

Revisión

Detección de armas de fuego en videovigilancia mediante aprendizaje profundo: revisión de literatura

Firearm detection in video surveillance using deep learning: a literature review

Cristhian Briones-Montalvo ^{1*}, José Steven Cordero-Bazurto ², Orlando Erazo ³, Valeria Torres-Lindao ⁴ y José Luis Toapanta ⁵

¹ Universidad Técnica Estatal de Quevedo, Facultad de Ciencias de la Computación, Quevedo, Ecuador; <https://orcid.org/0009-0003-2642-2527>

² Universidad Técnica Estatal de Quevedo, Facultad de Ciencias de la Computación, Quevedo, Ecuador; <https://orcid.org/0009-0001-2961-6736>, jcorderob@uteq.edu.ec

³ Universidad Técnica Estatal de Quevedo, Facultad de Ciencias de la Computación, Quevedo, Ecuador; <https://orcid.org/0000-0001-5642-9920>, orazo@uteq.edu.ec

⁴ Universidad Técnica Estatal de Quevedo, Unidad de Admisión y Nivelación, Quevedo, Ecuador; <https://orcid.org/0000-0001-8778-2794>, vtorresl@uteq.edu.ec

⁵ Universidad Técnica Estatal de Quevedo, Facultad de Ciencias de la Computación, Quevedo, Ecuador; <https://orcid.org/0009-0005-7120-8276>, jtoapantag2@uteq.edu.ec

* Correspondencia: cbrionesm2@uteq.edu.ec

Cita: Briones-Montalvo, C., Cordero-Bazurto, J. S., Erazo, O., Torres-Lindao, V., & Toapanta, J. L. (2026). Detección de armas de fuego en videovigilancia mediante aprendizaje profundo: revisión de literatura. *Multidisciplinary Collaborative Journal*, 4(4), 1-25. <https://doi.org/10.70881/mcj/v4/n4/181>

Recibido: 15/08/2026
Revisado: 25/09/2026
Aceptado: 06/10/2026
Publicado: 12/10/2026



Copyright: © 2026 por los autores. Este artículo es un artículo de acceso abierto distribuido bajo los términos y condiciones de la Licencia Creative Commons, Atribución-NoComercial 4.0 Internacional. (CC BY-NC).

(<https://creativecommons.org/licenses/by-nc/4.0/>)

 <https://doi.org/10.70881/mcj/v4/n4/181>

Resumen: Los sistemas de videovigilancia dependen de observadores humanos con capacidad de atención limitada, lo que impide convertir la cobertura visual en una alerta oportuna ante un arma de fuego. El objetivo de este trabajo fue revisar y organizar la literatura sobre la detección de armas de fuego en videovigilancia mediante aprendizaje profundo, con énfasis en los conjuntos de datos, los modelos, la evaluación y el despliegue. Se recopilaron artículos publicados entre 2016 y 2026 con resultados experimentales sobre escenarios de videovigilancia y conjuntos de datos documentados. Los hallazgos muestran que el problema está condicionado ante todo por los datos: el tamaño reducido del arma dentro del cuadro constituye el principal cuello de botella, junto con la oclusión, la iluminación y los objetos visualmente próximos al arma. Además, los detectores de dos etapas, los de una etapa y los esquemas en cascada ofrecen compromisos distintos entre falsos positivos y velocidad, aunque las cifras publicadas no autorizan una ordenación entre arquitecturas, pues proceden de conjuntos propios y métricas heterogéneas. Se concluye así que estos sistemas identifican la apariencia de un arma y no determinan una amenaza, por lo que el protocolo de verificación humana posterior a la alerta resulta tan relevante como el propio detector.

Palabras clave: visión por computadora, redes neuronales convolucionales, detección de objetos, computación en el borde, seguridad

Abstract: Video surveillance systems rely on human observers with limited attention capacity, which prevents visual coverage from becoming a timely alert when a firearm is involved. The aim of this work was to review and organize the literature on firearm detection in video surveillance using deep learning, focusing on datasets, models, evaluation and deployment. Articles published between 2016 and 2026 reporting experimental results on surveillance scenarios with documented datasets were collected. The findings show that the problem is constrained above all by data: the small size a weapon occupies within a frame is the main bottleneck, together with occlusion, lighting and objects visually similar to a weapon. Two-stage detectors, one-stage detectors and cascaded schemes offer different trade-offs

between false positives and speed, although published figures do not warrant a ranking among architectures, since they come from custom datasets and heterogeneous metrics. It is concluded that these systems identify the appearance of a weapon rather than determine a threat, so the human verification protocol following the alert is as relevant as the detector itself.

Keywords: computer vision, convolutional neural networks, object detection, edge computing, security

1. Introducción

Los sistemas de videovigilancia basados en circuito cerrado de televisión se han convertido en un componente habitual de la seguridad en espacios públicos y privados, con despliegues que abarcan desde comercios y entidades financieras hasta instituciones educativas y vías urbanas (Olmos et al., 2018; Santos et al., 2024). Su operación descansa en un esquema en el que una o varias personas observan de forma simultánea múltiples transmisiones en busca de eventos anómalos (Grega et al., 2016; Jaleel et al., 2022; Zahrawi & Shaalan, 2023). Sin embargo, este esquema suele utilizarse para la revisión posterior de incidentes y no para la reacción inmediata (Fernandez-Testa & Salcedo, 2024; Yadav et al., 2023). De hecho, la capacidad de atención de un operador humano es limitada y decae con la duración de la jornada, de modo que el número de cámaras que puede vigilar de manera efectiva resulta reducido (Olmos et al., 2018; Pavithra et al., 2022). En consecuencia, el crecimiento en el número de cámaras instaladas no se traduce de forma automática en un aumento equivalente de la seguridad efectiva.

Esta limitación resulta particularmente crítica cuando la amenaza involucra un arma de fuego (Berardini et al., 2024; Santos et al., 2024). En este tipo de incidentes, el intervalo entre la aparición visible del arma y la consumación del hecho puede ser muy breve, por lo que el valor de la detección depende de que ocurra dentro de ese intervalo y no durante la revisión posterior del material grabado (Fernandez-Testa & Salcedo, 2024; Olmos et al., 2018). A ello se suma que el arma permanece oculta durante buena parte de la escena y ocupa una porción muy pequeña del cuadro, lo que dificulta su identificación incluso para un observador atento (Berardini et al., 2024; Lim et al., 2019; Santos et al., 2024). Entonces, el problema no radica en la ausencia de cobertura visual, sino en la imposibilidad de convertir esa cobertura en una alerta oportuna. Por lo tanto, se requiere de algún mecanismo capaz de analizar de forma continua el flujo de video y de avisar al operador cuando un arma aparezca en la escena.

Frente a esta necesidad, la visión por computadora basada en aprendizaje profundo ha sido la vía dominante de respuesta durante la última década. A diferencia de los enfoques clásicos, que dependían de descriptores diseñados manualmente y se degradaban ante cambios de ángulo, escala o iluminación, las redes neuronales convolucionales aprenden representaciones directamente de los datos (Salido et al., 2021; Yadav et al., 2023). Sobre esa base se han consolidado dos familias de

detectores, la de dos etapas, representada por Faster R-CNN, y la de una etapa, representada por YOLO y RetinaNet, a las que se suman esquemas híbridos organizados en cascada (Olmos et al., 2018; Santos et al., 2024; Zahrawi & Shaalan, 2023). Los trabajos publicados reportan valores de desempeño elevados, con valores de mAP superiores al 90 % y velocidades de decenas de fotogramas por segundo (Ahmed et al., 2022; Bhatti et al., 2021), además de despliegues sobre *hardware* restringido, orientados a reducir la latencia y el tráfico de red (Berardini et al., 2024; Fathy & Saleh, 2022; Qi et al., 2021). En conjunto, estos resultados sugieren viabilidad técnica en condiciones controladas, aunque no equivalen por sí solos a una validación operativa generalizable.

A pesar de tal disponibilidad, el repertorio no se traduce en un panorama claro para quien debe seleccionar o desplegar una solución. Cada grupo de investigación construye su propio conjunto de datos, con lo cual las cifras publicadas proceden de escenas de dificultad distinta y de métricas que no siempre miden lo mismo, de modo que no facilitan una comparación directa entre arquitecturas (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024). A esta fragmentación se añaden limitaciones que la literatura reporta de forma recurrente, pero dispersa, entre las que están la detección de objetos de tamaño reducido, la oclusión, las falsas alarmas producidas por objetos cotidianos y el sesgo hacia armas cortas visibles (Berardini et al., 2024; Lim et al., 2019; Rao et al., 2026). Aunque existen revisiones previas del campo, Yadav et al. (2023) contrastan los métodos clásicos con los basados en aprendizaje profundo, mientras que Santos et al. (2024) sistematizan los modelos aplicados a la detección de armas en imágenes y videos de vigilancia.

