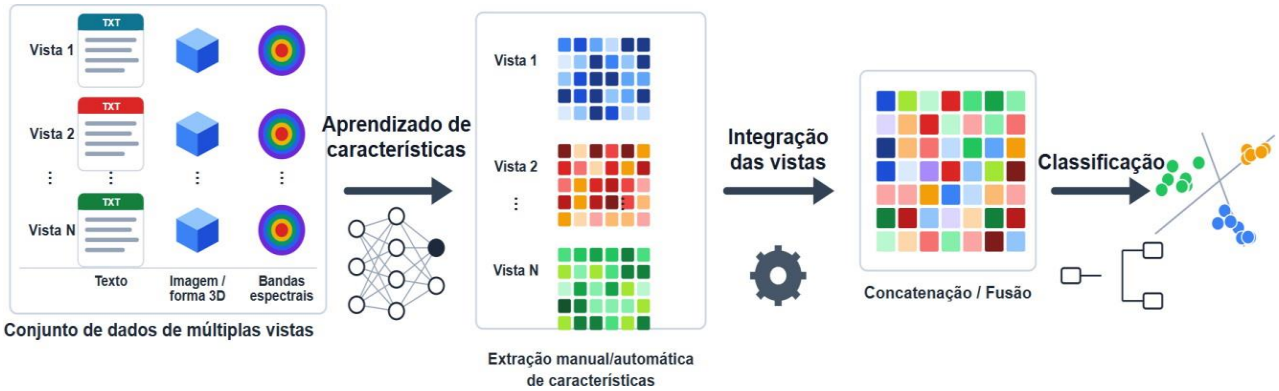
$$F1\ Score = 2 \times \frac{Precisão \times Recall}{Precisão + Recall} \quad (3)$$

2.7 Arquiteturas de Aprendizado Profundo para Múltiplas Câmeras

A *Multi-View Classification* (MVC) consiste em categorizar instâncias em classes predefinidas a partir de informações provenientes de múltiplas perspectivas. Diferentemente dos sistemas de visão única, que dependem de uma única projeção bidimensional (2D), a abordagem *multi-view* integra evidências complementares capturadas sob diferentes ângulos, conferindo maior robustez frente a oclusões, variações de iluminação e ambiguidades geométricas (Perez Soler *et al.*, 2025; Yang; Chung; Jang, 2026).

Berahmand *et al.* (2025) descrevem o processo em três etapas: (i) aprendizado de características, em que representações discriminativas são extraídas de cada vista; (ii) integração de vistas, que combina as informações entre perspectivas; e (iii) classificação, na qual o rótulo final é atribuído à instância, conforme ilustrado na Figura 6.

Figura 6 – Etapas do *Multi-View Classification*



Fonte : Berahmand *et al.* (2025).

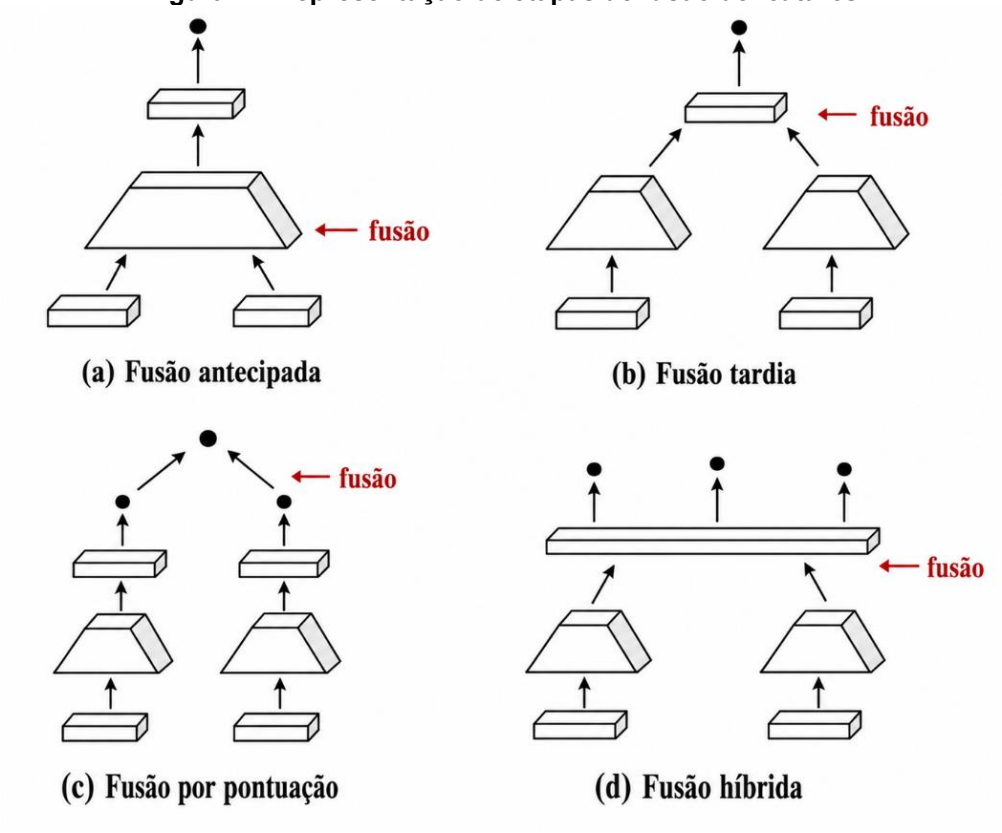
Yu *et al.* (2025) categorizam o aprendizado *multi-view* em quatro grupos: classificação supervisionada, classificação semi-supervisionada, agrupamento e agrupamento semi-supervisionado. Na vertente supervisionada, foco desta revisão, o objetivo é aprender representações que explorem tanto o consenso quanto a complementaridade entre as vistas para maximizar a acurácia de classificação (Yu *et al.*, 2025).

