Recibido 23 May. 2023 *ReCIBE, Año 12 No. 1, May. 2023*

Aceptado 05 Jul. 2023

Revisión de Problemas en la Detección de Objetos en Imágenes y Videos Digitales

A Survey on Problems of Detection of Objects in Digital Images and Videos

Miguel Ángel Gutiérrez Velázquez[1]
m19061419@chihuahua.tecnm.mx

Mario Ignacio Chacon-Murguia[1]
mario.cm@chihuahua.tecnm.mx

Alma Delia Corral-Saenz[1]
alma.cs@chihuahua.tecnm.mx

[1] Tecnológico Nacional de México / I.T. Chihuahua, Chihuahua, Chihuahua, México.

Resumen

En las últimas décadas, la detección de objetos ha sido una tarea muy importante en el área de visión por computadora, ya que la detección de objetos localiza y clasifica uno o más objetos en una imagen o videos. En este artículo, se presenta una revisión de artículos y se describen técnicas clásicas y de aprendizaje profundas utilizadas para la detección de objetos. Además, se realiza una revisión de trabajos recientes sobre la detección de objetos, enfocándose en cómo se solucionan algunos de sus problemas más relevantes. Los problemas que se abarcan son: oclusión, confusión, información contextual, cambios en la iluminación, objetos pequeños y cambios de escala, variación entre la misma clase y diferentes clases, y deformación y cambios de pose. Se espera que este artículo sirva para que los interesados en el área conozcan ideas y enfoques para resolver problemas existentes en la detección de objetos incluyendo los últimos avances.

Abstract

In recent decades, object detection has been a very important task in the area of computer vision, since object detection locates and classifies one or more objects in an image or video. In this article, a review of articles is presented, and classical and deep learning techniques used for object detection are described. In addition, a review of recent works on object detection is carried out, focusing on how some of its most relevant problems are solved. Issues covered are occlusion, confusion, contextual information, lighting changes, small objects and scale changes, variation between the same class and different classes, and deformation and pose changes. It is hoped that this article will help those interested in the area to learn about ideas and approaches to solving existing problems in object detection, including the latest advances.

Palabras clave: Detección de objetos, Problemas en la detección de objetos, Aprendizaje profundo.

Keywords: Object detection, Problems in object detection, Deep Learning.

1. Introducción

La detección de objetos es parte fundamental del proceso de la visión por computador, ya que brinda información de interés presente en una imagen. La detección de objetos es un área en constante evolución, lo que hace fundamental el mantenerse actualizado en los nuevos métodos e ideas relacionados a ella.

El campo de aplicación de la detección de objetos es amplio: detección de animales (Wang L., *et al.*, 2021; Li N., *et al.*, 2020; Yudin D., *et al.*, 2019; Kellenberger B., *et al.*, 2019; Singh A., *et al.*, 2020), detección de peatones (Han. B, *et al.*, 2020; Cygert S. y Czyzewski A., 2020), de billetes (Rodríguez A., *et al.*, 2020), diagnóstico médico (Gurbina M., *et al.*, 2019; Pathare S., *et al.*, 2020; Meda K., *et al.*, 2021), detección de desastres (Muhammad K., *et al.*, 2018; Radhika S., *et al.*, 2018), astronomía (Wu T., 2020), agricultura (Song C., *et al.*, 2020; Kaur M. y Min C., 2018), detección de objetos bajo el agua (Chen Z., *et al.*, 2020), detección de medios de transporte (Mo Y., *et al.*, 2019; Yilmaz B. y Karsligil M., 2020; Huang G., *et al.*, 2019), detección del horizonte (Zardoua Y., *et al.*, 2021), deportes (Guo. T., *et al.*, 2020; Bastanfard A., *et. al.*, 2019), etcétera. Sin embargo, a pesar de ser un área ampliamente estudiada durante varias décadas, la detección de objetos es un concepto que carece de un acuerdo universal (Liu L., *et al.*, 2019) ya que hay autores que limitan el alcance a la localización de objetos (Pandiya M., *et al.*, 2020), otros primero realizan una detección y después la localización de los objetos (Pathak A., *et al.*, 2018; Adreopoulos A. y Tsotsos J., 2013) y otros más, consideran la clasificación de los mismos como parte del proceso de detección (Vashisht M. y Kumar B., 2020; Zhang H. y Hong X., 2019). Asimismo, se incorporan variantes a estas etapas como realizar la localización mediante rectángulos delimitadores (Biswas S., *et al.*, 2021; Xiao Y., *et al.*, 2020) o bien, determinar primero si existen objetos de interés y de ser así, regresar la localización y la etiqueta de la categoría (Liu L., *et al.*, 2019).

Además de los enfoques de detección de objetos, existen trabajos publicados que describen posibles soluciones a problemas o aplicaciones específicas en la detección de objetos, como los mostrados en la Tabla . Los trabajos también presentan arquitecturas, métricas y conjuntos utilizados. La Figura 1 muestra de manera sistemática el contenido y organización de este artículo.

2. Análisis de estudios existentes

Debido a la gran cantidad de trabajos que hay en la literatura sobre detección de objetos, es de utilidad hacer una recopilación de los artículos más significativos en los avances del tema, sus métodos y las aplicaciones específicas.

A partir del análisis de los artículos se observaron las siguientes tendencias: análisis de aplicaciones específicas de detección de objetos, detección de objetos 3D, detección de objetos sobresalientes en una imagen, detección de objetos en imágenes RGB-D, detección de objetos pequeños, detección de objetos camuflados, detección de objetos en movimiento, detección de objetos con Few-Shot, detección de objetos con YOLO y análisis general del área de detección de objetos (Shetty A., et al., 2021; Deng J., et. al., 2020; Kaushal M., et al., 2018; Aziz L., et al., 2020; Jiao L., et al., 2019; Arulprakash E. y Aruldoss M., 2021; Abbas S., et al., 2022; Zou Z., et al., 2023); siendo estos artículos de análisis generales los recomendados para aquellos lectores que deseen introducirse al área de detección de objetos.

En la Tabla se presenta una descripción más detallada de los artículos mencionados en esta sección. Con base en el análisis previo, este artículo analiza y describe las ideas y enfoques implementados para resolver problemas presentados frecuentemente en la detección de objetos. Además, hace un hincapié en posibles aportes para mejorar y hacer más robusta la detección de objetos ante los problemas mencionados en este artículo, así como recomendaciones para una futura investigación.

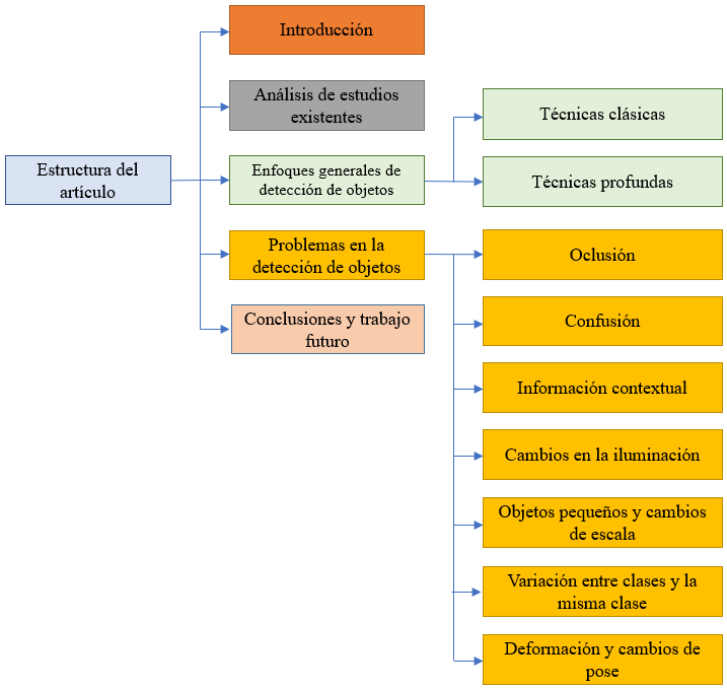

Figura 1. Estructura del artículo.

Artículo	Tema	Resumen
(Mohammed S., et al., 2021)	Detección de barcos	Analizan dos algoritmos: Multi-Input Convolutional Long Short Term Memory y Rotational Libra Region Based CNN para la detección de barcos.
(Moniruzzaman M., et al., 2017)	Detección de objetos debajo del agua	Mencionan detectores profundos en la detección de objetos marinos, así como enfoques a objetos específicos, como pescados, plancton o corales.
(Xu S., et al., 2023)	Detección de objetos debajo del agua	Presenta un análisis de los problemas en la detección de objetos bajo el agua, así como futuras tendencias y aplicaciones. Además, se analiza la relación entre el mejoramiento de imágenes y la detección de objetos.
(Ajmera F., et. al., 2021)	Detección de objetos en imágenes aéreas	Resumen de varios métodos tanto de aprendizaje profundo como con enfoques clásicos para la detección de objetos en imágenes aéreas. También se mencionan las ventajas, desventajas, precisión y conjuntos de datos.
(MIttal P., et al., 2020)	Detección de objetos en imágenes áreas	Se hace una revisión de algoritmos de aprendizaje profundo recientes en la detección de objetos en imágenes tomadas con UAV (Unmanned Aerial Vehicle) a baja altitud.
(Said N., et al., 2019)	Detección de desastres	Revisión de detección de desastres en redes sociales y en imágenes satelitales. Además, se discuten trabajos futuros y desafíos existentes.
(Geetha S., et al., 2021)	Detección de fuego	Mencionan métodos fundamentales de procesamiento de imágenes y CNN para la detección de humo y fuego, también conjuntos de datos existentes y desafíos y sugerencias de mejoras.
(Liang D., et al., 2020)	Detección de líneas	Organizan métodos dependiendo en los objetos a detectar y se presenta un nuevo conjunto de datos.
(Kaur P., et al., 2020)	Detección de tumores	Revisan varias técnicas de detección de tumores en imágenes MR junto con sus dificultados y fortalezas. Se cubren las estrategias más relevantes, restricciones, métodos, sus reglas de aplicación, y las direcciones futuras.

Tabla 1. Comparación de artículos de estudios sobre la detección de objetos.

Artículo	Tema	Resumen
(Naji S., et al., 2019)	Detección de piel	Se revisan métodos para la detección de la piel en imágenes a color. Además, se mencionan espacios de color, costos, riesgos, bases de datos, pruebas y una evaluación comparativa.
(Feng Y., et al., 2021)	Detección de rostros	Se presentan métodos, análisis en términos de precisión y eficiencia de detectores profundos de rostros. Se comparan conjuntos de datos y métricas de evaluación.
(Minaee S., et al., 2021)	Detección de rostros	Se hace revisión de métodos profundos para la detección de rostros y se proponen categorías de agrupamiento, se presentan arquitecturas y sus desempeños en evaluaciones comparativas populares, se presentan conjuntos de datos y se mencionan desafíos actuales y futuras direcciones de investigación.
(Ning C., et al., 2021)	Detección de peatones	Análisis del progreso en la detección de peatones con oclusión.
(Espinosa J., et al., 2020)	Detección de motocicletas	El artículo se enfoca en los algoritmos para la detección y seguimiento de motocicletas, se describe el desempeño en conjuntos de datos, además se presentan desafíos actuales, así como trabajo futuro.
(Jarunakarint V., et al., 2020)	Detección de motocicletas	Realizan experimentos en 20 modelos distintos de aprendizaje de máquina para detectar motocicletas.
(Nyein M. y Tint T., 2021)	Detección de vehículos	Se hace un análisis del desempeño de diversos métodos para la detección de vehículos mediante resultados experimentales utilizando conjuntos de datos habituales.
(Li Z., et al., 2021)	Detección de objetos 3D	Revisión del desarrollado en la detección de objetos 3D aplicados a la conducción inteligente, así como sus defectos y futuras investigaciones.
(Lang W., et al., 2021)	Detección de objetos 3D	Se enfocan en métodos populares de aprendizaje profundo para la detección de objetos 3D. Se dividen los enfoques en 4 categorías de acuerdo con los datos de entrada. Se hace un análisis de las innovaciones y se hace una comparación de sus desempeños. Así como los desafíos actuales y futuras direcciones.
(Arnold E., et al., 2019)	Detección de objetos 3D	Se revisan métodos utilizados en la conducción autónoma para la detección de objetos 3D. Se categorizan trabajos recientes basados en el tipo de sensor. Se indican problemas actuales y trabajos futuros.
(Kumar N., 2018)	Detección de objetos sobresalientes	Análisis de varios métodos de umbralización utilizados en el estado del arte en SOD. Además, se mencionan umbrales no explorados, se realizan, también, experimentos para mostrar el desempeño de umbralizaciones populares.
(Wang Q., et al., 2020)	Detección de objetos sobresalientes en videos	Hacen una clasificación de métodos del estado del arte, se mencionan los conjuntos de datos y las métricas de evaluación más utilizadas, se realizan experimentos para comparar el desempeño de los métodos y, finalmente, se mencionan trabajos futuros.
(Ullah I., et al., 2020)	Detección de objetos sobresalientes	Se hace una revisión de técnicas heurísticas y de aprendizaje profundo, además de campos relacionados a SOD (Eye-fixation-prediction, detección en imágenes RGB-D, detección de objetos coprominentes, en video). Además, se mencionan conjuntos de datos y problemas actuales y trabajos futuros.
(Wang W., et al., 2021)	Detección de objetos sobresalientes	Revisan algoritmos de SOD: su arquitectura, nivel de supervisión, paradigma de aprendizaje, detección nivel objeto/instancia. También métricas y conjuntos de datos, comparación de modelos SOD representativos. Se expone un conjunto de datos desafiante. Se analiza el desempeño de modelos SOD ante perturbaciones en la entrada y ataques adversarios. Se mencionan problemas actuales y direcciones para investigaciones futuras.
(Borji A., et al., 2019)	Detección de objetos sobresalientes	Muestran el progreso en SOD, sus inicios, conceptos fundamentales, las técnicas principales y conjuntos de datos y métricas de evaluación.

Tabla 1. Comparación de artículos de estudios sobre la detección de objetos [continuación].

Artículo	Tema	Resumen
(Borji A., et al., 2019)	Detección de objetos sobresalientes	Muestran el progreso en SOD, sus inicios, conceptos fundamentales, las técnicas principales y conjuntos de datos y métricas de evaluación.
(Han J., et al., 2018)	Detección de objetos sobresalientes	~~Se~~ Presentan definiciones, tareas, técnicas recientes, investigaciones esenciales, conjuntos de datos y métricas de evaluación y comparación y análisis de resultados experimentales. Y ~~se~~ hacen hincapié en la relación entre detección de objetos genéricos, sobresalientes y específicos a una clase.
(Zhou T., et al., 2021)	Detección de objetos sobresalientes	Se hace una revisión de modelos SOD enfocados a imágenes RGB-D, así como conjuntos de datos para ese dominio. Se hace una evaluación de diversos métodos. Finalmente, se mencionan desafíos existentes para investigaciones futuras.
(Liu Y., Sun P., et al., 2021)	Detección de objetos pequeños	Resumen de métodos de aprendizaje profundo para la detección de objetos pequeños. Se resumen desafíos y soluciones. Se incorporan técnicas de cuatro áreas de investigación: detección de objetos genéricos, detección de rostros, detección de objetos en imágenes aéreas y segmentación. Se hace comparación de diversos métodos de aprendizaje profundo utilizando tres bases de datos.
(Tong K., et al., 2020)	Detección de objetos pequeños	Revisión de métodos basados en aprendizaje profundo desde cinco aspectos: aprendizaje de características multiescala, aumento de datos, estrategias de entrenamiento, detección basada en el contexto y detección basada en redes tipo GAN. Se hace un análisis de algunos algoritmos típicos en los conjuntos de datos MS-COCO y PASCAL-VOC. Se concluye con posibles investigaciones futuras.
(Cheng G., et al., 2022)	Detección de objetos pequeños	Se efectúa un análisis de la literatura sobre la detección de objetos pequeños. Además, se presentan dos conjuntos de datos para la detección de objetos pequeños en escenarios aéreos y de conducción.
(Mondal A., 2021)	Detección de objetos camuflajeados	Análisis de técnicas utilizando procesamiento de imágenes, aprendizaje de máquina, gráficos de computadora y aprendizaje profundo. Se presentan conjuntos de datos y se mencionan futuras investigaciones posibles.
(Chapel M. y Bouwmans T., 2018)	Detección de objetos en movimiento	Revisión de métodos basados en la substracción del fondo, clasificación de trayectoria, matriz de descomposición y seguimiento de objetos de objetos en movimiento en secuencias de videos adquiridas por una cámara en movimiento.
(Chapel M. y Bouwmans T., 2020)	Detección de objetos en movimiento	~~Se~~ Analizan artículos sobre la detección de objetos en movimiento en videos tomados por una cámara en movimiento. Se proponen categorías para identificar los métodos existentes en el estado del arte de acuerdo con la representación de la escena. También se abordan métodos y enfoques para cámaras sin movimiento, conjuntos de datos y métricas de evaluación.
(Huang G., et al., 2023)	Detección de objetos con Few-Shot	Análisis y caracterización de detección de objetos con Few-Shot y aprendizaje autosupervisado.
(Köhler M., et al., 2023)	Detección de objetos con Few-Shot	Revisión del estado del arte sobre la detección de objetos con Few-Shot; así como conjuntos de datos utilizados, evaluaciones y comparativa de resultados. Además, se mencionan los problemas frecuentes y las posibles tendencias futuras.
(Diwan T., et al., 2023)	Detección de objetos con YOLO	Se presenta un análisis del detector YOLO; su evolución en las arquitecturas, comparaciones, aplicaciones y posibles investigaciones futuras.
(Terven J. y Cordova D., 2023)	Detección de objetos con YOLO	Análisis de las arquitecturas YOLO. Desde la primera versión de YOLO hasta YOLOv8.