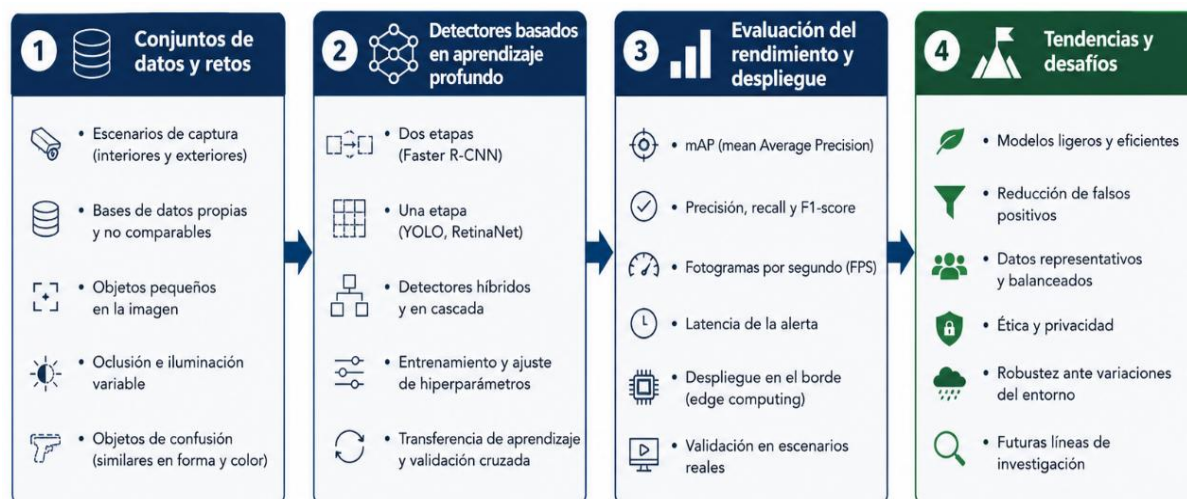
En este escenario, la presente revisión bibliográfica adopta un enfoque complementario al organizar la evidencia siguiendo la cadena que se extiende desde la construcción de los conjuntos de datos hasta el despliegue operativo. La falta de una síntesis articulada desde esta perspectiva dificulta determinar cuáles de las limitaciones observadas se deben a la arquitectura del detector y cuáles están relacionadas con los datos utilizados para su entrenamiento y evaluación. Así, el objetivo de este trabajo fue revisar la literatura sobre la detección de armas de fuego en sistemas de videovigilancia mediante aprendizaje profundo, con énfasis en los conjuntos de datos empleados, las familias de modelos aplicadas, los criterios de evaluación reportados y las condiciones de despliegue documentadas.

La contribución principal radica en organizar evidencia actualmente dispersa dentro de un marco común que permita separar los avances consolidados de las limitaciones persistentes, así como identificar las brechas que orientan el trabajo futuro en este dominio. La organización de la revisión responde a la secuencia con que se construye un sistema de detección, de manera que cada sección aporta las condiciones que la siguiente presupone. Se inicia con los conjuntos de datos y los retos que estos imponen, dado que el material disponible fija el techo de desempeño alcanzable. Luego se abordan las familias de modelos de aprendizaje profundo aplicadas al problema y, después, los criterios de evaluación y los contextos de implementación

en los que dichos modelos han sido desplegados. Por último, se examinan las tendencias y los desafíos abiertos, para dar paso a la discusión y a las conclusiones del estudio. La Figura 1 ilustra esta secuencia junto con las ideas principales identificadas en cada etapa.

Figura 1

Organización de la revisión y aspectos relevantes de cada etapa



2. Conjuntos de datos y desafíos de la detección de armas de fuego

La detección automática de armas de fuego en videovigilancia es, ante todo, un problema condicionado por los datos. Los modelos de aprendizaje profundo aprenden a partir de ejemplos etiquetados, de modo que la naturaleza de las cámaras, los escenarios de captura y la calidad del etiquetado determinan en buena medida el techo de desempeño alcanzable. Un conjunto de datos, en este dominio, es una colección de imágenes o fotogramas de video en los que la posición del arma se ha etiquetado manualmente, habitualmente mediante un rectángulo delimitador o *bounding box*. Estas etiquetas permiten entrenar el detector y, después, medir su acierto.

La literatura reciente coincide en que el aprendizaje profundo ya permite detectar armas de fuego en tiempo casi real, pero también en que el rendimiento cae de forma apreciable en escenas difíciles (Santos et al., 2024). Por ello, esta sección organiza el estado del campo desde la perspectiva de los datos: sobre qué material trabajan los modelos y qué dificultades introduce ese material.

2.1. Sistemas de videovigilancia y escenarios de captura

Un sistema de videovigilancia, comúnmente denominado CCTV, consiste en un conjunto de cámaras que transmiten video hacia un centro de monitoreo donde uno o varios operadores humanos observan las escenas en busca de eventos anómalos. Este esquema, concebido originalmente para la revisión posterior de incidentes, resulta limitado cuando se espera una reacción inmediata. La detección automática

de armas surge precisamente para reducir la dependencia de esos operadores, quienes se fatigan y no pueden vigilar de manera continua múltiples cámaras a la vez.

Los escenarios de captura reportados en la literatura son heterogéneos y condicionan el diseño de los modelos. Se encuentran cámaras fijas de interiores, como comercios y cajeros automáticos, donde la distancia al sujeto es corta y la iluminación es relativamente estable, y también cámaras de exteriores y de vía pública, con profundidad variable y condiciones lumínicas cambiantes. Esta variedad implica que un mismo detector debe enfrentar armas que ocupan desde una fracción considerable del cuadro hasta apenas unas decenas de píxeles (Berardini et al., 2024; Lim et al., 2019; Santos et al., 2024).

Un aspecto adicional del contexto de captura es el propósito operativo del sistema. Varios trabajos plantean sus propuestas como alarmas tempranas, orientadas a avisar de una amenaza probable, más que como sistemas de reconocimiento forense fino (Olmos et al., 2018; Rao et al., 2026; Salido et al., 2021). Esta distinción es relevante porque cambia el criterio de éxito: no se busca clasificar el modelo exacto del arma, sino disparar una alerta útil en un plazo compatible con la respuesta de seguridad.

2.2. Conjuntos de datos públicos y privados

La disponibilidad de conjuntos de datos estandarizados y suficientemente realistas para vigilancia es uno de los problemas centrales identificados en la literatura (Bhatti et al., 2021; Lim et al., 2019; Qi et al., 2021; Salazar González et al., 2023). La práctica dominante en cuanto a datos consiste en que cada grupo de investigación construye su propia base de datos. Los trabajos revisados reportan la recolección de imágenes desde internet, la extracción de fotogramas de videos de YouTube, el uso de grabaciones reales de cámaras CCTV y la combinación de datos reales con datos sintéticos (Bhatti et al., 2021; Qi et al., 2021; Santos et al., 2024). Estas colecciones propias resuelven el problema inmediato del entrenamiento, pero introducen uno nuevo: cada modelo se evalúa sobre un conjunto distinto.

Una característica poco discutida de estas colecciones propias es la procedencia del material fotografiado. Junto a las imágenes recolectadas de internet y a los fotogramas extraídos de video, una parte apreciable del material se obtiene mediante escenificaciones controladas, en las que un actor porta el arma ante una cámara dispuesta para el efecto y en las que el objeto empleado es con frecuencia una réplica y no un arma real (Grega et al., 2016; Lim et al., 2019). A ello se suma la generación de imágenes sintéticas mediante motores de videojuego, así como el registro de simulacros institucionales de ataque (Salazar González et al., 2020).

La motivación de estas prácticas es comprensible, dado que resulta inviable obtener por vías legítimas un volumen suficiente de grabaciones reales de incidentes armados. En cambio, su consecuencia rara vez se explicita, ya que el detector aprende y se evalúa sobre una representación del arma cuya equivalencia con el caso real se asume sin verificación. El caso mejor documentado es el de Salazar-González y sus colegas (Salazar González et al., 2020), quienes combinaron grabaciones

procedentes de una instalación CCTV real con imágenes sintéticas generadas para el efecto y analizaron de forma explícita el impacto de ese material sobre el entrenamiento del detector.

La consecuencia metodológica de esta fragmentación es que los resultados publicados no son directamente comparables entre sí. Dos detectores pueden reportar valores de precisión muy próximos y no haber sido medidos sobre escenas de dificultad muy distinta. La propia literatura señala que la generalización sigue limitada, en buena parte, por esta heterogeneidad de los conjuntos de datos (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024).