Quando as câmeras ocupam posições fixas e conhecidas, as imagens são capturadas a partir de um conjunto predefinido de pontos de vista (por exemplo, frente, lateral e traseira), garantindo número e angulação constantes entre amostras; tal arranjo se distingue dos cenários não estruturados (*unstructured*), nos quais as vistas provêm de posições arbitrárias e variáveis (Black; Souvenir, 2024).

2.8 Tipos de fusão em *multi-view classification*

A estratégia de fusão define como as informações provenientes das diferentes vistas são integradas ao longo do *pipeline* de classificação. Berahmand *et al.* (2025) e Black e Souvenir (2024) caracterizam quatro abordagens principais: fusão antecipada (*early fusion*), fusão tardia (*late fusion*), fusão por pontuação (*score fusion*) e fusão híbrida (*hybrid fusion*). Cada abordagem apresenta compensações específicas quanto ao nível de abstração em que a integração ocorre, à capacidade de capturar interações entre vistas e ao custo computacional (Yu *et al.*, 2025). A Figura 7 sintetiza essas quatro estratégias de forma esquemática.

Figura 7 – Representação de etapas de fusão de *features*



Fonte: Adaptado de Black e Souvenir (2024).

Na fusão antecipada, os dados brutos ou mapas de características de baixo nível de cada vista são concatenados antes do processamento profundo, permitindo que o modelo trate o conjunto fundido como entrada única. Yu *et al.* (2025) observam que essa abordagem pressupõe contribuição equivalente entre vistas e treina um único modelo com função objetivo conjunta. Sua principal vantagem é permitir o aprendizado de interações entre vistas desde as camadas iniciais. Entretanto, como a fusão ocorre antes da extração de características discriminativas, informações irrelevantes podem ser incorporadas prematuramente, diluindo os atributos relevantes (Yang; Chung; Jang, 2026).

Na fusão tardia, características de alto nível são aprendidas de forma independente para cada vista e combinadas apenas nos estágios finais do *pipeline* (Black; Souvenir, 2024). A integração pode ocorrer por concatenação simples (Wang

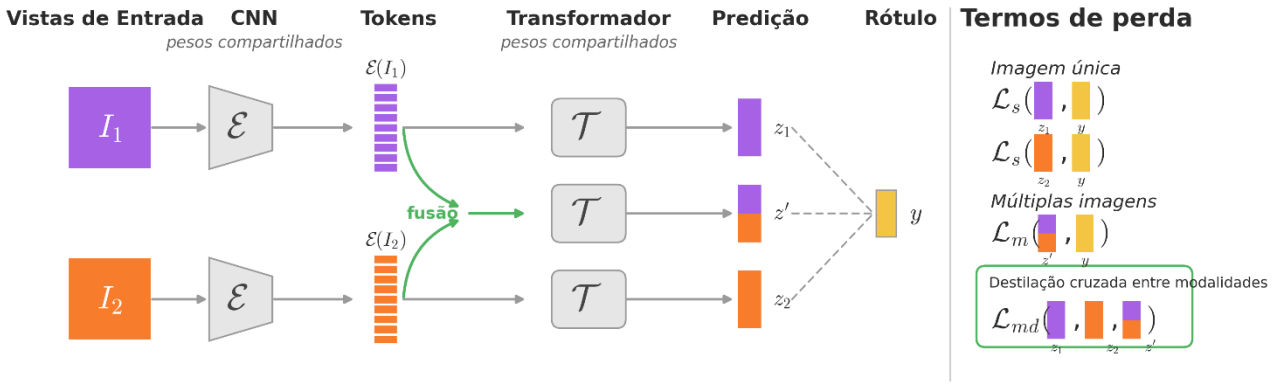
et al., 2026), *pooling* (Su *et al.*, 2015), grafos (Wei; Yu; Sun, 2020) ou *Transformers*.

Yu *et al.* (2025) destacam a analogia com aprendizado por *ensemble*, em que modelos especializados por vista têm suas predições combinadas por média, votação ou *stacking*. A vantagem reside na especialização independente de cada vista, sem contaminação cruzada; a limitação é a perda de interações de baixo nível entre perspectivas.

Caracterizada por Black e Souvenir (2024) como uma forma extrema de fusão tardia, a fusão por pontuação atribui um classificador completo a cada vista e combina suas distribuições de probabilidade elemento a elemento para produzir a predição final. Yang, Chung e Jang (2026) descrevem essa combinação por meio de *pooling* das distribuições de classe ou de modelos probabilísticos — como a distribuição de Dirichlet (Yang; Chung; Jang, 2026) —, que incorporam incerteza ao processo. A distinção em relação à fusão tardia convencional reside no grau de independência entre os modelos: enquanto a fusão tardia pode combinar características antes da camada classificadora, a fusão por pontuação opera exclusivamente sobre as saídas finais de classificadores independentes.