Tabla 1. Comparación de artículos de estudios sobre la detección de objetos [continuación].

Artículo	Tema	Resumen
(Shetty A., et al., 2021)	Modelos de detección de objetos	Se realiza un análisis comparativo entre técnicas basadas en aprendizaje profundo (de una y dos etapas) para la detección de objetos.
(Deng J., et al., 2020)	Modelos de detección de objetos	Revisan modelos profundos (de una y dos etapas), se reportan conjunto de datos y desafíos existentes.
(Kaushal M., et al., 2018)	Modelos de detección de objetos	Se hace una recopilación de trabajos con aprendizaje profundo, lógica difusa, algoritmos evolutivos e híbridos para la detección y seguimiento de objetos. Se remarcan los conjuntos de datos. Se presentan desafíos existentes y análisis que pueden guiar investigaciones futuras.
(Aziz L., et al., 2020)	Modelos de detección de objetos	Se revisan modelos de una y dos etapas. Se presentan cinco conjuntos de datos con sus métricas de evaluación. El artículo se centra en cinco aplicaciones: seguridad, milicia, transporte, medicina y vida diaria.
(Jiao L., et al., 2019)	Modelos de detección de objetos	Analizan detectores de una y dos etapas. Se presentan conjuntos de datos, se realiza, también, un análisis de métodos generales para mejorar la detección de objetos. Se mencionan diversas aplicaciones y, finalmente, se presentan tendencias y desafíos interesantes.
(Arulprakash E. y Aruldoss M., 2021)	Modelos de detección de objetos	Examinan arquitecturas profundas, mencionando sus ventajas y desventajas. Se presentan conjuntos de datos y sus métricas de evaluación. También se mencionan ampliamente problemas en la detección de objetos.
(Abbas S., et al., 2022)	Modelos de detección de objetos	Se presentan una gran cantidad de modelos de aprendizaje profundo recientes, también se muestran conjuntos de datos y métricas de evaluación y se realiza una comparación de diversos modelos.
(Zou Z., et al., 2023)	Modelos de detección de objetos	Muestran la evolución de los algoritmos de detección de objetos durante la etapa de 1990-2019. Se presentan detectores fundamentales, conjuntos de datos, métricas, técnicas de aceleración, aplicación y desafíos y mejoras en los últimos años.

Tabla 1. Comparación de artículos de estudios sobre la detección de objetos [continuación].

3. Enfoques generales de detección de objetos

En esta sección se mencionarán brevemente algunos enfoques en la detección de objetos mediante técnicas clásicas y mediante técnicas de aprendizaje profundo. Con el paso del tiempo, los enfoques para la detección de objetos han ido cambiando debido a los avances que ha habido en los algoritmos del aprendizaje de máquina, sin embargo, los enfoques clásicos y mejoras a estos, siguen siendo utilizados. De igual manera, dado el incremento en las capacidades computacionales, en la cantidad de datos de entrenamiento, el desarrollo de algoritmos y de detectores profundos, se ha propiciado que las técnicas profundas sean ampliamente utilizadas en la actualidad.

3.1 Técnicas clásicas

Las técnicas clásicas son aquellas que no utilizan aprendizaje profundo. En esta sección se exponen ideas y métodos de este enfoque cuyas técnicas han predominado en la resolución de tareas de visión por computadora y, por ende, en la detección de objetos. Algunos ejemplos de estas técnicas se describen a continuación. *Color de pixel*: H. Shih y J. Chen (Shih H. y Chen J., 2021) consideran el color como característica para la detección de objetos. Utilizan el espacio de color HSI e YCbCr para detectar piel en imágenes de personas. De igual manera, A. Sáez et. al. (Saez A., et al., 2019) utilizan el color de un pixel y de sus vecinos para obtener la probabilidad de que un pixel pertenezca a cierta clase y así detectar lesiones en la piel mediante imágenes de dermatoscopia. *Template matching*: Es una técnica para detectar en una imagen, la región o regiones que más se asemejen a un patrón dado (Shou X., et al., 2019). *Características invariantes*: Son métodos más avanzados en la detección de objetos (Dash P. y Sigappi A., 2018; Ansari S., 2019; Gupta S., et al.), que extraen características invariantes ante el escalamiento, rotación, traslación como SIFT (Lowe D., 1999), SURF (Bay H., et al., 2006) y ORB (Rublee E., et al., 2011). *Patrones binarios locales (LBP)*: Son métodos basados en textura que han sido utilizados para la detección de rostros (Zhang B., et al., 2010; Zhang G., *et. al.,* 2004; Suruliandi A., et al., 2012; Sun N., et al., 2006). Ejemplo de este método y sus variantes es (Ojala T., et al., 1994). *Histograma de gradientes orientados (HOG)*: Es un método que se enfoca en la forma del objeto, y además brinda la información de la dirección de los gradientes. Algunas investigaciones que utilizan HOG para detectar objetos son (Seemanthini K. y Manjunath S., 2018; Nguyen N., *et al.,* 2019; Li J., et al., 2019). *Los filtros de Haar*: Fueron propuestos por P. Viola y M. Jones en 2004 (Viola P. y Jones M., 2004). Han sido utilizados principalmente en la detección de rostros (Arfi A., *et al.,* 2020; Rahmatulloh *A., et al.,* 2021). En (Arunmozhi A. y Park J., 2018; Adouani A., et al., 2019) se publican comparaciones de HOG, LBP y características de Haar en la detección de vehículos y en la detección de rostros, respectivamente. *Filtros de correlación*: Son una clase de clasificadores, los cuales producen picos en la correlación cuando se localiza un objeto de interés. Estos filtros han sido utilizados en la detección y el seguimiento de objetos, como ASRCF (Adaptive spatial regularization correlation filter) (Dai K., et al., 2019). Para más información véase (Liu S., et. al., 2021; Du S. y Wang S., 2022). En la Tabla 2 se resumen las citas de la información presentada en esta sección.

Artículo	Enfoque	Artículo	Enfoque
(Shih H. y Chen J., 2021;Saez A., et. al., 2019)	Color del pixel	(Seemanthini K. y Manjunath S., 2018; Nguyen N., et al., 2019; Li J., et. al., 2019)	Histograma de gradientes orientados
(Shou X., et al., 2019)	Template matching	(Arfi A., et al., 2020; Rahmatulloh A., et al., 2021)	Filtros de Haar
(Dash P. y Sigappi A., 2018; Ansari S., 2019; Gupta S., et al.)	Características invariantes	(Liu S., et al., 2021; Du S. y Wang S., 2022).	Filtros de correlación
(Zhang B., et. al., 2010; Zhang G., et. al., 2004; Suruliandi A., et al., 2012; Sun N., et. al., 2006)	Patrones binarios locales		

Tabla 2. Algunos enfoques clásicos en la detección de objetos.

3.2 Técnicas profundas

Como se mencionó anteriormente, la detección de objetos tiene dos etapas: la localización y la detección. Dado que las redes neuronales convoluciones profundas (DCNN del inglés Deep Convolutional Neural Network) presentan un alto desempeño en tareas de clasificación, los métodos de detección de objetos más empleados en la actualidad están basados en DCNN. Este tipo de arquitecturas de detección de objetos puede dividirse en dos categorías: arquitecturas de detección de objetos de una etapa y de dos etapas (Zou Z., et. al. 2023).

3.2.1 Detectores de una etapa

Se caracterizan por localizar y clasificar el objeto en una sola etapa. Tienen la ventaja de presentar mayor velocidad de detección con respecto a los detectores de dos etapas, sin embargo, la precisión es menor que estos. A continuación, se hace una revisión de diversos detectores de una etapa.

En 2013 P. Sermanet *et al.* (Sermanet P., et al., 2013) mostraron OverFeat que consiste en utilizar la técnica multiscale and sliding window en la última capa de pooling de la DCNN para extraer las regiones de interés (donde puede haber un objeto) y predecir los valores de clasificación para cada región. En 2015 W. Liu et. al (Lu W., et al., 2016) propusieron la red SSD (Single Shot MultiBox Detector), cuya mayor contribución consistió en las técnicas de multirreferencia y multiresolución. La técnica de multirreferencia consiste en tener referencias (regiones rectangulares en distintos lugares en la imagen) y realizar las predicciones en dichas referencias; por otro lado, la técnica de multiresolución radica en hacer detecciones a diferentes escalas en diferentes capas de la red. J. Redmon et al. (Redmon J., et al., 2016) propusieron YOLO. Este detector divide la imagen en regiones y predice los rectángulos delimitadores y las probabilidades de que el objeto en dicha región pertenezca a cierta clase. J. Redmon y A. Farhadi mejoraron YOLO en (Redmon J. y Farhadr A., 2017) y (Redmon J. y Farhadr A., 2018). En (Redmon J. y Farhadr A., 2017) introdujeron la normalización por lotes para acelerar la convergencia y mejorar la generalización de la red. En (Redmon J. y Farhadr A., 2018) utilizaron la red FPN para que detectar más clases y objetos de múltiples tamaños. Más versiones de la red YOLO han sido presentadas (Terven J. y Cordova D., 2023), desde YOLOv4 hasta YOLOv8 (Bochkovskiy A., et al., 2020; Jocher G., 2020; Li C. et al., 2022; Wang C., et al.,2022; Jocher G., et al., 2023). En (Lin T., Goyal P., et. al., 2017) se tiene la red RetinaNet, donde se introduce una función de pérdida llamada «focal loss» para abordar el desequilibrio entre las clases del fondo y del primer plano. Esta función remodela la función de pérdida de entropía cruzada de modo que disminuye la pérdida asignada a muestras bien clasificadas. Por último, M. Hajizadeh et al. (Hajizadeh M., et al., 2023) presentan la red MobileDenseNet para mejorar la detección de objetos pequeños con un sistema embebido.

3.2.2 Detectores de dos etapas

Los detectores de dos etapas realizan la localización del objeto en una etapa y su clasificación en otra. En estos métodos primero se generan las regiones de propuesta (regiones candidatas a contener el objeto de interés) y después se extraen características de las regiones para determinar si el objeto se localiza en alguna de esas regiones. Su mayor ventaja con respecto a los detectores de una etapa es su mejor exactitud, sin embargo, suelen ser más lentos que ellos. Enseguida se describen algunos de los principales métodos de detección de objetos de 2 etapas presentados en la literatura.

R. Girshick et al. (Girshick R., et al., 2014) introdujeron los detectores de objetos en la era del aprendizaje profundo con la red Regions with CNN Features (RCNN), la cual combina algoritmos de generación de regiones de interés, seguido por la extracción de características con una CNN. Esta red tiene la importancia de haber introducido a los detectores de objetos en la era de las redes neuronales, pero tiene algunas desventajas como la generación de 2000 regiones de propuesta, lo cual consume espacio y recursos. K. He et al. (He K., et al.,2015) presentaron la red Spatial Pyramid Pooling Networks (SPPNet), cuya mayor contribución fue introducir la capa Spatial Pyramid Pooling (SPP), la cual le permite a la CNN generar un vector de características de tamaño fijo sin importar el tamaño de la imagen o región de interés. En el 2015, R. Girshick (Girshick R., 2015) mostró la red Fast RCNN. Fast RCNN entrena simultáneamente el detector y la regresión del rectángulo delimitador. En el 2017 S. Ren et al. (Ren S., et al., 2017) propusieron la red Faster RCNN, que incorpora una red de generación de propuestas (RPN). En la red Faster RCNN los bloques de detección de propuestas, extracción de características, regresión del rectángulo delimitador, etc, son integrados en un marco unificado. Después se desarrollaron otros métodos como R-FCN (Dai J., et al.,2016) y Light head RCNN (Li Z., et al., 2017). En 2017 T. Lin et al. (Lin T., Dollar P., et al., 2017) presentaron la Feature Pyramid Networks (FPN), donde su mayor contribución consiste en generar características de alto nivel semántico en todas las escalas, haciendo de esta una red apta para la detección de objetos pequeños. K. He et al. propusieron la red Mask R-CNN en el 2017 (He K., Gkioxari et al., 2017), esta red es una mejora de la red Faster R-CNN que permite también realizar una segmentación. En el 2019 G. Gkioxari et al. (Gkixioari G., et al., 2019) presentaron la red Mesh RCNN, la cual está basada en Mask RCNN. La red Mesh RCNN además de realizar la detección de objetos, también da una representación tridimensional de dichos objetos.

En la Tabla 3 y en la Figura 2 se muestra el resumen de citas de los detectores profundos presentados en esta sección.

Detectores de una etapa		**Detectores de dos etapas**	
Artículo	**Red**	**Artículo**	**Red**
(Sermanet P., et al., 2013)	OverFeat	(Girshick R., et al., 2014)	RCNN
(Lu W., et al., 2016)	SSD	(He K., et al.,2015)	SPPNet
(Redmon J., et al., 2016)	YOLO	(Girshick R., 2015)	Fast RCNN
(Redmon J. y Farhadr A., 2017)	YOLOv2	(Ren S., et al., 2017)	Faster RCNN
(Redmon J. y Farhadr A., 2018).	YOLOv3	(Dai J., et al.,2016)	R-FCN
(Bochkovskiy A., et al., 2020)	YOLOv4	(Li Z., et al., 2017)	Light head RCNN
(Jocher G., 2020)	YOLOv5	(Lin T., Dollar P., et al., 2017)	FPN
(Li C. et al., 2022)	YOLOv6	(He K., et al., 2017)**¡Error! No se encuentra el origen de la referencia.**	Mask RCNN
(Wang C., et al.,2022)	YOLOv7	(Gkixioari G., et al., 2019)	Mesh RCNN
(Jocher G., et al., 2023)	YOLOv8		
(Lin T., Goyal P., et al., 2017)	RetinaNet		
(Hajizadeh M., et al., 2023)	MobileDenseNet		

Tabla 3. Enfoques con aprendizaje profundo para la detección de objetos.

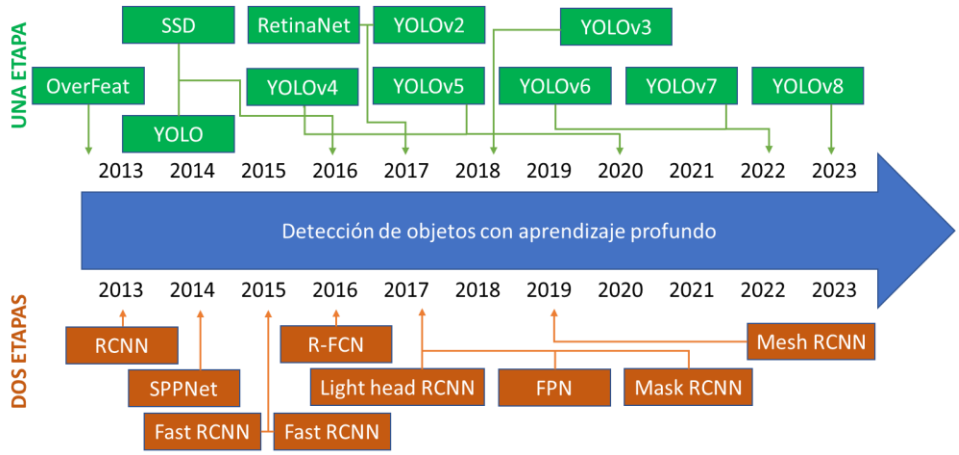

Figura 2. Línea de tiempo de arquitecturas de aprendizaje profundo para la detección de objetos.

4. Problemas en la detección de objetos

La detección de objetos es una tarea de visión por computadora que, a pesar de los avances en los últimos años, sigue presentando grandes dificultades. En esta sección se describen los problemas más comunes en la detección de objetos, y cómo han sido abordados recientemente en diversas investigaciones. Los problemas abordados en esta revisión son: oclusión, confusión, información contextual, cambios en la iluminación, objetos pequeños y cambios de escala, variación entre la misma clase y entre diferentes clases y deformación y cambios de pose.

4.1 Oclusión

Es un problema en el cual un objeto real o artificial oculta parcial o totalmente información de interés de una imagen.