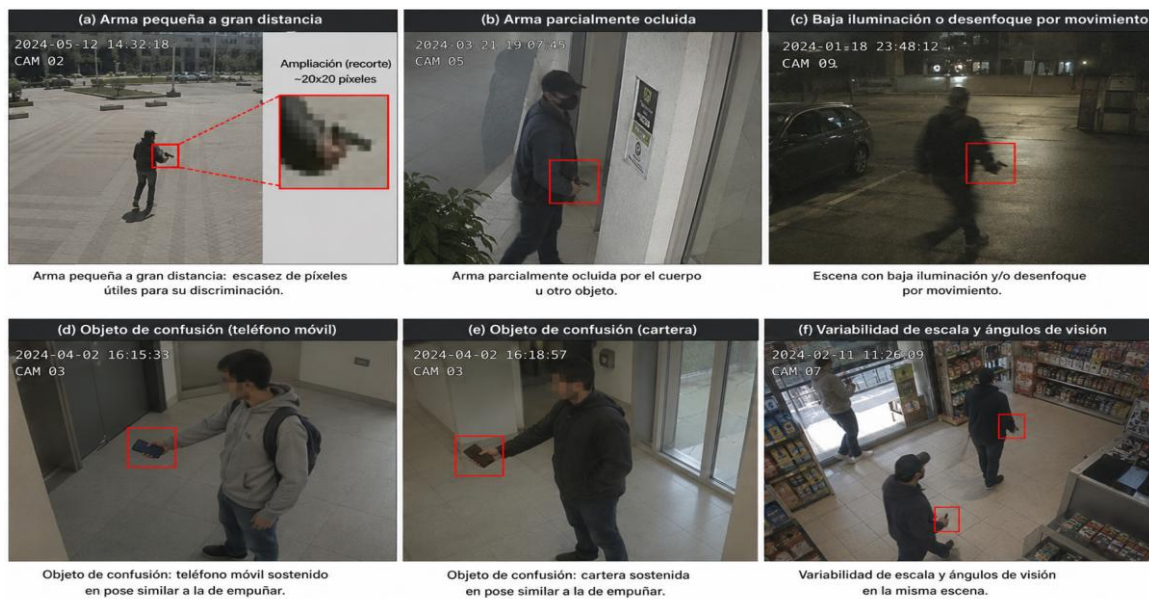
Debe notarse también un sesgo de alcance en el material disponible. La mayoría de los trabajos se concentra en pistolas o armas cortas visibles, y no en armas ocultas o parcialmente encubiertas (Olmos et al., 2018; Salido et al., 2021). Los conjuntos de datos reflejan esa elección, de modo que el desempeño reportado describe un subconjunto del problema real de seguridad y no su totalidad.

2.3. Retos asociados a los datos

La Figura 2 ilustra de forma conjunta los retos referentes a los conjuntos de datos de armas de fuego a partir de fotogramas representativos. El reto más repetido en la literatura es el tamaño reducido que ocupa el arma dentro del campo de visión. Una pistola observada por una cámara de vigilancia puede representar una región muy pequeña de la imagen, especialmente en CCTV de baja resolución o con profundidad variable, lo que deja al detector con muy pocos píxeles informativos (Berardini et al., 2024; Lim et al., 2019; Santos et al., 2024). Este factor se identifica de forma consistente como el principal cuello de botella del campo (Bhatti et al., 2021; Santos et al., 2024). La evidencia más clara de esta limitación procede de González-Barbosa et al. (2020), quienes construyeron un conjunto a partir de una instalación CCTV real y de imágenes sintéticas. Sus resultados muestran que, pese a los avances de los detectores, la detección de armas en CCTV continúa siendo un problema abierto, atribuible en buena medida a la pérdida de detalle propia de los objetos de tamaño reducido.

Figura 2

Manifestación visual de los retos asociados a los datos



Nota: imágenes sintéticas generadas mediante ChatGPT (modelo DALL·E 3, OpenAI); no corresponden a escenas reales ni constituyen datos experimentales.

Al problema en cuestión se suman degradaciones propias del entorno de captura. La oclusión, los cambios de ángulo, la iluminación pobre, el desenfoque, la compresión y los fondos complejos deterioran el rendimiento de los detectores en escenas reales (Abins et al., 2024; Bhatti et al., 2021; Corral-Sanz et al., 2024; Olmos et al., 2018). Frente a ello, algunos trabajos incorporan etapas de preprocesamiento orientadas a tratar imágenes borrosas o de baja luminosidad antes de la inferencia (Goenka & Sitara, 2022).

Un tercer reto proviene de los objetos que se parecen a un arma dentro de la escena. Teléfonos móviles, carteras y objetos similares sostenidos con la mano generan detecciones erróneas que se traducen en falsas alarmas. Por esta razón, varias propuestas incluyen deliberadamente objetos de confusión en el conjunto de entrenamiento, de modo que el modelo aprenda a distinguirlos (Bhatti et al., 2021; Goenka & Sitara, 2022; Salido et al., 2021).

Conviene distinguir este último caso del que plantean las armas de imitación. Un teléfono móvil o una cartera pueden confundirse accidentalmente con un arma por coincidencias parciales de forma y contexto de sujeción, de modo que su inclusión como clase negativa constituye una estrategia razonable para reducir falsos positivos (Bhatti et al., 2021). En cambio, una réplica realista ha sido fabricada para reproducir la apariencia externa de un arma. Por ello, no puede garantizarse su diferenciación respecto de un arma real utilizando exclusivamente información visual convencional. La Tabla 1 sistematiza estos retos junto con su implicación para la formulación de modelos.

Tabla 1
Retos asociados a los datos en la detección de armas de fuego en videovigilancia

Reto	Descripción	Implicación para el modelo	Citas
Objetos pequeños	El arma ocupa muy pocos píxeles en CCTV de baja resolución o a distancia variable.	Principal cuello de botella; pérdida de rasgos discriminantes.	(Berardini et al., 2024; Lim et al., 2019; Santos et al., 2024)
Oclusión y ángulo	El arma aparece parcialmente cubierta o en poses no canónicas.	Reduce el <i>recall</i> ; exige diversidad de vistas en el entrenamiento.	(Abins et al., 2024; Bhatti et al., 2021; Olmos et al., 2018)
Iluminación y desenfoque	Escenas con baja luminosidad, movimiento o fondos complejos.	Motiva etapas de preprocesamiento previas a la inferencia.	(Abins et al., 2024; Goenka & Sitara, 2022)
Objetos de confusión	Móviles, carteras y objetos sostenidos con la mano se asemejan a un arma.	Fuente principal de falsos positivos; se mitiga incluyéndolos como clase.	(Bhatti et al., 2021; Goenka & Sitara, 2022; Salido et al., 2021)
Armas de imitación y material simulado	Réplicas, armas de aire comprimido y armas de utilería resultan visualmente indistinguibles de un arma real; parte del material de entrenamiento procede de escenificaciones o de imágenes sintéticas.	Limita la validez ecológica y la capacidad de inferir si el objeto detectado corresponde a un arma real.	(Grega et al., 2016; Lim et al., 2019; Salazar González et al., 2020)
Escasez de conjuntos de datos	Ausencia de conjuntos estandarizados y realistas para vigilancia.	Obliga a construir bases propias y limita la comparabilidad.	(Bhatti et al., 2021; Lim et al., 2019; Qi et al., 2021)
Diversidad de escenarios	Interiores, exteriores, distintas cámaras y distancias.	Compromete la generalización fuera del conjunto de datos de origen.	(Fernandez-Testa & Salcedo, 2024; Santos et al., 2024)

3. Modelos de aprendizaje profundo para la detección de armas

El núcleo técnico del campo está constituido por modelos de detección de objetos basados en redes neuronales profundas. A diferencia de la clasificación de imágenes, que únicamente identifica qué elementos están presentes en una escena, la detección también debe localizarlos y delimitar su posición. En el caso de las armas de fuego, los estudios comparativos y experimentales muestran que el aprendizaje profundo ofrece ventajas frente a los enfoques clásicos en términos de velocidad y rendimiento (Akshaya et al., 2023; Salido et al., 2021; Yadav et al., 2023). Esta sección examina la evolución de las principales familias de modelos, en lugar de describir de manera aislada cada trabajo.