A fusão híbrida integra elementos das estratégias anteriores, combinando ramos de processamento independentes por vista com módulos adicionais que modelam interações entre vistas e aumentam a consistência das predições (Berahmand *et al.*, 2025). Essa abordagem busca preservar a especialização de cada vista, ao mesmo tempo em que captura interações classe a classe e vista a vista, frequentemente ausentes em fusões puramente independentes. A Figura 8 exemplifica essa estratégia por meio da arquitetura proposta por Yang, Chung e Jang (2026), na qual duas vistas (I_1 e I_2) são processadas por CNNs com pesos compartilhados, fundidas antes dos *Transformers* e otimizadas conjuntamente por funções de perda de imagem única, multi-imagem e destilação mútua entre vistas.

Figura 8 — Arquitetura de fusão híbrida com destilação mútua entre vistas



Fonte: Adaptado de Yang, Chung e Jang (2026).

3 METODOLOGIA

Esta revisão sistemática da literatura seguiu as diretrizes do PRISMA 2020 (PAGE *et al*., 2021), com protocolo previamente definido quanto à questão de investigação, bases consultadas, estratégia de busca, critérios de elegibilidade e procedimentos de triagem e extração.

3.1 Questão de Investigação

A questão central foi estruturada com base na estratégia PICO (PAGE *et al*., 2021), adaptada ao contexto de ciências da computação e engenharia de sistemas, conforme detalhado na Tabela 5:

> *"Quais arquiteturas de aprendizado profundo têm sido utilizadas para detecção de objetos em sistemas de visão com múltiplas câmeras"*

Tabela 5 – Estrutura PICO da questão de investigação

Elemento	Descrição	Aplicação nesta revisão
P – Contexto	Domínio de aplicação	Sistemas industriais com múltiplas câmeras: esteiras, linhas de produção e inspeção de qualidade
I – Intervenção	Tecnologia avaliada	CNNs, família YOLO, Transformers (DETR, ViT), arquiteturas híbridas e sistemas *end-to-end*

C – Comparação	Alternativas	Fusão tardia vs. precoce; detectores monoculares vs. multi-câmera; YOLO vs. Transformers
O – Desfecho	Métricas de avaliação	mAP, AP@IoU 0,5–0,95, FPS, precisão, revocação e F1-score

Fonte: Autores (2026).

3.2 Estratégia de Busca

As buscas foram executadas em maio de 2026 na base *ScienceDirect*, adotada como fonte primária por sua abrangência na área de visão computacional e aprendizado profundo. Durante o processo, identificou-se que parte dos artigos retornados exigia pagamento para acesso ao texto completo, o que inviabilizaria sua leitura e eventual inclusão na revisão. Diante dessa limitação, optou-se por buscar os mesmos artigos em bases complementares de acesso aberto *SpringerLink, Wiley Online Library e IET Digital Library*, priorizando, nesses casos, as versões disponíveis sem restrição.

Esse critério, embora não previsto no protocolo inicial, foi adotado de forma sistemática ao longo de toda a etapa de busca, garantindo que a seleção dos artigos fosse determinada pela relevância temática e não pela disponibilidade de acesso.

Os termos foram organizados em blocos temáticos derivados do PICO, combinados por operadores booleanos:

> *("deep learning" OR "convolutional neural network" OR CNN OR "vision transformer" OR ViT) AND ("multi-view classification" OR "cross-view" OR "multi-camera" OR "multiple cameras") AND ("object detection" OR "object recognition" OR "visual inspection")*

Cada base recebeu adaptações nos operadores conforme suas características técnicas. Após aplicação dos filtros de aderência temática e leitura preliminar dos títulos, foram separados 56 artigos para a etapa de triagem.

3.3 Critérios de Elegibilidade

O período de cobertura foi definido entre 2015 e 2026, abrangendo publicações

recentes sobre detectores baseados em aprendizado profundo. A Tabela 6 sintetiza os critérios adotados.

Tabela 6 – Critérios de elegibilidade

Dimensão	Inclusão	Exclusão
Tema	Arquiteturas de aprendizado profundo para detecção de objetos em sistemas mono ou multi-câmera	Estudos sem detecção de objetos ou sem aprendizado profundo como abordagem central
Contexto	Aplicações industriais, inspeção visual, linhas de produção ou vigilância multi-câmera	Domínios clínicos, biológicos ou sem relação com contexto industrial
Tipo de estudo	Artigos experimentais, comparativos e revisões com avaliação quantitativa	Editoriais, resumos sem texto completo e artigos de opinião
Período	2022–2026	Publicações anteriores a 2022, exceto marcos fundadores
Idioma	Inglês e português	Demais idiomas
Métricas	Ao menos uma métrica quantitativa: mAP, AP, FPS, precisão, revocação ou F1	Ausência de avaliação empírica de desempenho

Fonte: Os autores (2026).