La mayoría de los artículos analizados tienen como aplicación la detección de rostros y peatones ocluidos, tal situación se puede deber, principalmente, a un par de razones: 1) a la accesibilidad de los conjuntos de datos de entrenamiento cuyas clases son precisamente rostros o peatones, 2) y por el creciente interés en los vehículos autónomos donde es fundamental la correcta detección de personas, aunque estén ocluidas. El interés en los vehículos autónomos también se observa en (Liu Y., et al., 2021; Heo J., et al., 2022; Chilukuri D., et al., 2022), cuya aplicación consiste en detectar señales de tráfico y medios de transporte ocluidos. Por otra parte, existen metodologías en las cuales se analiza el mapa de características (Cen F., *et al.,* 2022; Wang Y., et al. 2022) para comparar con dichos mapas regiones ocluidas o no ocluidas; o para que la red profunda se enfoque en regiones del mapa de características donde no hay oclusión. Otras investigaciones tienen como objetivo la detección de ciertas figuras geométricas (Zhao M., et al., 2021; Dong W., et al., 2021) y a partir de tal objetivo desarrollan una metodología específica. También existen métodos en los cuales se considera que los objetos que ocluyen son ruido, y se elimina ese ruido (Xiang L., et al., 2023). A continuación, se realiza un análisis de cada artículo mencionado en esta sección.

En (Moncef S. y Othman A., 2021) se abordan tres tipos de oclusiones para la detección de rostros: barba, bigote y lentes. Mediante el método de Fuzzy C-Means se extraen la barba y el bigote y con operaciones morfológicas se detecta la presencia de lentes. También W. Zheng et. al. (Zheng W., et al., 2020) utilizan información contextual para solucionar el problema de la oclusión. Este enfoque se inspira en las características de los nervios ópticos de los seres humanos para reconocer rostros con oclusión. Además, se hace un aumento de datos para obtener imágenes con oclusión y se propone aplicar few-shot para el reconocimiento. En (Zhou C. y Yuan J., 2020) se detectan peatones con oclusión mediante la correlación entre algunas partes del cuerpo humano. Dado que no se sabe cuáles partes estarán ocluidas, se aplica un conjunto de detectores de partes. Para una región candidata en la imagen, los K detectores darán K probabilidades, con lo cual se integran esas probabilidades para dar el resultado final indicando qué tan probable es que esa región contenga un peatón.

Para esto se utilizan tres métodos: Max integration, TopS integration y L1 integration. X. Zhou y L. Zhang (Zhou X. y Zhang L., 2022) crearon la red SA-FPN para detectar personas en imágenes de lugares concurridos, donde suele presentarse oclusión. Ellos diseñaron una red, Scale-FPN, para solucionar el problema de variaciones de escala y un mecanismo de atención (attention-based lateral connection) el cual mejora la información semántica y le permite al detector enfocarse en características relevantes de peatones con oclusión. X. Song et. al. (Song X., et al., 2020) presentan la red Progressive Refinement Network (PRNet) para la detección de peatones ocluidos. La red PRNet imita el proceso de anotación progresiva del ser humano en la detección de peatones oclusión; este proceso tiene tres etapas: 1) anotación de la parte visible (se analiza la región visible del peatón), 2) realización de un proceso estadístico (se estima si la región visible corresponde a un peatón) y, 3) una anotación del cuerpo completo (se estima dónde estaría la parte no visible). La PRNet realiza el proceso descrito con anterioridad al entrenarse de la siguiente manera: primero se entrena el módulo Visible-part Estimation (VE) para la detección de la parte superior del cuerpo humano, que corresponde a la anotación de la parte visible, luego el módulo Anchor Calibration (AC) que imita el proceso estadístico y por último el módulo Full-body Refinement (FR), el cual es como la anotación del cuerpo completo. VE y FR toman ground truth de partes visibles y de cuerpo completo como referencia, respectivamente. VE se entrena con el objetivo de aprender los rectángulos delimitadores A1 en las partes visibles del cuerpo humano; luego AC se encarga de migrar los rectángulos delimitadores A1 a rectángulos delimitadores de cuerpo completo A2. Una vez se tiene A2 se entrena FR. En (Song X., et al., 2022), se tiene la red PRNet++ (basada en (Song X., et al., 2020)) para mejorar la habilidad del modelo ante diferentes estados de oclusión (es decir, que el modelo se desempeñe bien en la detección de peatones con poca oclusión y mucha oclusión).

PRNet++ tiene dos ramas: Easy-branch, que se encarga de la detección con poca oclusión y Hard-branch, que se enfoca en la detección con mucha oclusión. Ambas ramas utilizan PRNet (Song X., et al., 2020). En (Wu C. y Ding J., 2018) se presenta un algoritmo, gradient direction-based hierarchical adaptive sparse and low-rank (GD-HASLR) para el reconocimiento de rostros ocluidos. Un método llamado Pose-Driven Visibility Model (PDVM) se propone en (Zhou S., et al., 2020) para la detección de personas. Ese método contiene tres partes: construcción de información de humanos no ocluidos, representación de características globales representativas y características alineadas en parte. R. Biswas et. al (Biswas R., et al., 2021) presentan el método de Estructura de Vecindario Dominante de Frecuencia One-Shot (OSD-DNS). En este método se obtiene la identidad de una cara ocluida mediante un clasificador entrenado con caras no ocluidas. H. Wang et. al. (Wang H., et al., 2021) proponen un modelo, la red Shrinking and Reweightint (SRNet) que puede aprender características globales más robustas y altamente discriminatorias, y obtener umbrales suaves a través del módulo DRS (Deep Residual Shrinkage Module) para eliminar las características relacionadas con el ruido para la detección de peatones.

Además, se diseña un modelo para la coincidencia de características parciales de puntos clave que indican si el objeto es un humano o no (RMPM): se asignan pesos adaptativos a cada característica parcial, las características parciales sin información útil (oclusión) se debilitan. El modelo SRNet utiliza como red pre-entrenada la ResNet50. En (Liu T., et al., 2020) se presenta un enfoque de detección de peatones con una red acoplada que está formada por subredes. Una de esas subredes se enfoca en el problema de la oclusión al utilizar una región de interés regional deformable, la cual genera características robustas de peatones con oclusión, ya que se adapta mejor a las variaciones de posición de partes de los peatones. Este enfoque utiliza la RFCN (Region-based Fully Convolutional Network) debido a que puede localizar un objeto aun cuando exista una información parcial del mismo.

En (Xie J., et al., 2020) se tiene un método para la detección de peatones con oclusión llamado mask-guided attention network (MGAN), el cual genera una máscara de atención espacial utilizando información de regiones visibles del cuerpo. Este método utiliza la red preentrenada Faster R-CNN. En (Wank K., et al., 2020) se incrementa el conjunto de entrenamiento al anotar imágenes con oclusión y diversidad de poses para la detección de rostros, además se propone la red Region Attention Networks (RAN), que captura de manera adaptativa la importancia de las regiones faciales, se utilizan las redes profundas pre entrenadas ResNet-18 y VGG16. M. Flores et. al. (Flores M., et al., 2019) presentan un trabajo donde se divide la región de interés en 12 secciones, de esta manera se extraen las partes más relevantes; es decir, que tienen el potencial para identificar la forma humana. Para la extracción de características se usa el histograma de gradientes orientados (HOG) y para la clasificación se utilizan los algoritmos de máquina de vectores de soporte (SVM) e inferencia lógica (IL). J. Zuo et. al (Zuo J., et al., 2018) plantean un método con aprendizaje profundo para la detección de personas ocluidas, su método es Attention Framework of Person Body (AFPB), el cual consiste en dos componentes principales: un simulador de oclusión (OS), que genera muestras artificiales de personas ocluidas utilizando las muestras de personas sin ninguna oclusión; y un clasificador binario el cual es entrenado con las muestras sin oclusión y las muestras generadas con el OS. En (Zhou C. y Yuan J., 2020) se propone un enfoque que descubre patrones de oclusión que son representativos y discriminativos. Se basa en un modelo de partes deformables (DPM, del inglés Deformable Part Model) entrenado con muestras sin oclusión.

A partir de los patrones de oclusión, se buscan solamente los patrones que son representativas y que tienen un alto desempeño en la clasificación de manera individual. J. Dong et. al. (Dong J., et al., 2020) presentan un método para eliminar oclusiones basada en la red generativa adversaria (GAN). Este método consiste en dos GAN: G1, que representa un generador de oclusión, dada una imagen de un rostro y una imagen de oclusión, y G2 es la red que remueve la oclusión. Ambos generadores utilizan una arquitectura U-Net como codificador-decodificador. Y. Xu et. al. (Xu Y., et al., 2018) utilizan características profundas con forma de Omega (Ω), que son características de la región de cabeza y hombros de las personas para la detección de personas con oclusión. Se utiliza la red RFCN, la cual se entrena con muestras etiquetadas de todo el cuerpo humano, y otras imágenes etiquetadas de la parte de los hombros y la cabeza, haciendo el modelo más robusto ante posibles oclusiones. En (Wei W., et al., 2019) se utiliza la red profunda MobileNet para detectar peatones con oclusión mediante información de profundiad. C. Liu et. al (Liu C., et al., 2022) presentan la red Double Mask R-CNN para detectar personas en lugares concurridos, en donde existe oclusión. Esto lo logran al detectar puntos clave del cuerpo humano con la red MKFRCNN (Mask and Keypoint with Fast RCNN), de acuerdo con estos puntos clave se determina la visibilidad de las regiones donde están las personas. Si la visibilidad es menor que un cierto umbral, se añade una máscara para poder detectar las personas ocluidas.

Y. Liu et. al. en (Liu Y., Peng J., et al., 2021) proponen una red, TSingNet, para la detección de señales de tráfico ocluidas, la red aprende características sobre el contexto para la detección. Se utiliza la red ResNet50, y para obtener una información relevante del contexto, se reemplazan las capas de convolución de la última capa con tres bloques de fusión adaptativa de campo receptivo (ARFF). J. Heo et. al (Heo J., et al., 2022) crean la arquitectura Occlusion-aware Spatial Attention Transformer (OSAT), la cual está basada en Vision Transformer (ViT), CutMix augmentation y Occlusion Mask Predictor para resolver el problema de la oclusión. La utilización del ViT radica en implementar un mecanismo de atención para analizar medios de transporte con oclusión. En (Chilukuri D., et al., 2022) realizan una segmentación por instancias en imágenes de señales de tráfico con oclusión, posteriormente se realiza la detección. Gracias a la segmentación por instancias el modelo basado en la red SSD MobileNet atiende el problema de la oclusión.

En (Cen F., et al., 2021) se propone un enfoque para la clasificación de imágenes con oclusión al ajustar los parámetros de modelos pre entrenados con un conjunto de vectores de características profundas aumentado (DFVs). Los DFV son extraídos por una CNN. Se observa que el vector de diferencia (DV) entre los DFVs de una imagen ocluida y una sin oclusión está altamente relacionada con la oclusión. A partir de esto se propone un aumento de DFVs con pseudo-DFVs, que son generados de manera aleatoria al añadir DVs extraídos de un conjunto pequeño de pares de imágenes limpias y con oclusión. En (Li F., et al., 2022) se propone la red Generative Adversarial Occlusion Network (GAON), la cual se entrena junto con la red Faster R-CNN. Mientras la Faster R-CNN aprende a reconocer imágenes de varias clases, la GAON aprende a generar oclusiones en el mapa de características de las capas escondidas de la Faster R-CNN; de esta manera el detector es robusto ante la oclusión. En (Wang Y., et al., 2022) se presenta una red llamada OD-UTDNet. Esta red tiene un módulo (Dilaed Attention Cross Stage Partial) el cual es responsable de dos aspectos principales: 1) hace que la red OD-UTDNet se enfoque en las regiones sin oclusión y 2) mejora la capacidad de la red OD-UTDNet para extraer características en regiones con oclusión.

Una técnica basada en triángulos inscritos para la detección de círculos ocluidos se presenta con M. Zhao et. al. (Zhao M., et al., 2021), en (Dong W., et al., 2021) se detectan formas elípticas con oclusiones muy grandes mediante una modificación de la red R-CNN: Ellipse R-CNN, para manejar mejor la oclusión se integró la red U-Net para generar mapas de características decodificados que contienen información escondida. Además, se aprenden varios patrones de oclusiones.

L. Xiang et. al. (Xiang L., et al., 2023) consideran que en una región de interés la información que ocluye es ruido, a partir de esto diseñan un proceso iterativo llamado DINF (Dynamic instance noise filter) para mejorar la relación señal-ruido de las características de los objetos de interés. DINF primeramente transforma las características obtenidas en las regiones de interés, RDI, a un dominio en el cual los valores de la señal de ruido son cercanos a cero. Después, se utiliza una umbralización para remover el ruido. En la Tabla se hace un resumen de los artículos previos.

Artículo	Método	Artículo	Método
(Moncef S. y Othman A., 2021)	Con Fuzzy C-Means se extrae la barba y el bigote, y mediante operaciones morfológicas se detectan los lentes para la detección de rostros.	(Zuo J., et al., 2018)	Generación de muestras con oclusión a partir de muestras sin oclusión.
(Zheng W., et al., 2020)	Con la información contextual y el aumento de datos se detectan objetos con oclusión.	(Zhou C. y Yuan J., 2020)	Se basa en un modelo de partes deformables.
(Zhou C. y Yuan J., 2020)	Al utilizar un detector de partes humanas se detectan peatones con oclusión.	(Dong J., et al., 2020)	Dos redes generativas adversarias (GAN), una de ellas genera oclusión y la otra la elimina.

Tabla 4. Resumen de los artículos que atienden el problema de oclusión.

Artículo	Método	Artículo	Método
(Zhou X. y Zhang L., 2022)	Un mecanismo de atención mejora la información semántica, gracias a esto el detector se enfoca en características de peatones con oclusión.	(Xu Y., et al., 2018)	Con características de la región de la cabeza y hombros de humanos (características omegas) se hace más robusta le red R-FCN ante los objetos con oclusión.
(Song X., et al., 2020)	Inspirado en el proceso de anotación utilizado por el ser humano para la detección de personas con oclusión el cual consiste en: analizar la región visible, se estima si la región es de una persona y si sí se estima la región completa de la persona.	(Wei W., et al., 2019)	Con el uso de información de profundidad se resuelve el problema de la oclusión.
(Song X., *et. al.*, 2022)	La red PRNet++ tiene dos ramas, una rama se hace experta en la detección con poca oclusión y la otra se hace experta en la detección con alta oclusión.	(Liu C., et al., 2022)	Se detectan puntos clave del cuerpo humano para determinar la visibilidad de personas ocluidas.
(Wu C. y Ding J., 2018)	Con base al método *gradient direction-based hierarchical adaptive sparse* se detectan objetos con oclusión.	(Liu Y., Peng J., et al., 2021)	La red TSingNet aprende características contextuales para la detección de señales de tráfico con oclusión.
(Zhou S., et al., 2020)	A partir de características globales y la información en imágenes de personas sin oclusión se detectan personas con oclusión.	(Heo J., et al., 2022)	Implementación de un mecanismo de atención para analizar los objetos con oclusión.
(Biswas R., et al., 2021)	Se utiliza OSD-DNS para reconocer caras ocluidas.	(Chilukuri D., et al., 2022)	Segmentación por instancias.
(Wang H., et al., 2021)	Asignación de pesos adaptativos para debilitar las características con oclusión.	(Li F., et al., 2022)	Se presenta una red adversaria para generar oclusiones en el mapa de características de las capas más profundas de la red Faster R-CNN.
(Liu T., et al., 2020)	Se tienen subredes acopladas. Una de esas subredes utiliza una región de interés deformable, la cual extrae características robustas de peatones con oclusión.	(Wang Y., et al., 2022)	A partir de regiones sin oclusión se mejora la capacidad para extraer características en regiones con oclusión.
(Xie J., et al., 2020)	La red *mask guided attention network* se enfoca en regiones visibles del cuerpo humano.	(Zhao M., et al., 2021)	Mediante triángulos inscritos se detectan círculos ocluidos.

(Wang K., et al., 2020) y (Cen F., et al., 2021)	Se utiliza la diferencia de vectores con características en imágenes con oclusión y sin oclusión.	(Dong W., et al., 2021)	Con la integración de las redes Ellipse R-CNN y U-Net se detectan formas elípticas.
(Flores M., et al., 2019)	La imagen se divide en secciones para extraer las regiones que tienen el potencial de ser partes humanas. Las características se extraen con HOG.	(Xiang L., et al., 2023)	Se considera que la información en una región de interés se divide en dos: objeto de interés y ruido, donde el ruido es la información de los objetos que ocluyen. A partir de esto, se diseña un modelo para eliminar este ruido.

Tabla 4. Resumen de los artículos que atienden el problema de oclusión [continuación].

4.2 Confusión

La confusión, o cluttering, es un problema de detección de objetos que se presenta cuando el fondo de la imagen contiene color o textura semejantes al del objeto de interés. En la literatura existen varios trabajos desarrollados recientemente para lograr la detección de objetos en este tipo de escenarios.

La metodología más utilizada para resolver este problema consiste en diseñar arquitecturas para encontrar y mejorar los bordes de los objetos de interés (Skoviera R., et al., 2018; Xu X., et al., 2021; Ji G., et al., 2022; Chen T., et al., 2022; Zhang C., et al., 2022; Zhuge M., et al., 2022). La razón de esto se debe a la naturaleza misma del problema: el contorno (bordes) de los objetos que se confunden con el fondo (objetos camuflados) es difícil de identificar. Con lo cual, si se logra obtener correctamente los bordes de los objetos camuflados, la detección resultará más sencilla. Existen otras metodologías menos utilizadas basadas en información contextual, enfoque en primer plano, etc. (Liu S., et al., 2021; Huang X., et al., 2021; Zhang T., et al., 2020; Zhou S., et al., 2019; Wang K., et al., 2021; Kreim A., et al.; Chen G., et al., 2022). Sin embargo, se aprecia que la tendencia para la detección de objetos camuflados está en le detección de bordes. Tal como la oclusión, la detección de personas y vehículos también es una aplicación de interés en la detección de objetos camuflados (Zheng Y., et al., 2019; Wang Y., et al., 2020; Wei X., et al., 2020). A continuación, se presentan a más detalle estos artículos.