3.1. De los enfoques clásicos al aprendizaje profundo

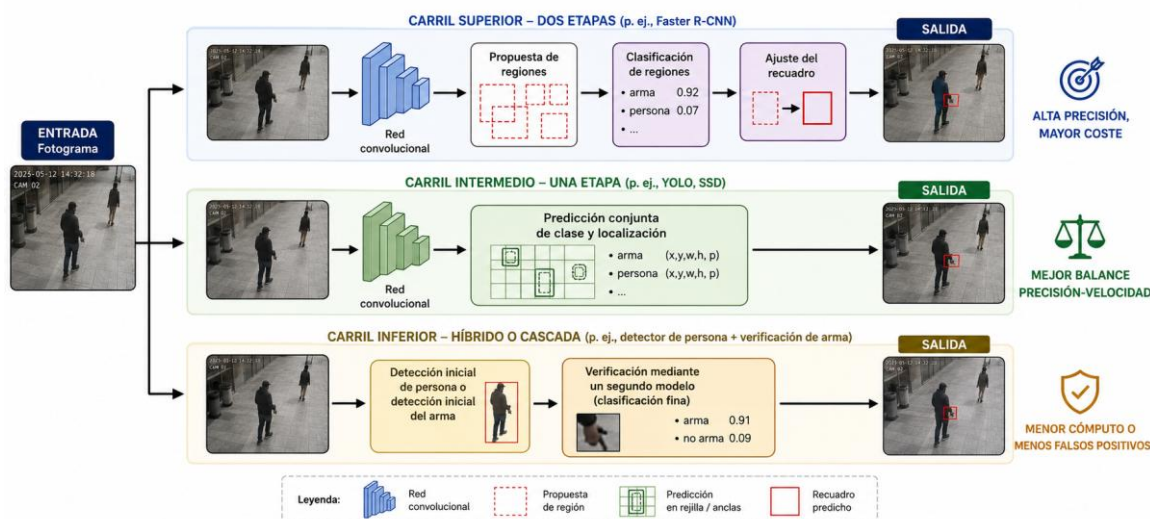
Los enfoques clásicos de visión por computadora abordaban la detección mediante descriptores diseñados a mano; es decir, reglas explícitas sobre bordes, texturas o formas que un clasificador utilizaba para decidir. Su principal limitación es que esos descriptores se degradan cuando cambian el ángulo, la escala o la iluminación,

condiciones que en videovigilancia son la norma. Los estudios comparativos disponibles muestran que los métodos de aprendizaje profundo superan a estos enfoques en el problema de la detección de armas (Salido et al., 2021; Yadav et al., 2023).

El aprendizaje profundo sustituye los descriptores manuales por representaciones aprendidas directamente de los datos. Las redes neuronales convolucionales extraen jerarquías de rasgos que van desde bordes simples hasta configuraciones complejas, y esa jerarquía es la que alimenta al detector. Sobre esa base común se han construido las dos grandes familias que dominan el campo, junto con esquemas híbridos que las combinan (Santos et al., 2024; Zahrawi & Shaalan, 2023). La Figura 3 contrasta el flujo de procesamiento de cada una de ellas y se ofrecen más detalles en los apartados posteriores.

Figura 3

Esquema comparativo de las familias de detectores aplicadas a la detección de armas



3.2. Detectores de dos etapas

Los detectores de dos etapas separan el problema en dos pasos sucesivos. En el primero, el modelo propone regiones de la imagen que podrían contener un objeto; en el segundo, clasifica cada una de esas regiones y ajusta su delimitación. Faster R-CNN es el representante más conocido de esta familia y una de las arquitecturas más utilizadas en la detección de armas, según las revisiones y estudios comparativos del campo (Olmos et al., 2018; Santos et al., 2024; Zahrawi & Shaalan, 2023).

La ventaja de este esquema es su capacidad para minimizar los falsos positivos, ya que la segunda etapa actúa como un filtro de verificación sobre las propuestas iniciales. Olmos et al. (2018) reportan que, sobre un conjunto de escenas de prueba, su sistema basado en Faster R-CNN activó la alarma en 27 de 30 casos con una latencia inferior a 0,2 segundos. Estos autores formulan el problema como una alarma

temprana, de modo que el criterio de éxito no es la exactitud agregada sino la fiabilidad del aviso, y ese comportamiento resulta especialmente valioso cuando el coste operativo de una falsa alarma es alto.

En una línea próxima, Arif et al. (2022) compararon R-CNN y R-FCN con los extractores de rasgos VGG y ResNet sobre un conjunto construido por los propios autores con cámaras de distinto tipo y resolución, con el propósito declarado de reducir el carácter idealizado de las imágenes de entrenamiento. Su trabajo muestra que, dentro de esta familia, la discusión se ha desplazado del detector hacia la representatividad del material sobre el que se lo mide. Su desventaja es el coste computacional asociado a ejecutar dos etapas por cada fotograma. En despliegues con múltiples cámaras en vivo, ese coste se multiplica y compite directamente con el requisito de operar en tiempo real. Esta tensión explica que la investigación posterior se haya desplazado en buena medida hacia arquitecturas de una sola etapa.

3.3. Detectores de una etapa

Los detectores de una etapa eliminan la fase de propuesta de regiones y predicen simultáneamente la clase y la posición del objeto en una única pasada por la red. La familia YOLO es el exponente principal de este enfoque y, junto con Faster R-CNN, concentra la mayor parte de los trabajos publicados en detección de armas (Olmos et al., 2018; Santos et al., 2024; Zahrawi & Shaalan, 2023). La evolución de esta familia puede seguirse a través de trabajos que han desplazado progresivamente el objetivo de optimización. Bhatti et al. (2021) entrenaron YOLOv4 sobre una base propia de CCTV y alcanzaron un mAP de 91,73 % y un F1 de 91 %, priorizando la operación en tiempo real. Ahmed et al. (2022) evaluaron además la viabilidad de despliegue: su variante Scaled-YOLOv4 obtuvo un mAP de 92,1 % a 85,7 FPS sobre una GPU RTX 2080 Ti y fue optimizada para Jetson Nano mediante TensorRT. Estudios posteriores también han explorado técnicas profundas para CCTV y configuraciones recientes de detección en tiempo real (Bamnolkar et al., 2025; Yeddula & Reddy, 2022).

Los trabajos posteriores confirman la competitividad de la familia sin resolver su comparabilidad. Yadav et al. (2024) reportaron con YOLOv7 un 96,80 % de exactitud y un 90,37 % de *recall*, mientras que Rao et al. (2026) optaron por una comparación interna entre versiones y midieron YOLOv8m frente a YOLOv5m sobre un mismo conjunto, con un mAP50 de 88,54 % frente a 81,69 %. Esta última estrategia es metodológicamente más informativa que las cifras aisladas, porque aísla el efecto de la versión del detector al mantener fijos los datos.

RetinaNet constituye otra alternativa de una etapa presente en la literatura del dominio (Salido et al., 2021). RetinaNet muestra que el desempeño no depende solo de la arquitectura. Salido et al. (2021) reportaron un AP de 96,36 % y un *recall* de 97,23 % con RetinaNet. En una línea complementaria, Lamas et al. (2022) incorporaron estimación de pose humana para mitigar falsos negativos y mostraron que la información contextual sobre el portador puede complementar al detector, especialmente cuando el arma es pequeña o está parcialmente ocluida. Estos

resultados sugieren que el contexto de la escena puede ser tan determinante como la arquitectura elegida.

Conviene subrayar una limitación transversal de los detectores de una etapa. Pese a su ventaja en velocidad, siguen enfrentando dificultades con objetos pequeños, que es exactamente la condición predominante en videovigilancia (Berardini et al., 2024; Lim et al., 2019; Santos et al., 2024). Las cifras elevadas reportadas deben leerse, además, en el marco de conjuntos de datos propios y no directamente comparables (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024).

3.4. Modelos híbridos y arquitecturas en cascada

Una línea de diseño distinta consiste en encadenar varios modelos, cada uno con una función acotada, en lugar de confiar todo el problema a un único detector. La idea es reducir el cómputo, la tasa de falsos positivos, o ambos, distribuyendo la decisión en fases. Estas propuestas se conocen como esquemas de doble etapa o arquitecturas en cascada, y aparecen tanto en despliegues de computación en el borde (edge computing) como en configuraciones distribuidas (Berardini et al., 2024; Fernandez-Testa & Salcedo, 2024).

Berardini et al. (2024) aplican esta lógica para aliviar el cómputo; es decir, detectan primero a la persona y buscan el arma solo dentro de esa región, lo que reduce el área de búsqueda y permite ejecutar el sistema en *hardware* limitado como Jetson Nano. Fernandez-Testa y Salcedo (2024) persiguen un objetivo distinto con el mismo principio, ya que combinan un detector de armas con una red 3DCNN alojada en la nube que confirma el evento mediante el análisis del contexto de robo. La diferencia entre ambos es informativa. El primero encadena modelos para reducir carga computacional, el segundo para reducir falsos positivos. En los dos casos, el segundo componente actúa como verificador de la hipótesis del primero.

Jaleel et al. (2022) aplican este mismo principio bajo computación en el borde, con un esquema en el que un detector de regresión de la familia YOLOv4 propone los candidatos y un clasificador basado en regiones, Faster R-CNN o R-FCN, los confirma antes de emitir la alarma. Los autores reportan una exactitud del 90 %, con 94 % de precisión, 86 % de *recall* y 89 % de F1, y atribuyen la mejora a esa etapa de verificación.

La lógica común de estos diseños es tratar la detección como una decisión escalonada y no instantánea. Esto conecta con las estrategias de reducción de falsos positivos y falsos negativos descritas en la literatura, que incluyen la incorporación de objetos de confusión, la estimación de pose y la confirmación mediante análisis temporal o de contexto (Bhatti et al., 2021; Fernandez-Testa & Salcedo, 2024; Lamas et al., 2022). En definitiva, se trata de compensar mediante el diseño del sistema aquello que la evidencia visual aislada no permite resolver. La Tabla 2 sintetiza las familias descritas.