3.4 Triagem, Seleção e Extração dos Dados

A partir dos 56 artigos identificados, nenhum duplicado foi detectado. Na triagem por títulos, resumos e palavras-chave, 18 artigos foram excluídos por não apresentarem aderência ao tema, por não abordarem sistemas multivisão, arquiteturas de aprendizado profundo para detecção de objetos ou contexto industrial.

Os 38 artigos restantes foram submetidos à leitura completa e todos atenderam aos critérios de inclusão, compondo o corpus elegível da revisão. Destes, 10 estudos foram selecionados para extração de dados e análise comparativa aprofundada, por apresentarem arquitetura, estratégia de fusão e métricas de avaliação claramente reportadas e diretamente relacionadas à questão de investigação; os demais 28 artigos compuseram a fundamentação teórica e contextual da revisão (Seção 2).

A extração de dados foi apoiada pelo Mendeley e compreendeu: título, autores,

ano, base de origem, objetivo, arquitetura utilizada, contexto de aplicação, configuração de câmeras, métricas reportadas, contribuições e limitações apontadas pelos autores. A análise foi conduzida de forma qualitativa, priorizando estudos com relação direta à questão de investigação e permitindo uma base comparativa entre as arquiteturas identificadas na literatura.

4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

A literatura revisada converge para uma arquitetura de três etapas sequenciais: detecção, fusão de representações e classificação. Essa estrutura segue os trabalhos estudados, com graus diferentes de implementação entre eles.

Um aspecto central que diferencia os trabalhos é a relação entre as diferentes vistas: em alguns sistemas, cada câmera resolve uma tarefa diferente e independente, de modo que a fusão agrega resultados de especialistas paralelos. Em outros, todas as câmeras observam o mesmo objeto com o mesmo objetivo classificatório, e a fusão combina perspectivas complementares para responder a uma única pergunta.

O segundo modelo é o que Berahmand *et al.* (2025) e Yu *et al.* (2025) formalizam como MVC estruturada, sustentada pelos princípios de consenso — em que as vistas concordam sobre características compartilhadas — e de complementaridade — em que cada vista revela informações que as demais não capturam isoladamente. Este modelo é demonstrado nos trabalhos de Zheng *et al.* (2023), Bhavanasi *et al.* (2025), Perez Soler *et al.* (2025) e Black e Souvenir (2024).

No *pipeline* completo de classificação *multi-view*, a primeira etapa consiste em localizar e recortar o objeto de interesse em cada câmera individualmente. Nesse estágio, os modelos de detecção de objetos operam de forma independente sobre cada fluxo de câmera, gerando as regiões de interesse que alimentam o módulo de fusão.

A maior parte dos trabalhos teóricos revisados parte de vistas já segmentadas

ou pré-renderizadas, omitindo a etapa de detecção. Su *et al*. (2015), Qi *et al*. (2016) e Wei, Yu e Sun (2020) utilizam projeções de modelos tridimensionais, enquanto Black e Souvenir (2024) e Yang, Chung e Jang (2026) avaliam conjuntos onde as imagens já correspondem à região de interesse.

Zheng *et al.* (2023) representam o caso mais explícito e mensurável desse estágio, com o Faster R-CNN com *backbone* ResNet-50 aplicado de forma independente a cada imagem bruta capturada pela plataforma *multi-câmera*, gerando caixas delimitadoras com *scores* de confiança para cada detecção em cada vista separadamente. Bhavanasi *et al.* (2025) capturam vistas de peças metálicas em braço robótico calibrado, e Perez Soler *et al.* (2025) extraem silhuetas geométricas a partir de câmeras fixas com posições conhecidas. Kulkarni, Jha e Kachhoria (2026) tornam a detecção ainda mais explícita, aplicando o detector YOLOv8 para localizar componentes e segmentar defeitos com 11 câmeras sincronizadas.

Em aplicações de *early fusion*, destaca-se o trabalho de Bhavanasi *et al.* (2025). O modelo proposto MV-UNet emprega um T-block baseado em homografia para alinhar espacialmente os mapas de características de câmeras calibradas antes da fusão, transformando perspectivas distintas em um espaço de referência comum. O sistema obtém F1 = 0,942 na detecção de defeitos em placas metálicas, superando variantes de fusão tardia (F1 = 0,936) e classificação monocular (F1 = 0,901) nas mesmas condições.

Qi *et al.* (2016) exploram variante da fusão antecipada em domínio 3D com o MVCNN-MultiRes, combinando renderizações multi-resolução antes do processamento convolucional profundo e atingindo 93,8% no *ModelNet40*.

Dentre as abordagens de *early fusion* analisadas, o alinhamento geométrico prévio via homografia mostrou-se o diferencial mais relevante, sendo especialmente adequado a cenários industriais com câmeras fixas e calibradas, onde a transformação entre vistas é determinística e estável.

Em aplicações de *late fusion*, Su *et al.* (2015) apresentam o MVCNN, que aplica uma CNN compartilhada a cada vista e realiza *view-pooling* elemento a elemento

antes de uma segunda CNN de classificação, atingindo 90,1% no *ModelNet40*. Wang *et al.* (2015) estendem a fusão tardia ao contexto multimodal com o MMSS, mantendo redes separadas por modalidade e extraindo tanto características compartilhadas quanto específicas, atingindo 88,5% no *RGB-D Object Dataset*.