R. Skoviera et. al. (Skoviera R., et al., 2018) proponen el uso de una red bio inspirada llamada Hierarchical Temporal Memory (HTM) para la detección de múltiples objetos en imágenes con confusión. El objetivo de la HTM es aprender a detectar representaciones invariantes de la entrada. Para mejorar el desempeño de la detección basado en HTM, se propone un sistema que usa información de tres HTM que se enfocan en diferentes características de la imagen: ejes, textura y color. El conjunto de datos fue generado para representar el cluttering. X. Xu et. al. (Xu X., et al., 2021) detectan el contorno de objetos en imágenes complejas mediante la red Boundary Guidance Network, la cual consiste en el ILD (Initial Localization Decoder), para encontrar la posición inicial de los objetos camuflajeados lo más exacto posible, y en el RRD (Residual Refinement Decoder), para refinar los detalles de los bordes de los objetos. En (Ji G., et al., 2022) se presenta la red ERRNet (Edge-based reversible re-calibration network) para la detección de objetos con confusión. Esta red presenta dos módulos: Selective Edge Aggregation y Reversible Re-calibration Unit para que la red logre una detección efectiva de los bordes y una comparación entre regiones potenciales camuflajeadas y el fondo. T. Chen et. al. (Chen T., et al., 2022) proponen una red llamada Boundary-guided Network (BgNet) que tiene el módulo Locating Module (LM) que se compone de una rama para obtener información contextual y otra rama que se utiliza para captar detalles de bajo nivel.

Las características de alto (información contextual) y bajo nivel se combinan, así si la información contextual es deficiente, las características de bajo nivel pueden dar información para lograr una correcta detección. También está el módulo Boundary-guided Fusion Module (BFM), el cual toma tres entradas: el mapa de características obtenido con LM y dos mapas de predicciones, uno que es la máscara del objeto y el otro la de los bordes; con estas entradas se hace un mejoramiento de características. Este mejoramiento tiene dos pasos, en el primero se calculan las características del fondo y del plano de interés y en el segundo se obtiene el resultado final. C. Zhang et. al. (Zhang C., et al., 2022) implementan dos módulos para la detección de objetos camuflajeados. Un módulo es el Neighbor Connection Mode (NCM) el cual concatena características de la capa previa, de la capa actual y de la capa siguiente; a estas características concatenadas se les llama características intermedias después de la fusión (CIF).

El otro módulo es el Hierarchical Information Transfer (HIT), cuya entrada son las CIF y su arquitectura consiste en 5 ramas: cuatro de ellas son convoluciones, suma y multiplicación de elementos, función sigmoide y al final de estas cuatro ramas se concatenan, y la otra rama implementa un aprendizaje residual y un mecanismo de atención para omitir características irrelevantes. El módulo NCM resalta los bordes y la información de posición de los objetos detectados mientras que el módulo HIT extrae características, con la combinación de ambos se realiza la detección. M. Zhuge et. al. (Zhuge M., et al., 2022) presentan la red CubeNet, que utiliza características jerárquicas de múltiples capas para la detección de objetos camuflajeados. La red CubeNet consiste en dos square fusión decoder (SFD) y en un sub edge decoder (SED). El SFD utiliza características de alto y bajo nivel extraídas por bloques encoder-decoder. El SED está localizado entre cada SFD para mejorar la representación de los bordes.

S. Liu et. al (Liu S., Liu D., et al., 2021) proponen un mecanismo para la actualización de plantillas para mejorar el seguimiento de objetos en fondos complejos. Este mecanismo utiliza múltiples plantillas para estimar la ubicación de los objetos. X. Huang et. al (Huang X., et al., 2021) presentan una red neuronal bio inspirada basada en un filtrado temporal neurodinámico y un filtrado espacial de Gabor. T. Zhang et. al (Zhang T., et al., 2020) proponen un enfoque que consiste en incrementar el conjunto de entrenamiento a partir de muestras no etiquetadas para localizar y clasificar animales salvajes en imágenes.

Dicho enfoque se basa en tres pasos: 1) entrenar un modelo (llamado modelo maestro) con muestras etiquetadas manualmente, como en el aprendizaje supervisado, 2) se ejecuta el modelo maestro sobre muestras no etiquetadas para seleccionar más muestras y 3) se entrena un modelo (llamado modelo estudiante) con las muestras etiquetadas manualmente y las muestras seleccionadas en el paso 2. S. Zhou et. al (Zhou S., et al., 2019) exponen una red neuronal profunda para la detección de personas camuflajeadas. La red aprende características discriminativas del primer plano de cada imagen. Una red neuronal llamada Foreground Attention Neural Network (FANN), mejora el primer plano y debilita el fondo.

Una subred guía la atención de la red, donde un codificador es utilizado para reconstruir una máscara binaria, y un decodificador enfoca su atención en las personas en el primer plano. K. Wang et. al (Wang K., et al., 2021) detectan objetos camuflajeados a partir de cómo el mecanismo visual humana lo hace: normalmente es difícil encontrar los objetos camuflajeados en una primera observación. Para eso se pasa a una segunda etapa, donde se realiza un análisis de la primera etapa. Con esto en mente se propone el modelo D2C-Net, que contiene dos módulos: Dual-branch Features Extraction (DFE) y Gradually Refined Cross Fusion (GRCF). El módulo DFE simula el mecanismo de dos etapas del sistema de visión del ser humano: primero se realiza una concatenación densa para agregar características de varios niveles para expandir el campo receptivo.

En el módulo GRCF se combinan características para mejorar el rendimiento de la detección. En (Kreim A., et al.) se propone un conjunto de datos llamado PTAW217Synth para la generación de muestras sintéticas para el seguimiento de personas en ambientes complicados (lluviosos, nevados, con mucha neblina), para mejorar el desempeño de algoritmos profundos de seguimiento de personas. En (Chen G., et al., 2022) se tiene la red Context-aware Cross-level Fusion Network (C2F-Net), dicha red fusiona características del contexto y características de nivel cruzado (esto es, a partir de las características más profundas se refinan las características de bajo nivel) para la detección de objetos camuflajeados.

En (Zheng Y., et al., 2019) se utiliza una red convolucional para detectar personas con camuflaje. Inicialmente se construye el conjunto de datos de gente camuflajeada en escenas naturales. Luego, con una red neuronal convolucional basada en la red VGG16, se extraen características semánticas y se introducen conexiones en la fase de deconvolución para lidiar con la influencia de patrones de camuflaje y el ruido en el fondo. Finalmente, se realiza una segmentación con super pixeles y se aplican restricciones espaciales de suavizado para mejorar la detección. En (Wang Y., et al., 2020) se detectan tanques en un entorno complejo (con mucha confusión entre el objeto y el fondo) con el detector YOLO utilizando pocas imágenes de entrenamiento. X. Wei et. al presentan en (Wei X., et al., 2020) el modelo profundo PftNet para la detección de personas en minas. Este modelo está basado en una red de transferencia de características en paralelo y está compuesto por dos módulos interconectados: el módulo de identificación, para ajustar la ubicación y el tamaño de los rectángulos delimitadores, y de localización, para que el modelo se adapte a varias escalas y relaciones de aspecto. En la Tabla 5 se comparte un resumen de los artículos revisados en esta subsección.

Artículo	Método	Artículo	Método
(Skoviera R., et al., 2018)	Red bioinspirada llamada Memoria Temporal Jerárquica (HTM). Se utiliza información de 3 HTM que se enfocan en características de ejes, textura y color. Además, el conjunto de datos fue generado para representa *cluttering*.	(Zhang T., et al., 2020)	Se utiliza la red Faster R-CNN y se consiguen rectángulos delimitadores a los cuales se les aplican transformaciones aleatorias.
(Xu X., et al., 2021)	Detección de contornos.	(Zhou S., et al., 2019)	Una red neuronal se enfoca en el primer plano y debilita el fondo para la detección de personas.
(Ji G., et al., 2022)	Detección de bordes y una comparación entre los objetos y el fondo.	(Wang K., et al., 2021)	Inspirado en la visión humana.
(Chen T., et al., 2022)	Con dos módulos: LM, para obtener información contextual e información de bajo nivel; y BFM para calcular las características del fondo y del objeto de interés.	(Kreim A., et al.)	Conjunto de datos para la detección de personas en ambientes complejos.
(Zhang C., et al., 2022)	Mediante los módulos NCM e HIT, para resaltar los bordes y la información de la posición de los objetos y la extracción de características, respectivamente.	(Chen G., et al., 2022)	A partir de características más profundas se mejoran las características de bajo nivel.
(Zhuge M., et al., 2022)	Bloques *encoder-decoder* extraen características de alto y bajo nivel, además se mejora la información de los bordes.	(Zheng Y., et al., 2019)	Con características semánticas y conexiones en la etapa de deconvolución para atender el camuflaje y el fondo.
(Liu S., Liu D., et al., 2021)	Múltiples *templates*.	(Wang Y., et al., 2020)	YOLO.
(Huang X., et al., 2021)	Red bioinspirada con un filtrado temporal neurodinámico y filtro de Gabor.	(Wei X., et al., 2020)	Transferencia de características.

Tabla 5. Resumen de los artículos que atienden el problema de confusión.

4.3 Información contextual

La información contextual se refiere a que el sistema de visión utilice la relación entre objetos de la imagen para poder comprender mejor la escena y detectar objetos. Con lo cual, el cómo adquirir información contextual en una imagen para poder localizar y clasificar objetos es un problema de interés en el área de detección de objetos.

Tras realizar un análisis de los artículos mostrados en la Tabla 6 en los cuales se presentan metodologías para la detección de objetos utilizando información contextual se encontró que todos siguen una metodología similar. Esto no es de extrañar, pues la información contextual se basa en encontrar la relación entre objetos de la imagen. Por eso, las metodologías consultadas tienen como idea fundamental el encontrar relaciones en la imagen. Estas relaciones se encuentran de diversas maneras: extrayendo características locales y globales, combinando las características locales y globales, utilizando características con más sentido semántico, proporcionando información de pixeles vecinos, usando objetos que se detecten con alta precisión para detectar otros objetos. Sin embargo, resulta que la extracción de características globales y locales y su combinación es la línea a para seguir para la obtención de información contextual. Los artículos revisados se describen a continuación.

En (She X., et al., 2021) se presenta una red profunda híbrida, ScieNet, con extracción de información contextual asistida por una red spike (SNN). ScieNet tiene una SNN con un algoritmo Spiking-timing-dependent plasticity (STDP) que extrae la información contextual. Luego se tiene una red profunda para la clasificación cuya entrada es la información contextual obtenida con STDP. S. Xie et. al. (Xie S., et al., 2020) presentan un método que integra características de diferentes campos receptivos para obtener una información contextual, mediante una red llamada diverse receptive field network (DRFNet). La utilización de varios módulos de campos receptivos en la DRFNet hace que regiones discriminativas en la vecindad del objeto de interés obtengan respuestas satisfactorias para la detección.

Esto indica que los módulos receptivos obtienen información contextual y detalles locales, con lo cual el detector puede distinguir mejor el fondo del objeto durante el entrenamiento de la red. X. Liang et. al (Liang X., et al., 2020) proponen el detector Feature Fusion and Scaling-based single shot detector (FS-SSD) para la detección de objetos pequeños con información contextual en imágenes aéreas. Además de las características profundas aprendidas por el FS-SSD, se mejora la precisión de la detección utilizando contexto espacial. La idea general del análisis del contexto espacial es utilizar los objetos detectados con un alto grado de precisión para objetos menos fiables. Q. Zhang et. al (Zhang Q., et al., 2020) muestran la red Dense Attention Fluid Network (DAFNet) para la detección de objetos sobresalientes en una imagen utilizando información contextual. DAFNet es una arquitectura encoder-decoder (basada en la red VGG16) donde en cada bloque de convolución se tiene un módulo llamado Global Context-aware Attention (GCA). El módulo GCA consiste en dos componentes: Global Feature Aggreation (GFA) y Cascaded Pyramid Attention (CPA). El módulo GFA toma las características obtenidas por la red implementada y produce características que codifican la información contextual global; este módulo tiene como objetivo lograr que las características se alineen y que exista un refuerzo mutuo entre los patrones sobresalientes al agregar relaciones semánticas globales entre pares de pixeles.

El módulo CPA se utiliza para abordar la variación en escala de los objetos; tiene como entrada las características obtenidas con el módulo GFA, el módulo CPA produce un mapa de atención refinado progresivamente. S. Wang et. al. (Wang S., et al., 2020) proponen un método que imita la percepción visual humana para la detección de navíos. En ese trabajo se desarrollan tres señales visuales para modelar el contexto de las imágenes: señal de escasez, debido a que los navíos solamente ocupan una pequeña porción de la escena; señal de contraste, ya que los navíos se diferencian notablemente del fondo; y señal de concentración, que se enfoca en la estructura local de los navíos. Una vez se tienen esas tres señales visuales, se combinan en un mapa de visibilidad para proveer ayudas visuales para la detección. En (Han L., et al., 2021) se propone la red Context and Structure Mining Network (CSMN) para la detección de objetos en video usando información del contexto. La CSMN consiste en un módulo de codificación de la información contextual espaciotemporal (stCIE) y un módulo de agregación de características de objetos basado en la estructura. La idea de stCIE consiste en que a cada mapa de características (obtenido con la red ResNet-101) de un pixel del objeto de interés se le agregan los pixeles vecinos tanto en el espacio como en el tiempo. En (Zhang L., et al., 2020) se propone aumentar una red neuronal feedforward utilizando un mecanismo de refinamiento en varias etapas.

En la primera etapa se construye una red maestra para generar un mapa de predicción en el cual faltan las estructuras más detalladas. En las etapas siguientes, la red de refinamiento con conexiones recurrentes a la red maestra combina información de contexto local a través de las etapas para refinar el mapa anterior. La información contextual propuesta se basa en que las redes neuronales en las capas más profundas tienen una fuerte consistencia semántica debido al largo campo receptivo, mientras que en las capas menos profundas las características codifican estructuras y detalles con poca consistencia semántica debido a su pequeño campo receptivo.

Para aprovechar lo antes mencionado, se aplica el módulo Channel Attention Mechanism (CAM), donde las características de alto nivel proveen una guía para extraer características discriminativas en las capas más bajas. También se implementa una capa global average pooling (GAP) en la cima de la red maestra para reunir información contextual global con un campo receptivo largo e introduce consistencia semántica en los siguientes módulos CAM. H. Luo et. al. (Luo H., et al., 2021) proponen el método Feature Fusion Semantic Information Enhancement (FFSI). Este método tiene dos módulos: el módulo de mejoramiento de información contextual (CIE, Context Information Enhancement) y el módulo de mejoramiento del campo receptivo (RFE, Receptive Field Enhancement).

El CIE resalta la ubicación de los objetos al establecer una relación entre la información contextual local y global. Mientras que el RFE mejora el campo receptivo al utilizar convolución dilatada para adaptar la detección de objetes de diferentes escalas, especialmente objetos pequeños. J. Guo et. al (Guo J., et al., 2020) se propone un detector de objetos con información contextual (EGCI-Net), el cual es una red de detección que tiene dos partes: una red backbone para extraer características y un modelo piramidal para enriquecer la información de contexto global de forma jerárquica. La información de contexto se obtiene con el módulo Pyramid Feature Pool (PFPM), que consiste en dos partes: el módulo pyramid pooling (PPM) y el módulo multi-scale feature pool (MFP). El PPM genera una priorización contextual mediante la fusión de un pooling de promedios multiescala, y el MFP genera información contextual piramidal mediante el uso de un pooling multiescala en el contexto global. Y. Zhu et. al (Zhu Y., et al., 2019) muestran la red Attention CoupleNet (ACoupleNet). Esta red utiliza la red ResNet-101 para la extracción de características. La red ACoupleNet funciona de la siguiente manera. Dada una imagen, se extraen características de los objetos utilizando atención en cascada.

Luego, con una red de regiones de propuesta (RPN) se generan las regiones donde puede estar un objeto. Cada propuesta va a dos ramas FCN (Fully Connected Networks) distintas: FCN local y FCN global, para extraer la información de la estructura global y aprender las partes específicas de objetos, respectivamente. Por último, en la salida se acoplan las dos ramas para predecir la clase del objeto. La FCN local captura diversas partes específicas (por ejemplo, en un rostro humano detecta la nariz, boca, etc., las cuales corresponden con regiones con respuesta alta en el mapa de características), lo cual refleja de manera efectiva las propiedades locales del objeto visual (siendo útil para lidiar con la oclusión o cuando el contorno del objeto está incompleto), mientras que la FCN global describe al objeto utilizando características de la región completa. Y. Gong et. al (Gong Y., et al., 2020) presentan la red Context Aware (CA-CNN) para la detección de objetos (la red neuronal convolucional es una Faster RCNN basada en VGG 16). La CA-CNN está constituida por la generación de propuestas, extracción de características de contexto, fusión de características y clasificación. Para extraer la información de contexto se propone una capa de extracción de regiones de interés, RDI, con contexto.