Tabla 2

Familias de modelos de aprendizaje profundo aplicadas a la detección de armas de fuego

Familia	Principio de funcionamiento	Fortaleza y limitación reportadas	Citas
Enfoques clásicos	Descriptores diseñados manualmente más un clasificador.	Superados por el aprendizaje profundo en velocidad y exactitud.	(Salido et al., 2021; Yadav et al., 2023)
Dos etapas (Faster R-CNN)	Propuesta de regiones seguida de clasificación y ajuste.	Se emplean cuando se prioriza la verificación; requieren mayor coste computacional.	(Olmos et al., 2018; Santos et al., 2024; Zahrawi & Shaalan, 2023)
Una etapa (YOLO)	Predicción conjunta de clase y localización en una sola pasada.	Adecuados para priorizar velocidad; mantienen dificultades con objetos pequeños.	(Ahmed et al., 2022; Bhatti et al., 2021; Rao et al., 2026; Yadav et al., 2024)
Una etapa (RetinaNet)	Detector denso de una pasada sobre la imagen completa.	Desempeño competitivo; puede complementarse con información contextual.	(Salido et al., 2021)
Híbridos y en cascada	Encadenamiento de modelos con funciones acotadas.	Reducen cómputo o falsos positivos; añaden complejidad de despliegue.	(Berardini et al., 2024; Fernandez-Testa & Salcedo, 2024)

4. Evaluación del desempeño y aplicaciones

Evaluar un detector de armas exige responder dos preguntas simultáneas: qué tan bien acierta y qué tan rápido lo hace. Ambas dimensiones están en tensión en videovigilancia porque una alerta correcta y tardía pierde valor operativo. Esta sección revisa las métricas utilizadas en el campo, compara el desempeño reportado y describe los contextos de implementación en los que estos modelos han sido desplegados.

4.1. Métricas utilizadas

La precisión promedio o AP resume la relación entre precisión y *recall* obtenida al variar el umbral de confianza del detector. El mAP corresponde al promedio del AP entre las clases evaluadas y, según el protocolo, puede calcularse con un único umbral de intersección sobre unión de 0,50 (mAP50) o promediarse entre umbrales de 0,50 a 0,95 en pasos de 0,05 (mAP50-95). En la literatura revisada se reportan, por ejemplo, un mAP de 91,73 % (Bhatti et al., 2021), un mAP de 92,1 % (Ahmed et al., 2022) y un mAP50 de 88,54 % (Rao et al., 2026). Estos valores solo son comparables cuando se emplean el mismo conjunto de evaluación, definición de métrica y protocolo experimental.

Junto al mAP se emplean la precisión, el *recall* y el F1. La precisión indica qué proporción de las detecciones emitidas es correcta y se relaciona directamente con la carga de falsas alarmas. El *recall* indica qué proporción de las armas presentes es detectada efectivamente. Por ejemplo, si el sistema emite 100 alertas y 94

corresponden realmente a armas, su precisión es del 94 %; si existen 100 armas en las imágenes y detecta 86, su *recall* es del 86 %. El F1 equilibra ambas y aparece de forma recurrente en la literatura; por ejemplo, con un valor de 91 % en un sistema basado en YOLOv4 (Bhatti et al., 2021) o como criterio de comparación entre RetinaNet y YOLOv3 (Salido et al., 2021).

Otra métrica relevante es la velocidad de inferencia, expresada en fotogramas por segundo. Su importancia deriva del propósito del sistema, que en la mayoría de los trabajos revisados es la alerta temprana (Olmos et al., 2018; Rao et al., 2026; Salido et al., 2021). Los valores reportados varían de forma considerable según el *hardware*, desde 85,7 FPS en una GPU de escritorio (Ahmed et al., 2022) hasta 4,43 FPS en un despliegue distribuido con optimización TensorRT (Fernandez-Testa & Salcedo, 2024).

Existe además una dimensión que la literatura trata como crítica sin reducirla siempre a una métrica única: la latencia de la alerta. Olmos et al. (2018) informan activación de alarma en menos de 0,2 segundos en 27 de 30 escenas, formulando así el criterio operativo en términos de tiempo de respuesta y no de exactitud agregada. Esta manera de reportar resulta difícil de contrastar con la de los estudios que solo publican mAP, y evidencia que el campo no ha convergido en un criterio común de evaluación.

4.2. Comparación de rendimiento: precisión frente a velocidad

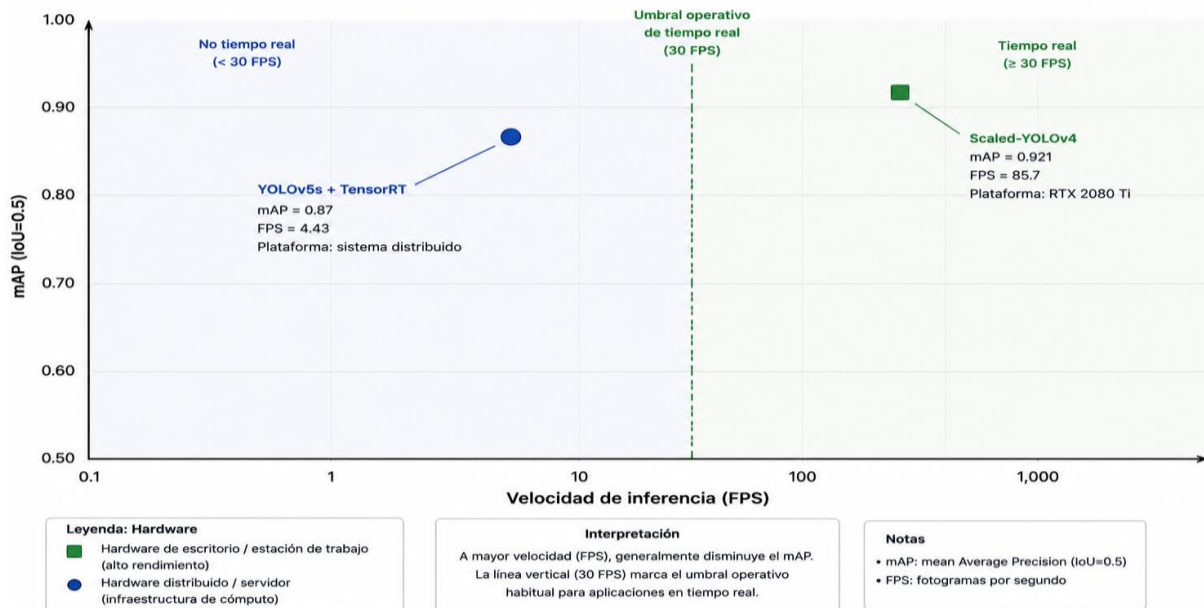
El compromiso entre precisión y velocidad organiza buena parte de los resultados publicados. En los estudios revisados, los detectores de dos etapas suelen emplearse cuando se prioriza la verificación de las propuestas y el control de falsas detecciones, mientras que la familia YOLO se selecciona con frecuencia en aplicaciones que priorizan la velocidad de inferencia (Ahmed et al., 2022; Bhatti et al., 2021; Olmos et al., 2018). Por tanto, la elección de la arquitectura debe responder al coste relativo de una falsa alarma, una omisión y una alerta tardía, y no a una supuesta superioridad universal.

La comparación entre estudios recientes ilustra el problema. Yadav et al. (2024) reportan un 96,80 % de exactitud con YOLOv7 y Salido et al. (2021) un 96,36 % de AP con RetinaNet, cifras que parecerían situar a ambos modelos por encima del 88,54 % de mAP50 que Rao et al. (2026) obtuvieron con YOLOv8m. Sin embargo, esa lectura sería incorrecta, porque cada valor procede de un conjunto de datos distinto e incluso de métricas que no miden lo mismo. Solo la comparación interna de Rao et al. (2026), realizada sobre datos fijos, permite una conclusión sobre superioridad relativa entre dos modelos.

Esta advertencia metodológica es importante para interpretar el estado del campo. Las diferencias observadas entre arquitecturas pueden ser menores que las diferencias entre los conjuntos de datos sobre los que fueron medidas (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024). En consecuencia, la Figura 4 y la Tabla 3 deben leerse como evidencia del rango de desempeño alcanzable y no como una comparación controlada.

Figura 4

Relación entre el mAP reportado y la velocidad de inferencia en dos configuraciones de referencia



Nota: Elaborada a partir de los datos reportados por Ahmed et al. (2022) y Fernandez-Testa y Salcedo (2024). Los puntos proceden de conjuntos de datos distintos y no son directamente comparables entre sí (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024).