Wei, Yu e Sun (2020) incorporam estrutura relacional entre vistas com *o View-GCN* (*Graph Convolutional Network*), onde as vistas são nós de um grafo com arestas por proximidade no espaço de câmeras, atingindo 97,6% no *ModelNet40*. Zheng *et al.* (2023) elevam a acurácia global de detecção de 76,28% para 96,98% em flores, de 71,64% para 99,09% em frutos imaturos e de 69,81% para 97,17% em frutos maduros, quantificando diretamente o ganho do princípio de complementaridade.

Dentre as abordagens de *late fusion*, o *View*-GCN demonstrou o maior ganho ao incorporar relações estruturais entre vistas por meio de grafos, aspecto ausente nas demais arquiteturas dessa categoria.

Em trabalhos baseados em *score fusion*, Kulkarni, Jha e Kachhoria (2026) adotam *max-pooling* sobre os *scores* de confiança das câmeras, onde o *score* final de cada tarefa é o máximo entre os *scores* produzidos pelas câmeras designadas. Destaca-se também o trabalho de Black e Souvenir (2024) que incorpora a fusão por pontuação como componente do MV-HFMD, e Yang, Chung e Jang (2026), que estendem essa abordagem com ponderação por incerteza, atribuindo menor peso às câmeras com maior incerteza na predição.

A *score fusion* mostrou-se mais adequada a sistemas com câmeras especializadas em tarefas distintas, nos quais a independência entre os classificadores é uma vantagem, não uma limitação.

Buscando aproveitar as vantagens de cada estratégia, os trabalhos mais recentes combinam dois ou mais mecanismos em seu *pipeline* (Berahmand *et al.*, 2025). Black e Souvenir (2024) propõem o MV-HFMD, combinando *late fusion* com *score fusion* e destilação mútua, superando abordagens especializadas em conjuntos médicos (CheXpert), forenses (Hotels-8k) e geográficos (GLDv2).

Yang, Chung e Jang (2026) levam essa direção adiante com o HMDMV,

adotando *backbone* híbrido CNN-*Transformer* com pesos compartilhados e destilação mútua hierárquica, superando o MV-HFMD em 3,48 pontos percentuais. Perez Soler *et al.* (2025) propõem o VCFF, combinando descritores de Fourier com *Dynamic Time Warping*, atingindo AP = 1,0 com oito câmeras fixas em inspeção industrial.

A fusão híbrida concentrou os melhores resultados entre todos os trabalhos revisados, evidenciando que a combinação de estratégias complementares supera abordagens isoladas, especialmente quando associada a mecanismos de atenção entre vistas.

Independentemente da estratégia de fusão adotada, o pipeline encerra-se em uma etapa de decisão que converte as representações fundidas no resultado final, e essa etapa assume formas distintas conforme a arquitetura. Nos sistemas voltados à classificação supervisionada, predominam camadas densas seguidas de softmax (Su et al., 2015; Wei; Yu; Sun, 2020), classificadores softmax aplicados sobre os tokens do Transformer (Black; Souvenir, 2024; Yang; Chung; Jang, 2026) e redes leves como a EfficientNet (Kulkarni; Jha; Kachhoria, 2026).

Nos sistemas de rastreamento, por outro lado, a decisão não é uma classificação, mas um agrupamento por similaridade: Zheng et al. (2023) empregam o DBSCAN (Density-Based Spatial Clustering of Applications with Noise) para associar, entre as vistas, as detecções correspondentes a um mesmo objeto. A Tabela 7 sintetiza os principais trabalhos revisados, organizados por estratégia de fusão.

Tabela 7 – Comparativo dos trabalhos revisados

Trabalho	Arquitetura	Fusão	Dataset
Su et al. (2015)	MVCNN	Tardia	ModelNet40
Qi et al. (2016)	MVCNN-MultiRes	Antecipada	ModelNet40
Wang et al. (2015)	MMSS	Híbrida	RGB-D Object Dataset

Wei; Yi; Sun (2020)	View-GCN	Tardia	ModelNet40
Zheng et al. (2023)	Faster R-CNN + FaceNet	Tardia	Agrícola
Bhavanasi et al. (2025)	MV-UNet	Antecipada	Placas metálicas
Kulkarni et al. (2026)	YOLOv8 + EfficientNet	Pontuação	Inspeção veicular
Perez Soler et al. (2025)	VCFF	Híbrida	Inspeção industrial
Black; Souvenir (2024)	MV-HFMD	Híbrida	CheXpert/Hotels-8k/GLDv2
Yang et al. (2026)	HMDMV	Híbrida	Múltiplos

Fonte: Autores (2026).

Esses padrões reforçam que a escolha da estratégia de fusão deve acompanhar a relação entre as vistas, especialistas paralelos favorecem fusões simples, como a por pontuação, enquanto vistas complementares exigem mecanismos capazes de explorar relações inter-vista, como destacado pela fusão híbrida. A fusão híbrida concentra os melhores resultados, e a incorporação de *Transformers* como *backbone* compartilhado (Yang; Chung; Jang, 2026) ou mecanismo de atenção inter-vista (Wei; Yu; Sun, 2020) representa o vetor de maior impacto recente.