La RPN (red de generación de propuestas) conserva solamente las 256 mejores propuestas como RDI, cuyo tamaño es exactamente el de los objetos reales, por esto se generan las RDI con contexto para extraer información de contexto. G. Zhang et. al (Zhang G., et al., 2019) presenta una red, CAD-Net, para la detección de objetos con conciencia del contexto. La CAD-Net consiste en: 1) una red de contexto global (GCNet, Global Context Net) que aprende la correlación entre los objetos de interés y su correspondiente escena global, es decir, la correlación entre las características de los objetos y las características de la imagen total (la GCNet utiliza la red ResNet-101) y 2) la red Pyramid Local Context (PLCNet) que aprende la coocurrencia de características multiescala y/o la coocurrencia de objetos alrededor del objeto de interés (PLCNet utiliza Fast RCNN). W. Zhang et. al (Zhang W., et al., 2021) proponen un modelo llamado global context aware (GCA) para fortalecer la correlación espacial entre el fondo y el plano de interés fusionando la información de contexto global.

El módulo que se encarga de la información del contexto consiste en dos submódulos: módulo de atención y módulo de desvinculación de tareas. El módulo de atención se encarga de mapear la información de contexto global a un espacio de mayor dimensión y el módulo de desvinculación da el resultado de la clasificación y posicionamiento. En (Yuan Y., et al., 2019) se propone un método para la detección de señales de tráfico en ambientes complejos (se utiliza la red MobileNet como backbone).

El proceso principal del método propuesto consiste en dos componentes: el módulo de aprendizaje de características multiresolución, el cual combina diferentes niveles características semánticas con capas de deconvolución densamente conectadas; el otro módulo es el Vertical Spatial Sequence Attention (VSSA), que codifica la información vertical para una clasificación más exacta de las señales de tráfico. En (Siris A., et al., 2021) se propone un enfoque de aprendizaje con conciencia del contexto utilizando la red Mask-RCNN como backbone.

Específicamente, se proponen dos módulos: el módulo Semantic Scene Context Refinement (SSCR), para mejorar las características contextuales de los objetos sobresalientes de la imagen y el módulo Contextual Instance Transformer (CIT) para aprender las relaciones contextuales entre los objetos y la escena. H. Xie et. al (Xie H., et al., 2019) presentan un método de detección de peatones, Deconvolution Integrated Faster R-CNN (DIF R-CNN), que consta de tres componentes: generación inicial de mapas de características (que utiliza la red Inception-ResNet), módulo de deconvolución, que genera un mapa de características sintético y aporta información contextual, y por último el módulo de detección y región de propuesta. K. Zhang et. al. (Zhang K., et al., 2021) proponen una red semántica con conciencia del contexto (SCANet) para la detección de objetos en multiescala. Se diseñan dos módulos: Receptive Field-Enhancement (RFEM), que extrae características a diferente escala, y el módulo Semantic Context Fusion (SCFM), que combina características de alto nivel semántico con características de bajo nivel. En (Shi P., et al., 2023) proponen el detector Swin Deformable Tansformer-BiPAFPN-YOLOX.

Este detector tiene un mecanismo de atención llamado Reconstructed Deformable Self-Attention, el cual guía, mediante las regiones importantes en el mapa de características, las relaciones existentes entre parches de la imagen (así obtiene información contextual). Además, se tiene una red, BiPAFPN, para fusionar características multiescala y, por último, se utiliza la red YOLOX para realizar la clasificación y regresión usando las características obtenidas previamente. En la Tabla 6 se condensa la información previamente mencionada.

Artículo	Método	Artículo	Método
(She X., et al., 2021)	Red ScieNet que extrae información contextual y es asistida por una red *spike*.	(Gong Y., et al., 2020)	Extracción de regiones de interés con contexto.
(Xie S., et al., 2020)	Integración de distintas características de diferentes campos receptivos.	(Zhang G., et al., 2019)	Correlación entre los objetos de interés y la escena global.
(Liang X., et al., 2020)	Los objetos detectados con alta precisión se utilizan para detectar objetos menos fiables.	(Zhang W., et al., 2021)	Modelo *global context aware* que fusiona la información de contexto global.
(Zhang Q., et al., 2020)	Módulo consciente del contexto compuesto por una adición de características y atención en cascada piramidal.	(Yuan Y., et al., 2019)	El módulo de secuencia de atención espacial vertical para codifica la información vertical para la detección de señales de tráfico.
(Wang S., et al., 2020)	Inspirada en la visión humana para la detección de navíos.	(Siris A., et al., 2021)	Mejoramiento de las características contextuales y aprendizaje de las relaciones contextuales entre los objetos y la escena.
(Han L., et al., 2021)	A cada mapa de características de un pixel se le añaden los pixeles vecinos.	(Xie H., et al., 2019)	Generación de características de contexto con un módulo de deconvolución.
(Zhang L., et al., 2020; Luo H., et al., 2021)	Las características de alto nivel semántico son una guía para la extracción de características en las capas más bajas.	(Zhang K., et al., 2021)	Se tiene un módulo para mejorar el campo receptivo y otro que fusiona la información semántica.
(Guo J., et al., 2020)	Información del contexto con un módulo de características piramidales.	(Shi P., et al., 2023)	Se encuentran relaciones que existen entre parches de la imagen.
(Zhu Y., et al., 2019)	Una FCN local que captura diversas partes específicas y una FCN global que describe al objeto utilizando las características de toda la región.		

Tabla 6. Resumen de los artículos que atienden el problema de información contextual.

4.4 Cambios en la iluminación

El problema en el cambio en la iluminación consiste en que un objeto puede verse completamente diferente ante variaciones en la iluminación de una escena. Dado que la visión por computadora lidia con imágenes naturales, donde la iluminación es variable, es fundamental que el sistema de visión sea robusto ante los cambios en la iluminación. Algunos han demostrado que incrementar el conjunto de entrenamiento, de tal modo que los objetos de interés estén con diferente iluminación, hace más robusto al sistema de visión ante cambios de iluminación (Wu D., et al., 2020; Olarewaju M., 2021; Badeka E., et al., 2020; Li G., et al., 2020; Wu Y., et al., 2020), siendo esta la metodología más seguida, pues el modelo de sistema de visión es capaz de aprender a detectar objetos ante distintas iluminaciones. En otras investigaciones se ha propuesto mejorar la iluminación de las imágenes (Xiao B., et al., 2021; Zhou K., et al., 2020) o utilizar características invariantes a la iluminación (Mohanty S., et al., 2019) y características manuales (Kumar S., et al., 2021). También se ha explorado el uso de imágenes infrarrojas (Dai X., et al., 2021). En otros casos se encuentra que ante ciertos conjuntos de datos hay redes profundas que son robustas al cambio de iluminación (Afif M., et al., 2020). A continuación, se hace una breve descripción de los trabajos antes mencionados.

D. Wu et. al (Wu D., et al., 2020) utilizan la red YOLO v4 para detectar flores. El conjunto de entrenamiento con el cual se ajustaron los parámetros de la red contiene imágenes tanto de flores con iluminación frontal como trasera, haciendo que la red pueda clasificar flores correctamente ante varios cambios de iluminación. M. Olarewaju (Olarewaju M., 2021) modifica la red YOLOv3, para proponer la red YOLO-Tomato-A, la cual detecta tomates en imágenes naturales. Todas las imágenes del conjunto de muestras fueron obtenidas en ambientes y condiciones naturales: tomates con oclusión, traslape y distintas variaciones de iluminación. Este hecho hace que la red sea robusta ante la detección de tomates en imágenes con distinta iluminación. Un método similar se sigue en (Badeka E., et al., 2020), donde se entrena la red YOLOv3 con imágenes obtenidas en ambientes complicados, con oclusión, variaciones de iluminación, etc. En (Li G., et al., 2020) se utiliza la red Faster R-CNN para la detección de peatones en imágenes obtenidas por un dron. El conjunto de muestras para el entrenamiento contiene 1500 imágenes con varias condiciones climáticas, imágenes de día y de noche, haciendo el modelo más robusto ante cambios de iluminación. Asimismo, en (Wu Y., et al., 2020) se hace un aumento de datos en el conjunto de entrenamiento para la detección de señales de tráfico en distintas horas del día.

B. Xiao et. al. (Xiao B., et al., 2021) proponen un método para el seguimiento de máquinas de construcción en la noche. El método consta de cinco módulos: de mejoramiento de la iluminación, detección, seguimiento con filtros Kalman, asociación y asignación líneas. El módulo de mejoramiento de la iluminación es en el que se resuelve el problema de la variación de la iluminación en la detección de objetos, con lo cual nos enfocaremos únicamente en ese módulo. En este módulo se utiliza la red GLADNet (global illumination-aware and detail-preserving network) para la mejora de la iluminación en imágenes oscuras preservando la mayoría de los detalles. Esta red está dividida en dos etapas: estimación de la distribución de iluminación y la reconstrucción de detalles. En la primera etapa un mapa de características pasa por una red encoder-decoder para estimar la iluminación global en la imagen. La red encoder utiliza una CNN para submuestrear las características, mientras que la red decoder (que utiliza una CNN también) las sobremuestrea. La etapa de reconstrucción de detalles concatena los mapas de características de la primera etapa con los de la imagen de entrada. Luego, este mapa de características concatenado es procesado por tres capas convolucionales para preservar una mayor cantidad de detalles de la imagen de entrada.

La salida del módulo de mejoramiento de la iluminación es la entrada del módulo de detección. En (Zhou K., et al., 2020) se propone la red Modality Balance Network, MBNet, para la detección de peatones. La red MBNet consta de tres partes: extracción de características, modelo de alineación de características de iluminación (IAFA) y, por último, un mecanismo de iluminación. El módulo IAFA consiste en dos capas convolucionales y tres capas totalmente conectadas y adapta el modelo a diferentes condiciones de iluminación. El módulo IAFA minimiza la función de entropía cruzada entre los valores de iluminación predichos y los reales. En (Mohanty S., et al., 2019) se detectan objetos en movimiento en ambientes con variaciones de iluminación al extraer una característica invariante a la iluminación. Las premisas que se siguen en ese trabajo es que la intensidad de un pixel es el producto de la reflectancia y la iluminación de ese pixel, y que la iluminación en una región local de la imagen es suave, es decir, el valor de la iluminación de dos pixeles cercanos es parecido. S. Kumar et. al. (Kumar S., et al., 2021) realizan una detección del plano principal de una imagen, el cual es robusto ante variaciones de la iluminación. Para esto se extraen características de intensidad, color, homogeneidad local y entropía. Una vez que se tiene el vector de características, durante el modelado del fondo, se utiliza la distancia de Canberra para medir la similitud entre la media del vector de características del cuadro actual con el modelo. Finalmente, un algoritmo de selección adaptativo selecciona un valor de umbral para clasificar un pixel como fondo o plano principal.

En (Dai X., et al., 2021). se presenta la red TIRNet para la detección de objetos en imágenes infrarrojas térmicas (TIR). La utilización de las cámaras que obtienen imágenes TIR es debido a que las imágenes IR raramente son influenciadas por los cambios en la iluminación, así que, desde el tipo de imágenes utilizada, se soluciona en gran medida la detección de objetos ante variaciones en la iluminación. La TIRNet utiliza como backbone la red VGG, y una rama residual para obtener características más robustas para la regresión y clasificación. A partir de estas dos premisas se obtiene la invarianza ante la iluminación. M. Afifi et. al (Afif M., et al., 2020) detectan objetos en interiores utilizando el detector RetinaNet con varias redes como backbone, tales como ResNet, DenseNet y VGGNet. Este procedimiento presenta una alta capacidad de detección de los objetos interiores en condiciones de iluminación desafiantes. En la Tabla se resume la información recientemente descrita.

4.5 Objetos pequeños y cambios de escala

Este problema consiste en la dificultad de detectar objetos pequeños en una imagen, así como detectar el mismo objeto si aparece en diferentes tamaños.

Artículo	Método	Artículo	Método
(Wu D., et al., 2020)	El conjunto de entrenamiento tiene muestras con distinta iluminación.	(Mohanty S., et al., 2019) y (Kumar S., et al., 2021)	Extracción de características robustas ante cambios de iluminación.
(Wu D., et al., 2020)	El conjunto de entrenamiento tiene muestras con distinta iluminación.	(Mohanty S., et al., 2019) y (Kumar S., et al., 2021)	Extracción de características robustas ante cambios de iluminación.
(Olarewaju M., 2021; Badeka E., et al., 2020; Li G., et al., 2020; Wu Y., et al., 2020)	El conjunto de entrenamiento cuenta con imágenes con distinta iluminación.	(Dai X., et al., 2021)	Se utilizan imágenes térmicas.
(Xiao B., et al., 2021)	Se estima la distribución de la iluminación y se reconstruyen los detalles.	(Afif M., et al., 2020)	Se utiliza RetinaNet.
(Zhou K., et al., 2020)	Módulo de alineación de características de iluminación que minimiza una función de pérdida entre los valores de iluminación predichos y los reales.		

Tabla 7. Resumen de los artículos que resuelven el problema de la iluminación.

La idea más utilizada en las diversas metodologías es la combinación de características multiescala, que ayudan a obtener características del mismo objeto a diferentes tamaños y de información contextual (Bosquet B., et al., 2020; Leng J., et al., 2021; Zheng Q. y Chen Y., 2021; Zhang Y., et al., 2019; Zheng H., et al., 2020; Yan Z., et al., 2021; Li L., et al., 2021; Zhang S., et al., 2020; Fan M., et al., 2021; Luo H., et al., 2019; Ji S., et al., 2023). También hay metodologías que utilizan conjuntos de datos donde existen representaciones de diferente tamaño de los objetos de interés (Zhang H., et al., 2021; Kisantal M., et al., 2019). Una idea menos explorada ha sido la de predecir valores de pixeles para la detección de objetos pequeños (Lee G., et al., 2021) o el uso de Transformers (Xu S., Gu J., et al., 2023). A continuación, se describe trabajos que no solo atienden el problema de detección de objetos pequeños sino también cambios de escala.

Bosquet et. al (Bosquet B., et al., 2020) introducen la red neuronal convolucional STDnet, la cual es una red enfocada en la detección de objetos pequeños (los autores establecen un objeto pequeño como aquel que es menor a 16x16 pixeles). La red STDnet engloba 5 etapas: convoluciones iniciales, Region Context Network (RCN), convoluciones tardías, Region Proposal Network (RPN), y el clasificador. En las convoluciones iniciales la red STDnet aprende características básicas, luego, en la RCN se seleccionan las regiones con mayor probabilidad de contener objetos pequeños. Después se obtiene un mapa de características de las convoluciones iniciales, pero solamente para las regiones seleccionadas por la RCN.

A continuación, se extraen características con más información semántica en las convoluciones tardías, estas características son la entrada a la etapa de RPN, la cual localiza los objetos con mayor precisión. Finalmente, los objetos localizados son clasificados en su respectiva clase. Además de proponer la red STDnet, se presenta un conjunto de muestras, USC-GRAD-STDdb, que consiste en más de 56000 objetos pequeños etiquetados en escenarios desafiantes. Así pues, tanto con la red STDnet como con USC-GRAD-STDdb abordan el problema de la detección de objetos pequeños. La red profunda utilizada como backbone es la red ResNet-50. J. Leng et. al (Leng J., et al., 2021) proponen la red IENet (Internal-External Network), la cual utiliza tanto la apariencia del objeto pequeño como la información contextual para una detección más robusta. La IENet consiste en tres partes: aumento de características, generación de propuestas y clasificación. La etapa de aumento de características tiene un módulo, Bi-FMM (Bidirectional Feature Fusion Module), el cual detecta características internas del objeto. La etapa de generación de propuestas tiene el módulo CRM (Context Reasoning Module), el cual mejora la calidad de las generaciones de propuestas utilizando información contextual.

Por último, la etapa de clasificación incluye el módulo CFAM (Context Feature Augmentation Module), este módulo aprende relaciones entre las propuestas dadas por el CRM, y esas relaciones son utilizadas para producir información global asociada con las regiones propuestas para mejorar la clasificación. La red IENet fue evaluada con los conjuntos de datos MS COCO y WIDER FACE. Q. Zheng e Y. Chen (Zheng Q. y Chen Y., 2021) proponen una estrategia para el mejoramiento de características multiescala. El enfoque que ellos utilizan tiene dos partes: 1) el módulo multi-scale auxiliary enhancement network, el cual mejora las características al introducir características detalladas de bajo y medio nivel de la red original (VGG-16), y el módulo Adaptive Interaction, el cual mejora el desempeño del modelo al utilizar las características mejoradas. De este modo, al utilizar múltiples escalas, obtiene un buen desempeño en la detección de objetos pequeños. En (Zhang Y., et al., 2019) se enfocan en la detección de rostros que aparecen con un tamaño pequeño en una imagen. El método en ese trabajo consiste en generar una imagen clara y de alta resolución de un rostro con una red GAN, a partir de una imagen pequeña y borrosa. Además, se utiliza información contextual en combinación con la red GAN propuesta. La información contextual se utiliza debido a que los seres humanos, al buscar un rostro, no solamente se enfocan en el rostro, sino que buscan cabello, cuello, incluso el cuerpo, sobre todo en situaciones donde es difícil encontrar un rostro. H. Zheng et al (Zheng H., et al., 2020) abordan la detección de objetos a diferentes escalas de la siguiente manera: primeramente, se obtienen características basadas en gradientes orientados para obtener orientaciones locales discriminativas.