Tabla 3

Resultados representativos reportados en la literatura revisada

Estudio	Modelo	Desempeño reportado	Plataforma
Olmos et al.	Faster R-CNN	Alarma activada en 27/30 escenas en menos de 0,2 s	No especificada
Bhatti et al.	YOLOv4	mAP = 91,73 %; F1 = 91 % en CCTV en tiempo real	Tiempo real
Ahmed et al.	Scaled-YOLOv4	mAP = 92,1 %; 85,7 FPS	RTX 2080Ti; Jetson Nano con TensorRT
Salido et al.	RetinaNet / YOLOv3	AP = 96,36 %; recall = 97,23 %	No especificada
Yadav et al.	YOLOv7	Exactitud = 96,80 %; recall = 90,37 %	No especificada
Rao et al.	YOLOv8m vs YOLOv5m	mAP50 = 88,54 % frente a 81,69 %	No especificada
Fernandez-Testa y Salcedo	YOLOv5s + 3DCNN	mAP = 87 %; 4,43 FPS con confirmación en la nube	Sistema distribuido con TensorRT

Nota: Los valores se transcriben tal como fueron informados en cada fuente; el mAP del último estudio, informado en escala 0–1, se expresa aquí en porcentaje para uniformar la escala. La ausencia de información sobre la plataforma de ejecución en varios estudios evidencia diferencias en el nivel de detalle del reporte. Los valores proceden de conjuntos de datos distintos y no son directamente comparables entre sí (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024).

4.3. Implementación en tiempo real y computación en el borde

La computación en el borde permite ejecutar el modelo directamente en el dispositivo que captura el video, en lugar de enviar los fotogramas a un servidor remoto. Esta orientación es una dirección clara en el campo porque busca menor latencia, menor tráfico de red y mayor privacidad (Ahmed et al., 2022; Fathy & Saleh, 2022). Se trata de una respuesta directa a las limitaciones de escalar la inferencia centralizada a muchas cámaras.

La literatura documenta despliegues concretos de esta idea. Se han demostrado sistemas sobre Jetson Nano, cámaras IP inteligentes y arquitecturas mixtas de borde y nube con envío selectivo de cuadros para reducir el ancho de banda y confirmar los eventos (Berardini et al., 2024; Fathy & Saleh, 2022; Qi et al., 2021). La optimización del modelo para *hardware* restringido, por ejemplo, mediante TensorRT sobre Jetson Nano, forma parte de estas propuestas (Ahmed et al., 2022). Jaleel et al. (2022) plantean su sistema de vigilancia proactiva bajo esta misma orientación y sitúan el procesamiento cerca de la cámara para que la alerta alcance al personal de seguridad dentro de un plazo útil.

Los estudios reportan también velocidades considerablemente diferentes según la configuración empleada: Fernandez-Testa y Salcedo (2024) informan 4,43 FPS en un sistema distribuido con confirmación en la nube, mientras que Ahmed et al. (2022) alcanzan 85,7 FPS sobre una GPU de escritorio. Estas cifras sugieren que el despliegue en dispositivos restringidos puede introducir un compromiso entre autonomía y capacidad de procesamiento, aunque no permiten una comparación directa debido a las diferencias de modelos, datos y plataformas experimentales.

4.4. Casos de aplicación y limitaciones observadas

Las aplicaciones reportadas se orientan a acortar el tiempo entre la detección y la respuesta humana. Más allá del propio detector, existen propuestas que incorporan alertas por voz, llamadas automáticas y aplicaciones móviles para acelerar la respuesta operativa (Al-Mousa et al., 2023; Raja Ramesh et al., 2024). Estas capas de notificación reconocen que el valor del sistema no se agota en la inferencia, sino en la cadena completa hasta el operador. En esa misma dirección, Razzaq et al. (2023) integran la generación de alertas en tiempo real hacia las autoridades competentes como parte del propio sistema y no como una capa añadida a posteriori.

En conjunto, el aprendizaje profundo para la detección de armas de fuego funciona bien en prototipos y despliegues controlados, especialmente con YOLO y Faster R-CNN (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024). Sin embargo, las limitaciones observadas son consistentes entre estudios. La generalización sigue restringida por conjuntos de datos heterogéneos, armas pequeñas, oclusión y falsas alarmas (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024). A ello se añade el sesgo hacia armas cortas visibles y el planteamiento del sistema como alarma temprana antes que como reconocimiento fino (Olmos et al., 2018; Rao et al., 2026; Salido et al., 2021).

5. Tendencias y desafíos futuros

El estado del campo permite identificar direcciones de desarrollo que ya cuentan con evidencia inicial, junto con problemas que permanecen abiertos. Las tendencias más consolidadas apuntan a ejecutar el modelo en el propio dispositivo de captura mediante computación en el borde, a la reducción sistemática de falsos positivos y a la construcción de conjuntos de datos más representativos. Los desafíos, por su parte, se concentran en la generalización a escenarios distintos de los vistos durante el entrenamiento y en las implicaciones de instalar estos sistemas en espacios de uso público (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024).

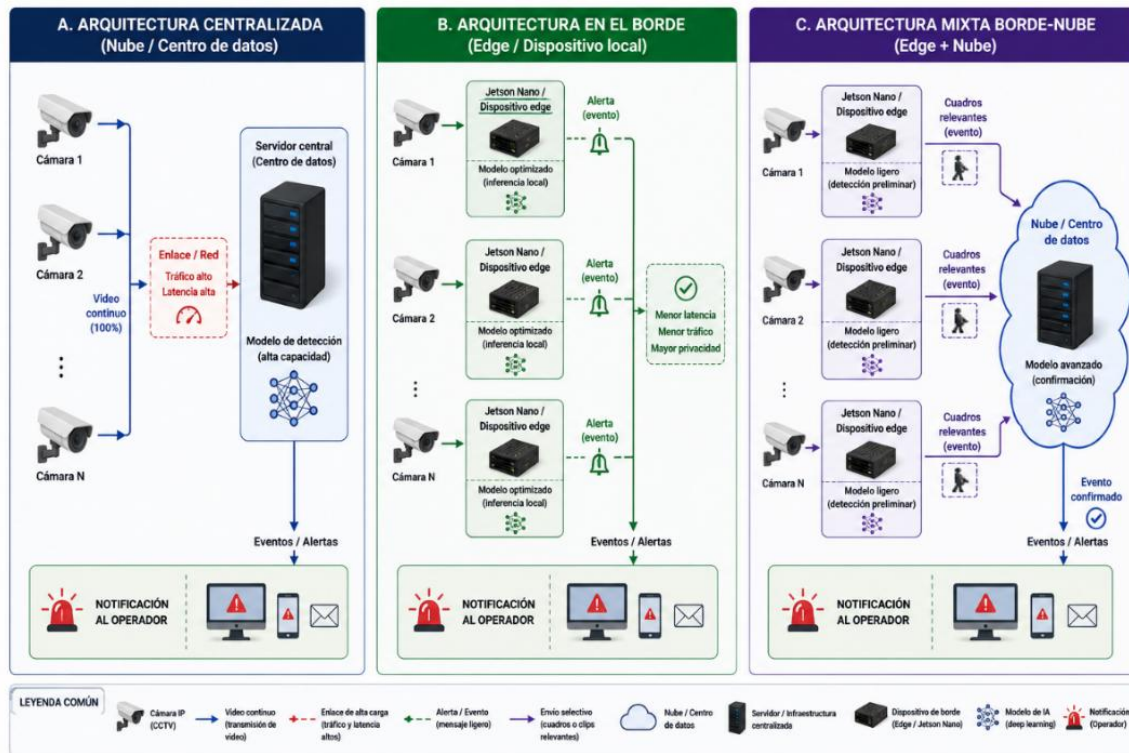
5.1. Modelos ligeros y computación en el borde

La primera tendencia consiste en llevar el modelo al dispositivo de captura mediante versiones ligeras y optimizadas. Esta dirección responde a tres motivaciones documentadas de forma explícita: reducir la latencia, disminuir el tráfico de red y aumentar la privacidad (Ahmed et al., 2022; Fathy & Saleh, 2022). Las demostraciones existentes sobre Jetson Nano, cámaras IP inteligentes y arquitecturas de borde y nube con envío selectivo de cuadros confirman su viabilidad técnica (Berardini et al., 2024; Fathy & Saleh, 2022; Qi et al., 2021).

El desafío pendiente en esta línea es cerrar la brecha de rendimiento respecto al *hardware* de escritorio. Los valores reportados en despliegues distribuidos, del orden de unos pocos fotogramas por segundo (Fernandez-Testa & Salcedo, 2024), contrastan con los obtenidos sobre GPU (Ahmed et al., 2022). Reducir esa distancia sin sacrificar mAP es una condición para el despliegue a escala. La Figura 5 contrasta las tres arquitecturas de despliegue documentadas en la literatura.