Em resposta à questão de investigação — *quais arquiteturas de aprendizado profundo têm sido utilizadas para detecção de objetos em sistemas de visão com múltiplas câmeras* —, os trabalhos revisados evidenciam que não existe uma arquitetura universalmente superior. CNNs com *backbones* como ResNet e VGG predominam na extração de características, enquanto detectores da família YOLO lideram nas aplicações industriais pela eficiência de inferência. Arquiteturas baseadas em *Transformers* emergem como tendência, sobretudo em sistemas que exigem modelagem de dependências globais entre vistas.

A estratégia de fusão híbrida, associada a mecanismos de destilação mútua e atenção cruzada entre vistas, concentra os melhores resultados recentes, indicando que a convergência entre paradigmas convolucionais e baseados em atenção representa a direção

mais promissora para sistemas multi-câmera estruturados.

5 CONSIDERAÇÕES FINAIS

Este trabalho identificou, por meio de revisão sistemática conduzida segundo o PRISMA 2020, as arquiteturas de aprendizado profundo aplicadas à classificação *cross-view* em sistemas multi-câmera com vistas estruturadas. Foram selecionados dez estudos publicados entre 2015 e 2026, analisados em três eixos: arquiteturas, estratégias de fusão e contextos de aplicação.

As arquiteturas identificadas dividem-se em três grupos: *backbones* convolucionais com fusão entre vistas (MVCNN, MVCNN-MultiRes, MMSS, MV-UNet), arquiteturas baseadas em grafos (View-GCN) e arquiteturas híbridas CNN-*Transformer* (MV-HFMD, HMDMV, VCFF). As híbridas concentraram os melhores resultados nos trabalhos recentes.

Quanto à fusão, a tardia predominou nos estudos mais antigos da amostra, como em Su *et al*. (2015), Wei; Yu; Sun (2020) e Zheng *et al*. (2023), pela compatibilidade com infraestruturas existentes. A antecipada apareceu em Qi *et al.* (2016) e Bhavanasi *et al.* (2025), mostrando-se vantajosa quando há câmeras sincronizadas e alinhamento geométrico conhecido, condição que permite alinhar as vistas antes da extração de características.

A híbrida, presente em Wang *et al.* (2015), Black e Souvenir (2024), Perez Soler *et al.* (2025) e Yang, Chung e Jang (2026), concentrou os melhores resultados recentes; mecanismos de atenção cruzada (*cross-attention*) e codificações posicionais de câmera surgem nesses trabalhos como a direção mais promissora.

Contudo,não existe arquitetura universalmente superior para sistemas multi-câmera. A diversidade de domínios, configurações de câmeras e conjuntos de dados exige que a escolha seja orientada pelos requisitos do cenário,como latência,acurácia, infraestrutura computacional e disponibilidade de dados anotados em múltiplas vistas.

Para trabalhos futuros, recomenda-se: (i) curadoria de *datasets* multi-câmera para inspeção industrial; (ii) investigação de Transformers com atenção cruzada em cenários de alta velocidade, comparando-os às variantes YOLO; (iii) desenvolvimento de protocolos de *benchmark* com métricas de acurácia, latência e robustez; e (iv) exploração de transferência de aprendizado e *few-shot learning* para reduzir a dependência de grandes volumes de dados.

REFERÊNCIAS

ABBAS, Kumail et al. Vision transformer-based multi-camera multi-object tracking framework for dairy cow monitoring. **Smart Agricultural Technology**, [s. l.], v. 12, p. 1-13, 2025.

AKELE, Yamral Kassanew; GURAU, Vladimir. A critical survey of machine learning-powered vision inspection in semiconductors manufacturing. **Array**, v. 30, p. 100888, 2026. DOI: 10.1016/j.array.2026.100888.

AMOSA, Temitope Ibrahim et al. Multi-camera multi-object tracking: a review of current trends and future advances. **Neurocomputing**, [s. l.], v. 552, p. 126558, 2023. DOI: 10.1016/j.neucom.2023.126558.

BERAHMAND, Kamal et al. A comprehensive survey on multi-view classification: methods, applications, and challenges. **ACM Transactions on Intelligent Systems and Technology**, [s. l.], v. 16, n. 6, 2025.

BHATT, P. M. et al. Image-based surface defect detection using deep learning: a review. **Journal of Computing and Information Science in Engineering**, [s. l.], v. 21, n. 4, art. 040801, 2021.

BHAVANA, N. et al. Cognizance through convolution: a deep learning approach for emotion recognition via convolutional neural networks. **Procedia Computer Science**, v. 259, p. 1336-1345, 2025. DOI: 10.1016/j.procs.2025.04.088.

BHAVANASI, Geethika et al. Enhanced vision-based quality inspection: a multiview artificial intelligence framework for defect detection. **Sensors**, [s. l.], v. 25, n. 6, 2025.

BLACK, Samuel; SOUVENIR, Richard. Multi-view classification using hybrid fusion and mutual distillation. In: IEEE WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION (WACV), 2024, [s. l.]. **Proceedings** [...]. IEEE, 2024. p. 269-279.

CAO, Jinfeng et al. Object detection based on CNN and Vision-Transformer: a survey. **IET Computer Vision**, [s. l.], v. 19, n. 1, p. 1-30, 2025.