Luego, se utiliza la convolución dilatada para aumentar la resolución espacial del mapa de características, lo que ayuda a la detección multiescala. Z. Yan et al. (Yan Z., et al., 2021) proponen un detector de una etapa llamado LocalNet. LocalNet tiene como backbone la red long neck ResNet, la cual preserva información más detallada en las etapas iniciales para mejorar la representación de objetos pequeños. Esto debido a que mejora los campos receptivos de las relaciones de aspecto múltiple (multiescala) durante la extracción de características.

En (Li L., et al., 2021) se emplea la red Attention Feature Pyramid Transformer Network (AFPN), la cual aprende a detectar objetos multiescala mediante unos mapas de características multiescala. S. Zhang et. al (Zhang S., et al., 2020) presentan la red AMS-Net, Assymetry multi-stage nerowrk, para la detección de peatones a diferentes escalas. En términos generales, la AMS-Net está compuesta por tres etapas: la etapa 1 tiene como entrada una imagen piramidal, esto es, una misma imagen a múltiples escalas diferentes, a la cual se le extraen características multiescala e información contextual para obtener las regiones de propuesta (donde puede estar un objeto); en la etapa 2 todas las regiones de propuesta entran a una CNN que utiliza un mecanismo de aumento de características globales para mejorar la clasificación, además, en esta etapa se rechazan muchos falsos positivos.

Finalmente, en la etapa 3 se mejora la localización de los objetos con una red más profunda. M. Fan et. al (Fan M., et al., 2021) exponen un método para la detección de objetos pequeños en imágenes infrarrojas. El método consiste en mejorar la intensidad de los objetos con características de intensidad locales, luego se realiza una detección de esquinas. A continuación, las regiones donde posiblemente haya un objeto entran al clasificador (CNN). En (Luo H., et al., 2019) se utiliza un algoritmo que fusiona información contextual y la red YOLOV3 llamado Contextual-YOLOV3, la detección de objetos pequeños se lleva a cabo al combinar una FPN multiescala con información contextual. S. Ji et. al. (Ji S., et al., 2023) proponen un modelo, MCS-YOLO v4, para la detección de objetos pequeños. Este modelo utiliza información contextual multiescala y se propone la función de costo Soft-CIOU. Las características de los objetos pequeños se combinan con características contextuales para hacer la extracción de características más robusta. La función de costo Soft-CIOU se utiliza porque en el proceso de ajustar los rectángulos delimitadores pequeños cambios tienen un gran impacto (ya que los objetos a detectar son pequeños).

H. Zhang et. al (Zhang H., et al., 2021) generan una base de datos en la cual se introducen numerosas muestras de imágenes UAV a diferentes distancias para que los objetos aparezcan con un tamaño distinto. De esta manera, algún detector podría detectar, al ser entrenado con esa base de datos, la misma clase de objeto a diferentes escalas. Similarmente, M. Kisantal et. al (Kisantal M., et al., 2019) realizan un aumento de datos en el conjunto MS COCO para que haya más objetos pequeños en las imágenes.

G. Lee et. al (Lee G., et al., 2021) comparten la red FEN (feature enhancement network) basada en la arquitectura U-Net, con el objetivo de enriquecer las características de las regiones donde están los objetos pequeños en imágenes que tienen ruido. Primeramente, se extraen características de imágenes con ruido, luego, de manera aleatoria, se eliminan algunos valores de esas las características extraídas. Y son estas características la entrada de la FEN. El objetivo de la FEN es predecir los valores eliminados de las características utilizando los valores de los pixeles cercanos. Es esto lo que permite obtener mejores características en objetos pequeños presentes en imágenes con ruido.

En (Xu S., Gu J., et al., 2023) se propone la red DKTNet. Esta red utiliza las redes ResNet50 y FPN para extraer características de la imagen de entrada. Estas características son la entrada de un dual-key transformer para aumentar la correlación entre Q, query, y V, value, lo cual mejora la discriminación de las características. En la Tabla 7 se presenta un resumen de la subsección sobre detección de objetos pequeños.

Artículo	Método	Artículo	Método
(Bosquet B., et al., 2020)	Se presenta la red enfocada en la detección de objetos pequeños Red STDnet, así como el conjunto de datos USC-GRAD-STDdb, con 56000 objetos pequeños etiquetados en ambientes desafiantes.	(Zhang S., et al., 2020; Luo H., et al., 2019)	Información contextual junto con características de múltiples escalas.
(Leng J., et al., 2021)	Red *Internal-External Network* (IENet), que utiliza tanto las características del objeto como información contextual.	(Fan M., et al., 2021)	Características de intensidad local y detección de esquinas en imágenes infrarrojas.
(Zheng Q. y Chen Y., 2021)	Características de múltiples escalas.	(Ji S., et al., 2023)	Información contextual y la función de costo Soft-CIOU.
(Zhang Y., et al., 2019)	Combinación de una red GAN (que obtiene una imagen de alta resolución) con información contextual para la detección de rostros.	(Zhang H., et al., 2021)	Base de datos de imágenes UAV a diferentes alturas.
(Zheng H., et al., 2020)	Convolución dilatada para aumentar la resolución espacial del mapa de características.	(Kisantal M., et al., 2019)	Aumento de datos en el conjunto MS COCO para que aparezcan más objetos pequeños.
(Yan Z., et al., 2021)	Mejoramiento de los campos receptivos de las relaciones de multiescala.	(Lee G., et al., 2021)	Red *feature enhancement newtork* (FEN), la cual mejora las características de las regiones donde se ubican los objetos pequeños.
(Li L., et al., 2021)	Características multiescala.	(Xu S., Gu J., et al., 2023)	*Transformer* para mejorar la discriminación de las características.

Tabla 7. Resumen de los artículos que atienden el problema de detección de objetos pequeños.

4.6 Variación entre clases y la misma clase

Este problema consiste en que puede haber dos clases distintas que se asemejen, o bien, una misma clase con objetos que se aprecien muy diferentes. Diversos estudios se han realizado para resolver este problema, donde la metodología más utilizada ha sido la modificación de las funciones de pérdida de las redes profundas (Sun B., et al., 2021; Yuan J., et al., 2020; Li C., et al., 2019) para maximizar las diferencias entre las clases. El uso de nuevos paradigmas ha sido de gran utilidad, como Few-Shot (Cao Y., et al., 2021; Li B., et al., 2021), en el cual puede realizar una correcta detección aun con pocas muestras. Una discriminación que parte de lo general a lo específico ha sido estudiada en (Shin D., et al., 2020), la idea central son los árboles de decisión. El aumento del conjunto de datos también ha sido de utilidad para resolver este problema (Li K., et al., 2020). Algunos de los trabajos relacionados se comentan a continuación.

En (Sun B., et al., 2021) se propone una función de pérdida, contrastive proposal encoding loss, la cual mide la similitud semántica entre las regiones de propuestas. X´+ representa las propuestas positivas, es decir, las regiones de la misma categoría (mismo objeto) y X* corresponde a las regiones negativas, que son aquellas que tienen diferentes categorías. La función de pérdida debe ser tal que score(f(X),f(X+)) >>score(f(X),f(X*)), es decir, se debe tener una varianza pequeña entre la misma categoría y mayor entre clases. Algo parecido se presenta en (Yuan J., et al., 2020) donde se utiliza una CNN para la detección de objetos. En ese trabajo se presenta una función de pérdida llamada Inter-class loss, la cual ayuda al algoritmo a aprender información discrepante entre categorías, lo cual hace que detecte de mejor manera objetos de la misma categoría. C. Li et. al. (Li C., et al., 2019), modifican la función de costo de la red YOLOv3, de una función logística a una softmax para maximizar la diferencia de las características entre las clases.

Y. Cao et. al. (Cao Y., et al., 2021) y B. Li et. al. (Li B., et al., 2021) presentan trabajos en los cuales se detectan objetos de clases nuevas a partir de pocas muestras mediante el enfoque de aprendizaje Few-Shot, y así resuelven el problema de la varianza entre clases. Y. Cao et. al. (Cao Y., et al., 2021) resuelven el problema de la varianza entre clases mediante asociación y discriminación. Durante la asociación se construye un espacio de características que imita un espacio de características de una clase base, el nuevo espacio se asocia a la clase base de acuerdo con su similitud semántica. En la discriminación se asegura la separabilidad de las clases con un margen de pérdida. En B. Li et. al. (Li B., et al., 2021) se presenta un margen de equilibrio de clases (CME) que optimiza tanto la partición del espacio de características y la reconstrucción de las clases nuevas. CME reserva un espacio adecuado de las nuevas clases al utilizar un margen de pérdida durante el aprendizaje.

En (Shin D., et al., 2020) se utiliza un enfoque basado en árboles de decisión donde la primera rama detecta clases amplias, esto quiere decir clases como «animales» o «vehículos», de esta manera se separa la diferencia entre clases, luego se detecta la clase de los objetos para cada nodo (clase), esto es, por ejemplo, de la clase «animales» surgen ramas distintas para la clase «perro», «gato», etc. Finalmente, se tienen subclases donde se detecta un único objeto de la clase.

K. Li et. al (Li K., et al., 2020) presentan en su artículo un conjunto de muestras de entrenamiento que, entre otras cualidades, tiene una adición de muestras para aumentar la diversidad entre la misma clase. En la Tabla 8 se recapitula la información anteriormente escrita.

Artículo	Método	Artículo	Método
(Sun B., et al., 2021; Yuan J., et al., 2020; Li C., et al., 2019)	Se propone una función de pérdida.	(Shin D., et al., 2020)	Enfoque basado en árboles de decisión.
(Cao Y., et al., 2021; Li B., et al., 2021)	Construcción de un mapa de características.	(Li K., et al., 2020)	Mejoramiento del conjunto de entrenamiento.

Tabla 8. Resumen de los artículos que atienden el problema de detección de objetos con variación entre la misma clase y entre diferentes clases.

4.7 Deformación y cambios de pose

Este problema consiste en que, si un objeto cambió de posición o se ve deformado por cualquier razón, podría provocar que un sistema de visión ya no sea capaz de detectarlo de manera correcta.

La tendencia para resolver este problema está en el aprender múltiples perspectivas del objeto (Zhang X., et al., 2022; Fang F., et al., 2019) y la extracción de características en regiones que correspondan a la forma real del objeto (Mordan T., et al., 2019; Zhang P., et al., 2020; Liu Y., et al., 2021; Dai J., et al., 2017). Es decir, las ideas de estas dos metodologías son similares: extraer características que ayuden a tener un mayor conocimiento de las formas de los objetos. Estos enfoques resultan de mayor interés para resolver este problema que otros presentados en la literatura (Das A., et al., 2022). A continuación, se presentan los trabajos consultados que resuelven el problema de deformación y cambios de pose.

X. Zhang et al. (Zhang X., et al., 2022) presentan una red sensible a la perspectiva, PSNet, para la detección de objetos. La red PSNet aprende múltiples espacios de perspectiva, donde en cada uno de estos, las características semánticas son desacopladas para hacerla robusta ante cambios de perspectiva.

El cambio de perspectiva indica que el objecto aparece en diversas poses, rotaciones, etcétera; con lo cual, es una manera de solucionar el problema de detectar un objeto que se vea diferente en distintas imágenes. En (Fang F., et al., 2019) se presenta un enfoque híbrido que combinan la red Faster R-CNN con un modelo de agrupamiento para la detección de objetos alargados. T. Mordan et. al (Mordan T., et al., 2019) parten de la base de que normalmente, en la detección de objetos, se utilizan rectángulos delimitadores, de los cuales se extraen características sin importan la forma real del objeto, debido a esto introducen la red DP-FCN, la cual aplica deformaciones en las regiones para aprender mejores representaciones que se ajusten de una manera más adecuada a la forma real del objeto.

Esta red tiene dos contribuciones importantes: invarianza a transformaciones locales y la obtención de información geométrica para describir más fielmente a los objetos. P. Zhang et. al (Zhang P., et al., 2020) detectan y siguen objetos sobresalientes, su enfoque utiliza mapas sobresalientes espaciotemporales, y para lidiar con objetos sensibles a escalas (y, por extensión de acuerdo con los autores, a objetos no rígidos) se implementa una FCN. Enfocándonos en la parte del contexto deformable, se utiliza una convolución deformable para la localización de los rectángulos delimitadores de objetos no rígidos presentada en (Dai J., et al., 2017).

Esta convolución es embebida en la estructura multiescala añadiendo un desplazamiento 2D de la convolución estándar, lo cual resulta en estructuras adaptativas a las formas geométricas en los kernels de convolución. Y. Liu et. al (Liu Y., Duanmu M., et al., 2021) proponen el módulo Multi-Scale Deformation Module (MSDM) para obtener pistas visuales multiescala y de formas variantes de objetos sobresalientes. Los autores utilizan una modificación de la red ResNet-50 como backbone. MSDM consiste en dos pasos: el primero es la extracción de contexto multiescala, donde se aprenden características multiescala, y el segundo es la extracción del contexto deformable. Combinando ambos pasos, el MSDM es capaz de capturar información a diferentes escalas y diversas formas.

En (Das A., et al., 2022) se presenta un detector de peatones en diferentes poses en donde se extraen características de peatones y se realiza una segmentación por instancias, después un codificador aprende características específicas en peatones dependiendo de su pose. En la Tabla se resume la información presentada sobre la detección de objetos ante deformaciones y cambios de pose.

5. Conclusiones y trabajo futuro

La detección de objetos es una de las etapas más importantes de visión por computadora. Por esta razón, se han desarrollado una gran cantidad de modelos de detección de objetos que presentan buenos resultados en ciertos conjuntos de muestras, en clases específicas, en ambientes restringidos, y utilizando

Artículo	Método	Artículo	Método
(Zhang X., et al., 2022)	Red PSNet, que aprende diferentes perspectivas.	(Liu Y., Duanmu M., et al., 2021)	Convolución deformable para la localización de objetos de no rígidos.
(Fang F., et al., 2019)	Red DP-FCN, que aplica transformaciones en las regiones en vez de utilizar solamente rectángulos delimitadores.	(Fang F., et al., 2019)	Combinación de la red Faster R-CNN con un módulo de agrupamiento.
(Zhang P., et al., 2020)	Mapas de características sensible a múltiples escalas.	(Das A., et al., 2022)	Segmentación por estancias y extracción características de peatones en función de su pose.

Tabla 10. Resumen de los artículos que atienden el problema de detección de objetos ante deformaciones y cambios de pose.

e implementado una de entre tantas ideas y posibilidades. Sin embargo, la detección de objetos sigue siendo una tarea compleja, ya que presenta diversos retos como los mencionados en este artículo.

A partirdel análisis realizado en este trabajo, resulta evidente que se debería incrementar el número de clases a detectar, pues existen aplicaciones en las cuales se tienen varias clases de interés. También se observó que, para resolver algunos retos como la oclusión, cambios de iluminación o detección de objetos pequeños y cambios de escala es típico aumentar el conjunto de datos para hacer al detector más robusto; sin embargo, sería más interesante y contribuiría más el aportar y aplicar ideas en los modelos que solucionen dichos problemas.

Por otra parte, como resultado del análisis de la detección de objetos con confusión, resulta evidente que es fundamental la detección en ambientes más diversos y distintos, y también es crucial el estudio de las relaciones del fondo con los objetos de interés. Además, debido a que la tendencia es la detección de bordes, el desarrollo en esta área será de gran utilidad.

Basándonos en el análisis de la detección de objetos utilizando información contextual se desprende que, a pesar de su correcta aplicación, es indispensable asegurar que exista una relación útil entre objetos y/o el objeto y el fondo, ya que no suele presentarse este análisis en los trabajos reportados en este artículo y es la relación que existe entre elementos de la imagen lo que permite utilizar la información contextual. Esta relación no solamente debe darse entre los objetos, sino en atributos presentes en la imagen (por ejemplo, si hubiera una relación entre el origen de una fuente de iluminación y la manera en que un objeto proyecta una sombra, se debería analizar si esta relación sirve como información contextual).

Considerando los métodos para detectar objetos con variación entre clases y entre la misma clase, se concluye que sería benéfico el uso de otros paradigmas de aprendizaje para la generalización entre clases y subclases, que, a pesar de no ser la metodología más seguida, es la que mayor impacto podría tener.

De acuerdo con las ideas presentadas en la detección de objetos con deformaciones y cambios de pose se concluye que la forma en la que el modelo extrae características que se asemejen más a la forma del objeto es un área interesante de investigación. Debido a esto, la implementación de redes profundas que encuentren relaciones espaciales (como las redes tipo cápsula) sería un gran beneficio para resolver este problema.