Figura 5

Arquitecturas de despliegue: inferencia centralizada, en el borde y mixta borde-nube



5.2. Reducción de falsos positivos y uso del contexto temporal

Una segunda tendencia agrupa las estrategias orientadas a evitar la alarma injustificada. La literatura reporta cuatro vías complementarias: incluir objetos de confusión como móviles o carteras en el entrenamiento, añadir información de pose, preprocesar imágenes borrosas o de baja luminosidad, y confirmar el arma mediante análisis temporal o de contexto de robo (Bhatti et al., 2021; Goenka & Sitara, 2022; Salido et al., 2021). Estas vías comparten la idea de que la apariencia aislada del objeto es una evidencia insuficiente.

El uso del contexto temporal resulta particularmente prometedor. La confirmación de eventos mediante una red 3DCNN alojada en la nube muestra que el movimiento y la secuencia aportan información que un solo fotograma no contiene (Fernandez-Testa & Salcedo, 2024). De manera análoga, la estimación de pose humana puede contribuir a mitigar falsos negativos cuando el arma es pequeña u ocluida (Lamas et al., 2022). Razzaq et al. (2023) siguen una orientación relacionada al acoplar un detector YOLOv5 de armas con un módulo de reconocimiento de violencia basado en ResNet y Bi-LSTM, de modo que la apariencia del objeto se contrasta con la dinámica de la escena antes de notificar a las autoridades.

5.3. Conjuntos de datos más representativos y etiquetado más rico

La escasez de conjuntos de datos estandarizados y suficientemente realistas continúa siendo un obstáculo estructural del campo (Bhatti et al., 2021; Lim et al., 2019; Qi et

al., 2021). Mientras cada grupo entrene y evalúe sobre bases propias, construidas con imágenes de internet, YouTube o grabaciones CCTV (Bhatti et al., 2021; Qi et al., 2021; Santos et al., 2024), los resultados publicados seguirán siendo difícilmente comparables. Por ende, la disponibilidad de conjuntos comunes de evaluación es una necesidad metodológica antes que una mejora incremental.

Un segundo aspecto de esta línea es la ampliación del alcance del problema. El predominio de pistolas y armas cortas visibles deja fuera escenarios relevantes de seguridad, como las armas ocultas (Olmos et al., 2018; Rao et al., 2026; Salido et al., 2021). Cubrir esos escenarios exigirá conjuntos de datos diseñados específicamente para condiciones de oclusión severa, así como para la variabilidad de escala y de iluminación reportada como crítica (Abins et al., 2024; Santos et al., 2024).

A la cobertura de armas ocultas debe añadirse el registro explícito de la procedencia del material en la documentación de cada conjunto de datos, indicando si el arma fotografiada es real, una réplica o una representación sintética. Esta información, hoy ausente o mencionada de forma incidental en la mayoría de los trabajos revisados, condiciona la interpretación de los resultados y su transferibilidad a condiciones abiertas.

5.4. Aspectos éticos y de privacidad

El despliegue de detectores automáticos sobre cámaras de espacios públicos plantea consideraciones que exceden lo técnico. La literatura del dominio aborda la privacidad principalmente desde la arquitectura del sistema, señalando que la inferencia en el borde la favorece al evitar la transmisión continua de video hacia servidores externos (Ahmed et al., 2022; Fathy & Saleh, 2022). En la misma dirección operan las arquitecturas de borde y nube con envío selectivo de cuadros, que limitan qué información sale del dispositivo (Fathy & Saleh, 2022; Qi et al., 2021).

Además, existe una dimensión ética asociada al error del sistema. Dado que la mayoría de las propuestas se conciben como alarmas tempranas y no como reconocimiento forense fino (Olmos et al., 2018; Rao et al., 2026; Salido et al., 2021), la decisión final permanece necesariamente en el operador humano. Esta condición otorga a la tasa de falsas alarmas un peso que no es solo operativo, sino también ético, en la medida en que cada alerta errónea moviliza una respuesta de seguridad sobre una persona.

5.5. Brechas de investigación

El balance del estado del arte muestra un campo maduro en prototipos y limitado en generalización. El aprendizaje profundo funciona bien en despliegues controlados, especialmente con YOLO y Faster R-CNN, pero la transferencia a condiciones abiertas sigue restringida por conjuntos de datos heterogéneos, armas pequeñas, oclusión y falsas alarmas (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024). La detección de objetos pequeños permanece como el principal cuello de botella no resuelto (Berardini et al., 2024; Bhatti et al., 2021; Santos et al., 2024).

De este balance se desprenden tres líneas prioritarias. La primera es la evaluación sobre conjuntos comunes que permita comparaciones controladas entre arquitecturas (Santos et al., 2024). La segunda es el tratamiento explícito de la escala reducida del arma en el cuadro, más allá del cambio de versión del detector (Berardini et al., 2024; Lim et al., 2019). La tercera es la integración del sistema en la cadena de respuesta operativa, donde ya existen propuestas de alertas por voz, llamadas automáticas y aplicaciones móviles (Al-Mousa et al., 2023; Raja Ramesh et al., 2024).

A estas tres líneas se suma una cuarta, hasta ahora poco atendida. Los trabajos revisados no incorporan de forma sistemática las armas de imitación como clase diferenciada ni reportan tasas de falsos positivos desagregadas para este tipo de objeto. Caracterizar empíricamente el comportamiento de los detectores ante réplicas realistas constituye, por tanto, una tarea pendiente con implicaciones directas para la interpretación operativa de las alertas.

6. Discusión

El análisis de la literatura muestra que el desempeño de los sistemas de detección de armas no depende únicamente de la arquitectura utilizada. El campo presenta un grado importante de desarrollo técnico en condiciones controladas, pero su transferencia a entornos operativos abiertos sigue condicionada por factores que anteceden a la elección del detector. Los cuatro ejes examinados, esto es, los conjuntos de datos, las familias de modelos, los criterios de evaluación y las condiciones de despliegue, no operan como problemas independientes, sino como restricciones encadenadas: el material disponible fija el techo de desempeño alcanzable, la métrica reportada determina qué parte de ese desempeño resulta visible y el *hardware* de destino delimita cuáles soluciones son efectivamente ejecutables. A continuación, se discute cada eje y, después, un aspecto transversal poco atendido por la literatura.

En lo que respecta a los conjuntos de datos, el hallazgo más consistente es que la práctica de construir bases propias resuelve un problema inmediato y genera uno estructural. Cada grupo compone su material a partir de imágenes de internet, fotogramas de video, grabaciones CCTV y, de manera creciente, imágenes sintéticas (Bhatti et al., 2021; Qi et al., 2021; Salazar González et al., 2020; Santos et al., 2024). La comunidad dispone así de numerosos conjuntos, pero carece de un referente común; por ello, la heterogeneidad del material se convierte en una variable no controlada que se confunde con el efecto de la arquitectura. A ello, se añade que el reto principal, el tamaño reducido del arma dentro del cuadro, es un atributo del material antes que, del modelo, puesto que ninguna versión posterior del detector modifica el número de píxeles informativos que entrega una cámara de vigilancia (Berardini et al., 2024; Lim et al., 2019; Santos et al., 2024).

En cuanto a las familias de modelos, la discusión no debería plantearse en términos de superioridad absoluta, sino de correspondencia entre el diseño y el criterio operativo. En los estudios revisados, los detectores de dos etapas se emplean cuando

se prioriza la verificación de propuestas y el control de falsos positivos (Arif et al., 2022; Olmos et al., 2018; Santos et al., 2024), mientras que los de una etapa se seleccionan con frecuencia por su balance entre desempeño y velocidad (Ahmed et al., 2022; Bhatti et al., 2021; Rao et al., 2026). Los esquemas en cascada muestran que ambas lógicas son combinables (Berardini et al., 2024; Fernandez-Testa & Salcedo, 2024; Jaleel et al., 2022). Lo relevante es que estas propuestas desplazan la decisión desde el detector aislado hacia la arquitectura completa del sistema.

La evaluación constituye el punto más frágil del campo. La coexistencia de mAP, exactitud, precisión, *recall*, F1 y velocidad de inferencia, medidos sobre conjuntos y *hardware* diferentes, produce una apariencia de progreso comparable que el procedimiento no sustenta (Fernandez-Testa & Salcedo, 2024; Santos et al., 2024). La evidencia que permite ordenar arquitecturas es la comparación interna sobre datos fijos, como la realizada entre YOLOv8m y YOLOv5m (Rao et al., 2026). Existe además una magnitud que la mayoría de los trabajos no reporta de manera uniforme: la latencia completa de la alerta, pese a que esta determina el valor operativo del sistema en el escenario declarado como objetivo (Olmos et al., 2018). Mientras el campo no adopte conjuntos y protocolos comunes, las cifras publicadas describirán el rango de desempeño alcanzable, pero no una jerarquía general entre propuestas.