CARION, Nicolas et al. End-to-end object detection with Transformers. In: EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), 2020, [s. l.]. **Proceedings** [...]. Cham: Springer, 2020. p. 213-229. (Lecture Notes in Computer Science, v. 12346).

CHAVDAROVA, Tatjana et al. WILDTRACK: a multi-camera HD dataset for dense unscripted pedestrian detection. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2018, [s. l.]. **Proceedings** [...]. IEEE, 2018. p. 5030-5039.

DALAL, Navneet; TRIGGS, Bill. Histograms of oriented gradients for human detection. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2005, [s. l.]. **Proceedings** [...]. IEEE, 2005. p. 886-893.

DAVILA, Daniel et al. MEVID: multi-view extended videos with identities for video person re-identification. In: IEEE WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION (WACV), 2023, [s. l.]. **Proceedings** [...]. IEEE, 2023. p. 1634-1643.

DEMIR, Necati. **Understanding Image Classification: Convolutional Neural Networks**. *Medium*, 6 jan. 2022. Disponível em: https://blog.demir.io/understanding-image-classification-cnn-a3c0606445b6. Acesso em: 22 jun. 2026.

DIWAN, Tausif; ANIRUDH, G.; TEMBHURNE, Jitendra V. Object detection using YOLO: challenges, architectural successors, datasets and applications. **Multimedia Tools and Applications**, [s. l.], v. 82, n. 6, p. 9243-9275, 2023.

DONG, Bowen; SHENG, Lu. Bags of tricks for learning depth and camera motion from monocular videos. **Virtual Reality & Intelligent Hardware**, v. 1, n. 5, p. 500-510, 2019. DOI: 10.1016/j.vrih.2019.09.004.

DOSOVITSKIY, Alexey et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 9., 2021, [s. l.]. **Proceedings** [...]. ICLR, 2021.

EVERINGHAM, Mark et al. The Pascal Visual Object Classes (VOC) challenge. **International Journal of Computer Vision**, [s. l.], v. 88, n. 2, p. 303-338, 2010.

FELZENSZWALB, Pedro F. et al. Object detection with discriminatively trained part-based models. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [s. l.], v. 32, n. 9, p. 1627-1645, 2010.

GIRSHICK, Ross et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2014, [s. l.]. **Proceedings** [...]. IEEE, 2014. p. 580-587.

GRIGORESCU, S. et al. A survey of deep learning techniques for autonomous driving. **Journal of Field Robotics**, [s. l.], v. 37, n. 3, p. 362-386, 2020. DOI: 10.1002/rob.21918.

HOU, Yunzhong; ZHENG, Liang; GOULD, Stephen. Multiview detection with feature perspective transformation. In: EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), 2020, [s. l.]. **Proceedings** [...]. Cham: Springer, 2020. p. 1-18. (Lecture Notes in Computer Science, v. 12352).

HÜTTEN, Nils et al. Deep learning for automated visual inspection in manufacturing and maintenance: a survey of open-access papers. **Applied System Innovation**, [s. l.], v. 7, n. 1, art. 11, 2024.

JAVAID, M. et al. Exploring impact of Industry 4.0 in machine vision technology. **Materials Today: Proceedings**, [s. l.], 2022.

KAMALI, Masoud et al. Training-free open-vocabulary monocular 3D object detection for industrial assets. **Automation in Construction**, v. 184, p. 106841, 2026. DOI: 10.1016/j.autcon.2026.106841.

KHAN, Asharul Islam; AL BADI, Ali; ALQAHTANI, Mohammed. Explainable artificial intelligence for computer vision and quantum machine learning. **Procedia Computer Science**, v. 258, p. 3723-3730, 2025. DOI: 10.1016/j.procs.2025.04.627.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. ImageNet classification with deep convolutional neural networks. **Communications of the ACM**, [s. l.], v. 60, n. 6, p. 84-90, 2012.

KULKARNI, Yash; JHA, Raman; KACHHORIA, Renu. **Multi-view camera system for variant-aware autonomous vehicle inspection and defect detection**. [S. l.]: arXiv, 2026. Disponível em: http://arxiv.org/abs/2509.26454. Acesso em: 18 maio 2026.

LI, Jiapan et al. MIF-YOLO: an enhanced YOLO with multi-source image fusion for autonomous dead chicken detection. **Smart Agricultural Technology**, [s. l.], v. 12, art. 101104, 2025.

LIN, T.-Y. et al. Microsoft COCO: common objects in context. In: EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), 2014, Zurique. **Proceedings** [...]. Cham: Springer, 2014. p. 740-755.

LIU, Jun et al. NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [s. l.], v. 42, n. 10, p. 2684-2701, 2020.

LIU, Wei et al. SSD: single shot MultiBox detector. In: EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), 2016, [s. l.]. **Proceedings** [...]. Cham: Springer, 2016. p. 21-37. (Lecture Notes in Computer Science, v. 9905).

LU, Zhengyang et al. **Differentiable NMS via Sinkhorn matching for end-to-end fabric defect detection**. [S. l.]: arXiv, 2025. Disponível em: http://arxiv.org/abs/2505.07040. Acesso

em: 18 maio 2026.

MUNSIF, Muhammad et al. **Multi camera connected vision system with multi view analytics: a comprehensive survey**. [S. l.]: arXiv, 2025. Disponível em: http://arxiv.org/abs/2510.09731. Acesso em: 18 maio 2026.