Con base en el análisis presentado en este artículo se mantiene el énfasis en que existen varias soluciones a un mismo problema, y como se aprecia en diversos artículos publicados, algunos solucionan un problema al resolver otro. Con lo cual se hace evidente que la integración es un proceso fundamental en la detección de objetos.

Por lo anterior, concluimos que un posible tema de la investigación futura en la detección de objetos se debería concentrar, entre otros enfoques y/o métodos que se comentan en diversos estudios (como los mencionados en la sección 2), en tratar de adoptar ideas y mecanismos realizados por el ser humano al detectar objetos. Esto se propone porque la visión humana tiene un buen desempeño para reconocer patrones, obtener información del contexto y localizar objetos en escenas complicadas.

Agradecimientos

Los autores agradecen el financiamiento del Tecnológico Nacional de México/I.T. Chihuahua bajo el proyecto 16431.23-P para realizar este proyecto.

REFERENCIAS

Abbas S., et al. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing*, 103514.

Adouani A., et al. (2019). Comparision of Haar-like, HOG and LBP approaches for face detection in video sequences. *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)* (págs. 266-271). Estambul, Turquía: IEEE.

Adreopoulos A. y Tsotsos J. (2013). A Computational Learning Theory of Active Object Recognition Under Uncertainty. *International Journal of Computer Vision*, 95-143.

Afif M., et al. (2020). An Evaluation of RetinaNet on Indoor Object Detection for Blind and Visually Impaired Persons Assistance Navigation. *Neural Processing Letters*, 2265-2279.

Ajmera F., et al. (2021). Survey on Object Detection in Aerial Imagery. *IEEE Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (págs. 1050-1055). Tirunelveli, India: IEEE.

Ansari S. (2019). A Review on SIFT and SURF for Underwater Image Feature Detection and Matching. *2019 IEEE International Conference on Electrical, Computer and Communication Technologies* (págs. 1-4). IEEE.

Arfi A., et al. (2020). Real Time Human Face Detection and Recognition Based on Haar Features. *2020 IEEE Region 10 Symposium (TENSYMP)* (págs. 517-521). IEEE.

Arnold E., et al. (2019). A Survey on 3D Object Detection Methods for Autonomous Driving Applications. *IEEE Transactions on Intelligent Transportation Systems*, 3782-3795.

Arulprakash E. y Aruldoss M. (2021). A study on generic object detection with emphasis on future research. *Journal of King Saud University-Computer and Information Sciences*.

Arunmozhi A. y Park J. (2018). Comparision of HOG, LBP and Haar-Like Features for on-Road Vehicle Detection. *2018 IEEE International Conference on Electro/Information Technology (EIT)* (págs. 362-367). IEEE.

Aziz L., et al. (2020). Exploring Deep Learning-Based Architecture, Strategies, Applications and Current Trends in Generic Object Detection: A Comprehensive Review. *IEEE Access*, 170461-170495.

Badeka E., et al. (2020). Harvest Crate Detection for Grapes Harvesting Robot Based on YOLOv3 Model. *2020 Fourth International Conference on Intelligent Computing in Data Sciences (ICDS)*, (págs. 1-5).

Bastanfard A., et al. (2019). Improving Tracking Soccer Players in Shaded Playfield Video. *2019 5th Iranina Conference on Signal Processing and Intelligent Systems (ICSPIS)* (págs. 1-8). IEEE.

Bay H., et al. (2006). SURF: Speeded Up Robust Features. *European Conference on Computer Vision (ECCV)* (págs. 404-417). Springer.

Biswas R., et al. (2021). A new perceptual hashing method for verification and identity classification of occluded faces. *Image and Vision Computing*, 104186.

Biswas S., et al. (2021). Beyond document object detection: instance-level segmentation of complex layouts. *International Journal on Document Analysis and Recognition*, 269-281.

Bochkovskiy A., et al. (2020). YoloV4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Borji A., et al. (2019). Salient object detection: A survey. *Computational Visual Media*, 117-150.

Bosquet B., et al. (2020). STDnet: Exploiting high resolution feature maps for small object detection. *Engineering Applications of Artificial Intelligence*, 103615.

Cao Y., et al. (2021). Few-Shot Object Detection Via Assosciation and Discrimination. *Advances in Neural Information Processing*.

Cen F., et al. (2021). Deep feature augmentation for occluded image classification. *Pattern Recognition*, 107737.

Chapel M. y Bouwmans T. (2018). New trends on moving object detection in video images captured by a moving camera: A survey. *Computer Science Review*, 157-177.

Chapel M. y Bouwmans T. (2020). Moving objects detection with a moving camera: A comprehensive review. *Computer Science Review*, 100310.

Chen G., et al. (2022). Camouflaged Object Detection via Context-aware Cross-level Fusion. *IEEE Transactions on Circuit Systems and Systems for Video Technology*.

Chen T., et al. (2022). Boundary-guided network for camouflaged object detection. *Knowledge-Based Systems*, 108901.

Chen Z., et al. (2020). Underwater salient object detection by combining 2D and 3D visual. *Neurocomputing*, 239-259.

Cheng G., et al. (2022). Towards Large-Scale Small Object Detection: Survey and Benchmarks. *arXiv preprint arXiv:2207.13096*.

Chilukuri D., et al. (2022). A robust object detection system with occlusion handling for mobile devices. *Computational Intelligence*.

Cygert S. y Czyzewski A. (2020). Towards Robust Pedestrian Detection With Data Augmentation. *IEEE Access*, 126674-126683.

Dai J., et al. (2016). R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 379-387.

Dai J., et al. (2017). Deformable convolutional networks. *Proceedings of the IEEEE International Conference on Computer Vision*, (págs. 764-773).

Dai K., et al. (2019). Visual Tracking via adaptive spatially-regularized correlation filters. *Proceedings of the IEEE conference on computer vision and pattern recognition* (págs. 4670-4679). IEEE.

Dai X., et al. (2021). TIRNEt: Object detection in thermal infrared images for autonomous driving. *Applied intelligence*, 1244-1261.

Dalal N. y Triggs B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recogntion (CVPR'05)* (págs. 886-603). IEEE.

Das A., et al. (2022). Deep Multi-Task Networks For Occluded Pedestrian Pose Estimation. *arXiv preprint arXiv:2206.07510*.

Dash P. y Sigappi A. (2018). Detection and Classification of Retinal Diseases in Spectral Domain Optical Coherence Tomography Images based on SURF descriptors. *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA)* (págs. 1-6). IEEE.

Deng J., et al. (2020). A review on reseatch on object detection based on deep learning. *Journal of Physics: Conference Series*, 12028.

Diwan T., et al. (2023). Object detection using YOLO: challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications*, 9243-9275.

Dong J., et al. (2020). Occlusion-Aware GAN for Face De-Occlusion in the Wild. *2020 IEEE International Conference on Multimedia and Expo (ICME)* (págs. 1-6). IEEE.

Dong W., et al. (2021). Ellipse R-CNN: Learning to Infer Elliptical Object from Clustering and Occlusion. *IEEE Transactions on Image Processing*, 2193-2206.

Du S. y Wang S. (2022). An Overview of Correlation-Filter-Based Object Tracking. *IEEE Transactions on Computational Social Systems*, 18-31.

Espinosa J., et al. (2020). Detection of Motorcycles in Urban Traffic Using Video Analysis: A Review. *IEEE Transactions on Intelligent Transportation Systems*, 6115-6130.

Fan M., et al. (2021). Infrared small target detection based on region proposal and CNN classifier. *Signal, Image and Video Processing*, 1927-1936.

Fang F., et al. (2019). Combining Faster R-CNN and Model-Driven Clustering for Elongated Object Detection. *IEEE Transactions on Image Processing*, 2052-2065.

Feng Y., et al. (2021). Detect Faces Efficiently: A Survey and Evaluations. *Transactions on Biometrics, Behavior, and Identity Source*.

Flores M., et al. (2019). Pedestrian Detection Under Partial Occlusion by using Logic Inference, HOG and SVM. *IEEE Latin America Transactions*, 1552-1559.

Geetha S., et al. (2021). Machine Vision Based Fire Detection Techniques: A Survey. *Fire Technology*, 591-623.

Girshick R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (págs. 1440-1448). IEEE.

Girshick R., et al. (2014). Rich feature hierarchies for accurate object detection and semantics segmentation. *Proceedings of the IEEE conference on computer vision and pattern recogntion* (págs. 580-587). IEEE.

Gkixioari G., et al. (2019). Mesh r-cnn. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (págs. 9785-9795). IEEE.

Gong Y., et al. (2020). Context-Aware Convolutional Neural Network for Object Detection in VHR Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 34-44.

Guo J., et al. (2020). Object detector with enriched global context information. *Multimedia Tools and Applications*, 29551-29571.

Guo T., et al. (2020). Detection of Ice Hockey Players and Teams via a Two-Phase Cascaded CNN Model. *IEEE Access*, 195062-195073.

Gupta S., et al. (s.f.). Improved object recognition results using SIFT and ORB feature detector. *Multimedia*.

Gurbina M., et al. (2019). Tumor Detection and Classification of MRI Brain Image using Different Wavelet Transforms and Support Vector Machines. *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)* (págs. 505-508). IEEE.

Hajizadeh M., et al. (2023). MobileDenseNet: A new approach to object detection on mobile devices. *Expert Systems with Applications*.

Han B., et al. (2020). Small-Scale Pedestrian Detection Based on Deep Neural Network. *IEEE Transactions on Intelligent Transportation Systems*, 3046-3055.

Han J., et al. (2018). Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey. *IEEE Signal Processing Magazine*.

Han L., et al. (2021). Context and Structure Mining Network for Video Object Detection. *International Journal of Computer Vision*, 2927-2946.

He K., et al. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recogntion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1904-1916.

He K., et al. (2017). Mask r-cnn. *Proceedings of the IEEE international conference on computer vision* (págs. 2961-2969). IEEE.

Heo J., et al. (2022). Occlusion-aware spatial attention transformer for occluded object recognition. *Pattern Recognition Letters*, 70-76.

Huang G., et al. (2019). Ship detection based on squeeze excitation skip-connection path networks for optical remote sensing images. *Neurocomputing*, 215-223.

Huang G., et al. (2023). A Survey of Self-Supervised and Few-Shot Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4071-4089.

Huang X., et al. (2021). A bioinspired retianl neural network for accurately extracting small-target motion information in clutter backgrounds. *Image and Vision Computing*, 104266.

Jarunakarint V., et al. (2020). Survey and Experimental Comparision of Machine Learning Models for Motorcycle Detection. *2020 5th International Conference on Information Technology (InCIT)* (págs. 320-325). IEEE.

Ji G., et al. (2022). Fast Camouglaged Objec Detection via Edge-based Reversible Re-calibration Network. 108414.

Ji S., et al. (2023). An improved algorithm for small object detection based on YOLO v4 and multi-scale contextual information. *Computers and Electrical Engineering*, 108490.

Jiao L., et al. (2019). A Survey of Deep Learning-Based Object Detection. *IEEE Access*, 128837-128868.

Jocher G. (22 de 4 de 2020). *YOLOv5 by Ultralytics*. Obtenido de https://github.com/ultralytics/yolov5

Jocher G., et al. (22 de 4 de 2023). *YOLO by Ultralytics*. Obtenido de Available: https://github.com/ultralytics/ultralytics

Kaur M. y Min C. (2018). Automatic Crop Furrow Detection for Precision Agriculture. *2018 IEEE 61st International Mildwest Symposium on Circuits and Systems (MWSCAS)* (págs. 520-523). IEEE.

Kaur P., et al. (2020). A survey on brain tumor detection techniques for MR images. *Multimedia Tools and Applications*, 21771-21814.

Kaushal M., et al. (2018). Soft computing based object detection and tracking approaches: State-of-the-Art survey. *Applied Soft Computing*, 423-464.

Kellenberger B., et al. (2019). Half a Percent of Labels is Enough: Efficient Animal Detection in UAV Imagery Using Deep CNNs and Active Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 9524-9533.

Kisantal M., et al. (2019). Augmentation for small object detection. *arXiv preprint*.

Köhler M., et al. (2023). Few-Shot Object Detection: A Comprehensive Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21.

Kreim A., et al. (s.f.). Using syntethic data for person tracking under adverse weather conditions. *Image and Vision Computing*, 104187.

Kumar N. (2018). Thresholding in salient object detection: a survey. *Multimedia Tools and Multimedia*, 19139-19170.

Kumar S., et al. (2021). An improved scheme for multifeature-based foreground detection using challenging conditions. *Digital Signal Processing*, 103030.

Lang W., et al. (2021). A survey of 3D object detection. *Multimedia Tools and Applications*, 29617-29641.

Lee G., et al. (2021). Self-Supervised Feature Enhancement Networks for Small Object Detection in Noisy Images. *IEEE Signal Processing Letters*, 1026-1030.

Leng J., et al. (2021). Realize your surroundings: Expliting context information for small object detection. *Neurocomputing*, 287-299.

Li B., et al. (2021). Beyond Max-Margin Class Margin Equilibrium for Few-Shot Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (págs. 7363-7372).

Li C. et. al. (2022). YOLOV6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv preprint arXiv:2209.02976*.

Li C., et al. (2019). Face Detection Based on YOLOv3. *Recent Trends in Intelligent Computing, Communication and Devices*, 277-284.

Li F., et al. (2022). Faster R-CNN with Generative Adversarial Occlusion Network for Object Detection. *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, (págs. 536-541).

Li G., et al. (2020). Faster R-CNN Deep learning Model for Pedestrian Detection from Drone Images. *SN Computer Science*.

Li J., et al. (2019). Multi-scale HOG feature used in object detection. *Tenth International Conference on Graphics and Image Processing (ICGIP).*

Li K., et al. (2020). Object detection in optical remote sensing images: A survey and a new benchmarck. *ISPRS Journal of Photogrammetry and Remote Sensing*, 296-307.

Li L., et al. (2021). Scale-Insensistive Object Detection via Attention Feature Pyramid Transformer Network. *Neural Processing Letters.*

Li N., et al. (2020). Detection of Animal Behind Cages Using Convolutional Neural Network. *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* (págs. 242-245). IEEE.

Li Z., et al. (2017). Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*.

Li Z., et al. (2021). A survey of 3D object detection algorithms for intelligent vehicles development. *Artifical Life and Robotics*.

Liang D., et al. (2020). Lane Detection: A Survey with New Results. *Journal of Computer Science and Technology*, 493-505.

Liang X., et al. (2020). Small Object Detection in Unmanned Aerial Vehicle Images Using Feature Fusion and Scaling-Based Single Shot Detector With Spatial Context Analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 1758-1770.

Lin T., et al. (2017). Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recogntion (CVPR)* (págs. 2117-2125). IEEE.

Lin T., Goyal P., et al. (2017). Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision* (págs. 2980-2988). IEEE.

Liu C., et al. (2022). Double Mask R-CNN for Pedestrian Detection in a Crowd. *Mobile Information Systems*.

Liu L., et al. (2019). Deep Learning for Generic Object Detection: A Survey. *International Jouney of Computer Vision*, 261-3128.

Liu S., Liu D., et al. (2021). Effective template upadte mechanism in visual tracking with background clutter. *Neurocomputing*, 616-625.

Liu S., Liu D., Seivastava G., et al. (2021). Overview of correlation filter based algorithms in object tracking. *Complex and Intelligent Systems*, 1895-1917.

Liu T., et al. (2020). Coupled Network for Robust Pedestrian Detection With Gated Multi-Layer Feature Extraction and Deformable Occlusion Handling. *IEEE Transactions on Image Processing*, 754-766.

Liu Y., Duanmu M., et al. (2021). Exploring multi-scale deformable context and channel-wise attention for salient object detection. *Neurocomputing*, 92-103.

Liu Y., Peng J., et al. (2021). TSingNet: Scale-aware and context-rich feature learning for traffic sign detection and recogntion in the wild. *Neurocomputing*, 10-22.

Liu Y., Sun P., et al. (2021). A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems wirh Applications*, 114602.

Lowe D. (1999). Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision* (págs. 1150-1157). IEEE.

Lu W., et al. (2016). SSD: Single hot MultiboBox Detector. *European Conference on Computer Vision (ECCV)*, (págs. 21-37).

Luo H., et al. (2019). Contextual-YOLOv3: Implement Better Small Object Detection Based Deep Learning. *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, (págs. 134-142).

Luo H., et al. (2021). Object Detection Method Based on Shallow Feature Fusion and Semantic Information Enhancement. 21839-21851.

Meda K., et al. (2021). Artificial intelligence research within reach: an object detection model to identify rickets on pediatric wrist radiographs. *Pediatric Radiology*, 782-791.

Minaee S., et al. (2021). Going Deeper Into Face Detection: A Survey. *arXiv preprint*.

MIttal P., et al. (2020). Deep learning-based object detection in low-altitude UAV datasets: A survey. *Image and Vision Computing*, 204046.

Mo Y., et al. (2019). Highlight-assisted nighttime vehicle detection using a multi-level fusion network and label hierarchy. *Neurocomputing*, 13-23.