NASIR, Muhammad Fahad et al. Multi-camera fusion and bird-eye view location mapping for deep learning-based cattle behavior monitoring. **Artificial Intelligence in Agriculture**, [s. l.], v. 15, n. 4, p. 724-743, 2025. DOI: 10.1016/j.aiia.2025.06.001.

PADILLA, Rafael; NETTO, Sergio L.; DA SILVA, Eduardo A. B. A survey on performance metrics for object-detection algorithms. In: INTERNATIONAL CONFERENCE ON SYSTEMS, SIGNALS AND IMAGE PROCESSING (IWSSIP), 27., 2020, Niterói. **Proceedings** [...]. IEEE, 2020. p. 237-242.

PAGE, Matthew J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. **BMJ**, v. 372, n. n71, 2021. DOI: 10.1136/bmj.n71.

PEREZ SOLER, Javier et al. Object-specific multiview classification through view-compatible feature fusion. **Sensors**, [s. l.], v. 25, n. 13, p. 1-23, 2025.

QI, Charles R. et al. Volumetric and multi-view CNNs for object classification on 3D data. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2016, [s. l.]. **Proceedings** [...]. IEEE, 2016. p. 5648-5656.

RADHAKRISHNAN, Jayabhaduri et al. Image processing based machine vision system for datum setting in computer numeric control machine. **Procedia Computer Science**, v. 263, p. 463-470, 2025. DOI: 10.1016/j.procs.2025.07.056.

REDMON, Joseph et al. You only look once: unified, real-time object detection. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2016, [s. l.]. **Proceedings** [...]. IEEE, 2016. p. 779-788.

REICHENSTEIN, Tobias et al. Role-based manufacturing systems: a concept towards intelligent and adaptive automation in manufacturing. **Procedia CIRP**, v. 134, p. 837-842, 2025. DOI: 10.1016/j.procir.2025.02.187.

RISTANI, Ergys et al. Performance measures and a data set for multi-target, multi-camera tracking. In: EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), 2016, [s. l.]. **Proceedings** [...]. Cham: Springer, 2016. p. 17-35. (Lecture Notes in Computer Science, v. 9914).

RUSSAKOVSKY, O. et al. ImageNet large scale visual recognition challenge. **International Journal of Computer Vision**, [s. l.], v. 115, n. 3, p. 211-252, 2015.

SARMENTO, Hugo; ORDOÑEZ SAAVEDRA, Néstor; ROSADO, António. **Revisão sistemática da literatura** [recurso eletrônico]. Mossoró: Edições UERN, 2024. 92 p. ISBN 978-85-7621-478-6.

SU, Hang et al. Multi-view convolutional neural networks for 3D shape recognition. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2015, [s. l.]. **Proceedings** [...]. IEEE, 2015. p. 945-953.

VASWANI, Ashish et al. Attention is all you need. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS), 30., 2017, [s. l.]. **Proceedings** [...]. [S. l.]: Curran Associates, 2017. p. 5998-6008.

VIOLA, P.; JONES, M. J. Robust real-time face detection. **International Journal of Computer Vision**, [s. l.], v. 57, n. 2, p. 137-154, 2004.

WANG, Liwei et al. Design and research of intelligent strawberry sorting robot based on machine vision technology. **Smart Agricultural Technology**, [s. l.], v. 13, art. 101695, 2026. DOI: 10.1016/j.atech.2025.101695.

WEI, Xin; YU, Ruixuan; SUN, Jian. View-GCN: view-based graph convolutional network for 3D shape analysis. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN

RECOGNITION (CVPR), 2020, [s. l.]. **Proceedings** [...]. IEEE, 2020. p. 1847-1856.

YANG, Jiwoong; CHUNG, Haejun; JANG, Ikbeom. Hierarchical mutual distillation for multi-view fusion: learning from all possible view combinations. **Pattern Recognition**, [s. l.], v. 178, art. 113432, 2026. DOI: 10.1016/j.patcog.2026.113432.

YANG, Yueheng et al. Lightweight attention-guided YOLO with level set layer for landslide detection from optical satellite images. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, [s. l.], v. 17, p. 3543-3559, 2024.

YU, Zhiwen et al. A review on multi-view learning. **Frontiers of Computer Science**, [s. l.], v. 19, n. 7, 2025.

ZAIDI, Syed Sahil Abbas et al. A survey of modern deep learning based object detection models. **Digital Signal Processing**, [s. l.], v. 126, p. 1-18, 2022.

ZHENG, Caiwang et al. Object-detection from multi-view remote sensing images: a case study of fruit and flower detection and counting on a central Florida strawberry farm. **International Journal of Applied Earth Observation and Geoinformation**, [s. l.], v. 123, art. 103457, 2023.

ZHU, Xizhou et al. Deformable DETR: deformable Transformers for end-to-end object detection. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 9., 2021, [s. l.]. **Proceedings** [...]. ICLR, 2021. p. 1-16.

ZOU, Zhengxia et al. Object detection in 20 years: a survey. **Proceedings of the IEEE**, [s. l.], v. 111, n. 3, p. 257-276, 2023.

Esta obra está licenciada com Licença Creative Commons Atribuição-Não Comercial 4.0 Internacional.