Mohammed S., et al. (2021). A Review on Object Detection Algorithms for Ship Detection. *IEEE 7th International Conference on Advanced Computing and Communication Systems (ICACCS)* (págs. 1-5). IEEE.

Mohanty S., et al. (2019). A New Approach for Movign Object Detection under Varying Illumination Enviroments. *2019 International Conference on Information Technology (ICIT)*, (págs. 420-424).

Moncef S. y Othman A. (2021). Efficient Techniques For Human Face Occlusions Detection and Extraction. *2021 International Conference of Women in Daa Science at Taif University (WiDSTaif)*, (págs. 1-5).

Mondal A. (2021). Camouflage design, assessment and breaking techniques: a survey. *Multimedia Systems*.

Moniruzzaman M., et al. (2017). Deep Learning on Underwater Marine Object Detection: A survey. *Advanced Concepts for Intelligent Vision Systems*, 150-160.

Mordan T., et al. (2019). End-to-end Learning of Latent Deformable Part-Based Representations for Object Detection. *International Journal of Computer Vision*, 1659-1679.

Muhammad K., et al. (2018). Early fire detection using convolutional neural networks during surveillance for effective disaster management. *Neurocomputing*, 30-42.

Naji S., et al. (2019). A survey on skin detection in colored images. *Artificial Intelligence Review*, 1041-1087.

Nguyen N., et al. (2019). A Novel Hardware Architecture for Human Detection using HOG-SVM Co-Optimization. *2019 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)* (págs. 33-36). IEEE.

Ning C., et al. (2021). Survey of pedestrian detection with occlusions. *Compley & Intelligent Systems*, 577-587.

Nyein M. y Tint T. (2021). A Review on Advanced Detection Methods in Vehicle Traffic Scenes. *IEEE 6th International Conference on Inventive Computation Technologies (ICICT)* (págs. 642-649). IEEE.

Ojala T., et al. (1994). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. *Proceedings of the 12th IAPR International Conference on Pattern Recognition* (págs. 582-585). IEEE.

Olarewaju M. (2021). Tomato detection based on modified YOLOv3 framework.

Pandiya M., et al. (2020). Analysis of Deep Learning Architectures for Object Detection - A critical Review. *IEEE-HYDCON* (págs. 1-6). IEEE.

Pathak A., et al. (2018). Deep Learning Approaches for Detecting Objects from Images: A Review. *Progress in Computing, Analytics and Networking, Advances in Intelligent Sustems and Computing*.

Pathare S., et al. (2020). Detection of Fractures in Long Bones for Trauma Centre Patients using Hough Transform. *2020 International Conference on Communication and Signal Processing (ICCSP)* (págs. 88-91). IEEE.

R., M. (1986). *Estados Unidos Patente nº 4567610.*

Radhika S., et al. (2018). Determination of Degree of Damage on Building Roofs Due to Wind Disaster from Close Range Remote Sensing Images Using Texture Wavelet Analysis. *IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium* (págs. 3366-3369). IEEE.

Rahmatulloh A., et al. (2021). Face Mask Detection Using Haar Cascade Classifier Algorithm based on Internet of Things with Telegram Bot Notification. *2021 International Conference Advancement in Daa Science, E-learning and Information Systems (ICADEIS)*, (págs. 1-5).

Redmon J. y Farhadr A. (2017). YOLO9000: Better, Faster, Stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pág. 72637271). IEEE.

Redmon J. y Farhadr A. (2018). YOLOv3: An Incremental Improvement. *arXiv preprint arXiv: 1804.02767*.

Redmon J., et al. (2016). You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recogntion (CVPR)* (págs. 779-788). IEEE.

Ren S., et al. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1137-1149.

Rodríguez A., et al. (2020). Reconocimiento de la denominación de billetes a través de una aplicación móvil con reconocimiento de imagen. *ReCIBE. Reviste electrónica de Computación, Informática, Biomédica y Electrónica*, 1-16.

Rublee E., et al. (2011). ORB: An efficient alternative to SIFT or SURF. *2011 International Conference on Computer Vision (ICCV)* (págs. 2564-2571). IEEE.

Saez A., et al. (2019). Statistical Detection of Colors in Dermoscopic Images With a Texton-Based Estimation of Probabilities. *Journal of Biomedical and Health Informatics*, 560-569.

Said N., et al. (2019). Natural disasters detection in social media and satellite imagery: a survey. *Multimedia Tools and Applications*, 31267-21302.

Seemanthini K. y Manjunath S. (2018). Human Detection and Tracking using HOG for Action Recognition. *Procedia Computer Science*, 1317-1326.

Sermanet P., et al. (2013). Overfeat: Integrated recogntion, localization and detection using convolutional networks. *arXiv preprint arXiv: 1312.6229*.

She X., et al. (2021). ScieNet: Deep learning with spike-asisted contextual information extraction. *Pattern Recognition*, 108002.

Shetty A., et al. (2021). A Review: Object Detection Models. *IEEE 6th International Conference for Convergence in Technology (I2CT)* (págs. 1-8). IEEE.

Shi P., et al. (2023). Object Detection Based on Swin Deformable Transformer-BiPAFPN-YOLOX. *Computational Intelligence and Neuroscience*.

Shih H. y Chen J. (2021). Multiskin Color Segmentation Through Morphological Model Refinement. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 225-235.

Shin D., et al. (2020). Dynamic MLML-tree based adaptive object detection using heterogeneous data distribution. *Multimedia Tools and Applications*, 6689-6708.

Shou X., et al. (2019). Automated Visual Inspection of Glass Bottle Bottom With Saliency Detection and Template Matching. *IEEE Transactions on Instrumentation and Measurements*, 4253-4267.

Singh A., et al. (2020). Animal Detection in Man-made Enviroments. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (págs. 1427-1438). IEEE.

Siris A., et al. (2021). Scene Context-Aware Salient Object Detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (págs. 4156-4166). IEEE.

Skoviera R., et al. (2018). Object recognition in clutter color images using Hierarchical Temporal Memory combined with salient-region detection. *Neurocomputing*, 172-183.

Song C., et al. (2020). Automatic Detection and Image Recognition of Precision Agriculture for Citrus Diseases. *2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)* (págs. 187-190). IEEE.

Song X., et al. (2020). Progressive Refinement Network for Occluded Pedestrian Detection. *European Conference on Computer Vision (ECCV)*, (págs. 32-48).

Song X., et al. (2022). PRNet++: Learning towards generalized occluded pedestrian detection via progressive refinement network. *Neurocomputing*, 98-115.

Sun B., et al. (2021). FSCE: Few-Shot Object Detection via Constrastive Proposal Encoding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (págs. 7352-7362). IEEE.

Sun N., et al. (2006). Gender Classification Based on Boosting Local Binary Pattern. *International Symposium on Neural Networks*, (págs. 194-201).

Suruliandi A., et al. (2012). Local binary pattern and its derivatives for face recognition. *IET Computer Vision*, 480-488.

Terven J. y Cordova D. (2023). A Comprehensive Review of YOLO: From YOLOv1 to YOLOv8 and Beyond. *arXiv preprint arXiv:2304.004501*.

Tong K., et al. (2020). Recent advances in small object detection based on deep learning: A review. *Image and Vision Computing*, 103910.

Ullah I., et al. (2020). A brief survey of visual saliency detection. *Multimedia Tools and Applications*, 34605-34645.

Vashisht M. y Kumar B. (2020). A Survey Paper on Object Detection Methods in Image Processing. *International Conference on Computer Science, Engineering and Applications (ICCSEA)* (págs. 1-4). IEEE.

Viola P. y Jones M. (2004). Robust Real-Time Face Detection. *International Jounral of Computer Vision*, 137-154.

Wang C., et al. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.

Wang H., et al. (s.f.). Pose-guided part matching network via shriking and reweighting for occluded person re-identification. *Image and Vision Computing*.

Wang K., et al. (2021). D2C-Net: A Dual-branch, Dual-guidance and Cross-refine Network for Camouflaged Object Detection. *IEEE Transactions on Industrial Electronics*.

Wang K., Peng X., Yang J., Meng D. y Qiao Y. (2020). Region Attention Networks for Pose and Occlusion Robust Facia Expression Recognition. *IEEE Transactions on Image Processing, 29*, 4057-4069.

Wang L., et al. (2021). Giant Panda Identification. *IEEE Transactions on Image Processing*, 2837-2849.

Wang Q., et al. (2020). Overview of deep-learning based methods for salient object detection in videos. *Pattern Recognition*, 107340.

Wang S., et al. (2020). Visual Context Aware Ship Detector for High-Resolution SAR Imagery. *2020 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (págs. 1778-1781). IEEE.

Wang W., et al. (2021). Salient Object Detection in the Deep Learning Era: An In-depth Survey. *Transactions on Pattern Analysis and Machine Intelligence*.

Wang Y., et al. (2020). A Camouflaged Object Detection Model Based on Deep Learning. *2020 IEEE International Conference on Artificial Intelligence and information Systems (ICAIIS)* (págs. 150-153). IEEE.

Wang Y., et al. (2022). Detecting Occluded and Dense Trees in Urban Terrestrial Views with High-quality Tree Detecton Dataset. *IEEE Transactions on Geoscience and Remote Sensing*.

Wank K., et al. (2020). Region Attention Networks for Pose and Occlusion Robust Facial Expression Recogntion. *IEEE Transactions on Image Processing*, 4057-4069.

Wei W., et al. (2019). Occluded Pedestrian Detection Based on Depth Vision Significance in Biomimetic Binocular. *IEEE Sensors Journal*, 11469-11474.

Wei X., et al. (2020). Pedestrian detection in underground mines via parallel feature transfer network. *Pattern Recognition*, 107195.

Wu C. y Ding J. (2018). Occluded face recognition using low-rank regression with generalized gradient direction. *Pattern Recognition*, 256-268.

Wu D., et al. (2020). Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural enviroemnt. *Computers and Electronics in Agriculture*, 105742.

Wu T. (2020). A Supernova Detection Implementation based on Faster R-CNN. *2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)* (págs. 390-393). IEEE.

Wu Y., et al. (2020). Real-time traffic sign detection and classification towards real traffic scene. *Multimedia Tools and Applications*, 18201-18219.

Xiang L., et al. (2023). DINF: Dynamic Instance Noise Filter for Occluded Pedestrian Detection. *arXiv preprint arXiv:2301.05565*.

Xiao B., et al. (2021). A vision-based method for automatic tracking of construction machines at nighttime based on deep learning illumination enhancement. *Automation in Construction*, 103721.

Xiao Y., et al. (2020). A review of object detection based on deep learning. *Multimedia Tools and Applications*, 23729-23791.

Xie H., et al. (2019). Context-aware pedestrian detection especially for small-sized instances with Deconvolution Integrated Faster RCNN (DIF R-CNN). *Applied Intelligence*, 1200-1211.

Xie J., et al. (2020). Mask-Guided Attention Network and Occlusion-Sensitive Hard Example Mining for Occluded Pedestrian Detection. *IEEE Transactions on Image Processing*, 3872-3884.

Xie S., et al. (2020). Diverse receptive field network with contect aggregation for fast object detection. *Journal of Visual Communication and Image Representation*, 102770.

Xu S., Gu J., et al. (2023). DKTNet: Dual-Key Transformer Network for small object detection. *Neurocomputing*, 29-41.

Xu S., Zhang M., et al. (2023). A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing*, 204-232.

Xu X., et al. (2021). Boundary guidance network for camouflage object detection. *Image and Vision Computing*, 104283.

Xu Y., et al. (2018). Rapid Pedestrian Detection Based on Deep Omega-Shape Features with Partial Occlusion Handing. *Neural Processing Letters*, 923-937.

Yan Z., et al. (2021). Detection-Oriented Backbone Trained from Near Scratch and Local Feature Refinement for Small Object Detection. *Neural Processing Letters*, 1921-1943.

Yilmaz B. y Karsligil M. (2020). Detection Of Airplane And Airplane Parts From Security Camera Images with Deep Learning. *2020 28th Signal Processing and Communications Applications Conference (SIU)* (págs. 1-4). IEEE.

Yuan J., et al. (2020). Gated CNN: Integrating multi-scale feature layers for object detection. *Pattern Recognition*, 107131.

Yuan Y., Xiong Z. y Wang W. (2019). VSSA-NET: Vertical Spatial Sequence Attention Network for Traffic Sign Detection. *IEEE Transactions on Image Processing*, 3423-2434.

Yudin D., et al. (2019). Detection of Big Animals on Images with Road Scenes using Deep Learning. *2019 International Conference on Artificial Intelligence Applications and Innovations* (págs. 100-103). IEEE.

Zardoua Y., et al. (2021). A Horizon Detection Algorithm for Maritime Surveillance. *arXiv preprint*.

Zhang B., et al. (2010). Local Derivative Pattern Versus Local Binary Pattern: Face Recognition With High-Order Local Pattern Descriptor. *IEEE Transactions on Image Processing*, 533-544.

Zhang C., et al. (2022). Camouflaged object detection via Neighbor Connection and Hieratchical Information Transfer. *Computer Vision and image Understanding*, 103450.

Zhang G., et al. (2004). Boosting Local Binary Pattern (LBP)-Based Face Recognition. *Chinese Conference on Biometric Recogntion*, (págs. 179-186).

Zhang G., et al. (2019). CAD-Net: A Context-Aware Detection Network for objects in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 10015-10024.

Zhang H. y Hong X. (2019). Recent progress on object detection: a brief review. *Multimedia Tools and Applications*, 27809-27847.

Zhang H., et al. (2021). An empirical study of multi-scale object detection in high resolution UAV images. *Neurocomputing*, 173-182.

Zhang K., et al. (2021). Semantic Context-Aware Network for Multiscale Object Detection in Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, 1-5.

Zhang L., et al. (2020). A Multistage Refinement Network for Salient Object Detection. *IEEE Transactions on image Processing*, 3534-3545.

Zhang P., et al. (2020). Non-rigid object tracking via deep multi-scale spatial-temporal saliency maps. *Pattern Recognition*, 107130.

Zhang Q., et al. (2020). Dense Attention Fluid Network for Salient Object Detection in Optical Remote Sensing Images. *IEEE Transactions on Image Processing*, 1305-1317.

Zhang S., et al. (2020). Asymmetric multi-stage CNNs for small-scale pedestrian detection. *Neurocomputing*, 12-26.

Zhang T., et al. (2020). Omni-supervised joint detection and pose estimation for wild animals. *Pattern Recognition Letters*, 84-90.

Zhang W., et al. (2021). Global context aware RCNN for object detection. *Neural Computing and Applications*, 11627-11639.

Zhang X., et al. (2022). PSNEt: Perspective-sensitive convolutional network for object detection. *Neurocomputing*, 384-395.

Zhang X., Liu Y., Huo C., Xu N., Wang L. y Pan C. (2022). PSNEt: Perspective-sensitive convolutional network for object detection. *Neurocomputing*, 384-395.

Zhang Y., et al. (2019). Detecting small faces in the wild based on generative adversarial network and contextual information. *Pattern Recognition*, 74-86.

Zhao M., et al. (2021). An occlusion-resistan circle detector using inscribed triangles. *Pattern Recognition*, 107588.

Zheng H., et al. (2020). Feature enhacnement for multi-scale object detection. *Neural Processing Letters*, 1907-1919.

Zheng Q. y Chen Y. (2021). Interactive multi-scale feature representation enhancement for small object detection. *Image and Vision Computing*, 104128.

Zheng W., et al. (2020). A novel approach inspired by optic nerve characteristics for few-shot occluded face recogntion. *Neurocomputing*, 25-41.

Zheng Y., et al. (2019). Detection of People With Camouflage Pattern Via Dense Deconvolution Network. *IEEE Signal Processing Letters*, 29-33.

Zhou C. y Yuan J. (2019). Multi Pattern Recognition-label learning of part detectors for occluded pedestrian detection. 99-111.

Zhou C. y Yuan J. (2020). Occlusion Pattern Discovery for Object Detection and Occlusion Reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2067-2080.

Zhou K., et al. (2020). Improving Multispectral Pedestrian Detection by Addresing Modality Imbalance Problems. *European Conference on Computer Vision (ECCV)*, (págs. 787-803).

Zhou S., et al. (2019). Discriminative Feature Learning With Foreground Attention for Person Re-Identification. *IEEE Transactions on Image Processing*, 4671-4684.

Zhou S., et al. (2020). Depth occlusion perception feature analysis for person re-identification. *Pattern Recognition Letters*, 617-623.

Zhou T., et al. (2021). RGB-D salient object detection: A survey. *Computational Visual Media*, 37-69.

Zhou X. y Zhang L. (2022). SA-FPN: An effective feature pyramid network for crowded human detection. *Applied Intelligence*.

Zhu Y., et al. (2019). Attention CoupleNet: Fully Convolutional Attention Coupling Network for Object Detection. *IEEE Transactions on Image Processing*, 113-126.

Zhuge M., et al. (2022). CubeNet: X-shape connection for camouflaged object detection. *Pattern Recognition*, 108644.

Zou Z., et al. (2023). Object Detection in 20 Years: A Survey. *Proceedings of the IEEE*, 257-276.

Zuo J., et al. (2018). Person Re-Identification. *2018 IEEE International Conference on Multimedia and Expo (ICME)* (págs. 1-6). IEEE.