Respecto de las condiciones de despliegue, la computación en el borde aparece como una dirección relevante, motivada por la reducción de la latencia, la disminución del tráfico de red y la protección de la privacidad (Ahmed et al., 2022; Fathy & Saleh, 2022; Jaleel et al., 2022). Sin embargo, su discusión no debe omitir el coste asociado. La distancia entre las decenas de fotogramas por segundo obtenidas sobre GPU de escritorio y los pocos fotogramas por segundo de un despliegue distribuido (Ahmed et al., 2022; Fernandez-Testa & Salcedo, 2024) indican que la descentralización implica un intercambio entre autonomía del dispositivo y capacidad de cómputo. La arquitectura elegida también determina qué información abandona la cámara y, por tanto, el perfil de privacidad del sistema desplegado (Fathy & Saleh, 2022; Jaleel et al., 2022).

Un aspecto que la literatura revisada aborda solo de manera tangencial es el tratamiento de las armas de imitación, presente de dos formas distintas. Por un lado, las réplicas, las escenificaciones y las imágenes sintéticas constituyen fuentes habituales del material de entrenamiento y prueba, debido a la dificultad de obtener grabaciones reales de incidentes armados (Lim et al., 2019; Qi et al., 2021; Salazar González et al., 2020). Por otro lado, las réplicas no suelen figurar como clase explícita ni como categoría separada en el análisis de errores, a diferencia de objetos cotidianos de confusión (Abins et al., 2024; Bhatti et al., 2021; Goenka & Sitara, 2022). Esta omisión limita la validez ecológica de la evidencia, pues la equivalencia entre el material simulado y las condiciones reales de CCTV suele asumirse sin una comprobación específica (Berardini et al., 2024; Lim et al., 2019; Santos et al., 2024).

Este análisis revela una limitación fundamental de los sistemas basados exclusivamente en información visual. Cuando una réplica reproduce de manera realista la apariencia externa de un arma, un detector que opera exclusivamente con información visual convencional no puede garantizar su diferenciación respecto de un arma real. Más que una deficiencia atribuible únicamente al entrenamiento o a la arquitectura, se trata de un límite de la información disponible para el sistema. En consecuencia, estos detectores identifican una apariencia compatible con un arma, pero no determinan por sí solos la existencia de una amenaza. Esta precisión es coherente con la formulación del dominio como alarma temprana y no como reconocimiento forense fino (Goenka & Sitara, 2022; Olmos et al., 2018; Rao et al., 2026). El problema se traslada así al protocolo de respuesta, donde la verificación humana posterior a la alerta resulta tan relevante como el detector.

7. Conclusiones

Esta revisión examinó el estado de la detección de armas de fuego en sistemas de videovigilancia mediante aprendizaje profundo a partir de cuatro ejes: los conjuntos de datos, las familias de modelos, los criterios de evaluación y las condiciones de despliegue. El campo presenta soluciones con resultados sólidos en condiciones controladas, pero su transferencia a entornos operativos abiertos permanece limitada por factores anteriores a la arquitectura del detector. El principal factor es la naturaleza del material disponible: el tamaño reducido que ocupa el arma dentro del cuadro se identifica como el cuello de botella no resuelto, junto con la oclusión, la variabilidad de ángulo e iluminación y la presencia de objetos visualmente similares.

En cuanto a las soluciones, los estudios revisados muestran que los detectores de dos etapas suelen seleccionarse cuando se prioriza la verificación de propuestas, mientras que los de una etapa se utilizan con frecuencia cuando la velocidad de inferencia es un requisito central; los esquemas en cascada, por su parte, distribuyen la decisión para reducir el cómputo o las falsas alarmas. Sin embargo, las cifras publicadas no permiten ordenar universalmente las arquitecturas, porque proceden de conjuntos propios y métricas heterogéneas: solo las comparaciones internas sobre datos fijos autorizan conclusiones sobre superioridad relativa. En el plano del despliegue, llevar la inferencia al dispositivo de captura responde a exigencias de latencia, ancho de banda y privacidad, aunque puede reducir la capacidad de procesamiento. Por ello, la computación en el borde debe entenderse como un compromiso explícito y no como una mejora incondicional.

De manera transversal, los resultados permiten precisar el alcance real de estos sistemas. La formulación del problema como alarma temprana, el entrenamiento con réplicas o material simulado y la dificultad de distinguir visualmente una imitación realista indican que estos sistemas identifican una apariencia compatible con un arma, pero no determinan por sí solos una amenaza, lo que orienta el diseño operativo hacia el protocolo de verificación humana posterior a la alerta. De ello se desprenden cuatro líneas prioritarias: adoptar conjuntos comunes de evaluación, tratar explícitamente la

escala reducida del objeto, documentar la procedencia del material de entrenamiento y caracterizar el comportamiento ante armas de imitación, e integrar el sistema en la cadena completa de respuesta operativa.

Finalmente, conviene señalar las limitaciones de esta revisión. El análisis se apoya en un conjunto acotado de trabajos y en el desempeño tal como fue reportado por sus autores, sin verificación experimental independiente. La heterogeneidad de los conjuntos de datos y de las métricas, señalada aquí como un hallazgo del campo, opera también como una restricción sobre la propia síntesis, puesto que impide agregar los resultados en una comparación cuantitativa única. Por ello, las cifras reseñadas deben leerse como indicativas del rango de desempeño alcanzable y no como una medición comparable entre propuestas.

Contribución de los autores: conceptualización, C.B.M. y O.E.; metodología, C.B.M. y O.E.; validación, O.E. y V.T.L.; análisis formal, C.B.M., J.S.C.B. y J.L.T.; investigación, C.B.M. y V.T.L.; recursos, J.S.C.B. y J.L.T.; curación de datos, J.S.C.B. y J.L.T.; redacción del borrador original, C.B.M., J.S.C.B. y V.T.L.; redacción, revisión y edición, V.T.L., J.L.T. y O.E.; visualización, J.S.C.B. y J.L.T.; supervisión, O.E. y V.T.L.; administración del proyecto, C.B.M. y O.E.

Financiamiento: esta investigación no recibió financiamiento de agencias públicas, del sector comercial ni de entidades sin fines de lucro.

Declaración de disponibilidad de datos: este trabajo no generó datos primarios. Las fuentes analizadas se identifican en las referencias bibliográficas. La matriz de extracción elaborada para la síntesis está disponible previa solicitud razonable al autor de correspondencia.

Declaración de conflictos de interés: los autores declaran no tener ningún conflicto de intereses.

Referencias Bibliográficas

- Abins, A., Priyadharshini, P., Rohidh, G. C., & Cheran, R. (2024). Weapon Recognition in CCTV Videos: Deep Learning Solutions for Rapid Threat Identification. *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, 1–8. <https://doi.org/10.1109/ic-ETITE58242.2024.10493569>
- Ahmed, S., Bhatti, M., Khan, M. G., Löfström, B., & Shahid, M. (2022). Development and Optimization of Deep Learning Models for Weapon Detection in Surveillance Videos. *Applied Sciences*, 12(12), 5772. <https://doi.org/10.3390/app12125772>
- Akshaya, P., Reddy, P. B., Panuganti, P., Gurusai, P., & Subhahan, A. (2023). Automatic weapon detection using Deep Learning. *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, 1–7. <https://doi.org/10.1109/RMKMATE59243.2023.10369889>
- Al-Mousa, A., Alzaibaq, O., & Abu-Hashyeh, Y. (2023). Deep Learning-Based Real-Time Weapon Detection System. *International Journal of Computing and Digital Systems*, 14(1), 531–540. <https://doi.org/10.12785/ijcds/140141>

- Arif, E., Shahzad, S. K., Mustafa, R., Jaffar, M. A., & Iqbal, M. W. (2022). Deep Neural Networks for Gun Detection in Public Surveillance. *Intelligent Automation & Soft Computing*, 32(2), 909–922. <https://doi.org/10.32604/iasc.2022.021061>
- Bamnolkar, A., Khakre, K., Gadge, A., Waje, M., Raut, V., & Khan, M. (2025). Cutting Edge Weapon Detection in Real-Time CCTV Videos. *2025 10th International Conference on Signal Processing and Communication (ICSC)*, 345–350. <https://doi.org/10.1109/ICSC64553.2025.10968883>
- Berardini, D., Migliorelli, L., Galdelli, A., Frontoni, E., Mancini, A., & Moccia, S. (2024). A deep-learning framework running on edge devices for handgun and knife detection from indoor video-surveillance cameras. *Multimedia Tools and Applications*, 83(7), 19109–19127. <https://doi.org/10.1007/s11042-023-16231-x>
- Bhatti, M. T., Khan, M. G., Aslam, M., & Fiaz, M. J. (2021). Weapon Detection in Real-Time CCTV Videos Using Deep Learning. *IEEE Access*, 9, 34366–34382. <https://doi.org/10.1109/ACCESS.2021.3059170>
- Corral-Sanz, P., Barreiro-Garrido, A., Moreno, A. B., & Sánchez, Á. (2024). On the influence of artificially distorted images in firearm detection performance using deep learning. *PeerJ Computer Science*, 10, e2381. <https://doi.org/10.7717/peerj-cs.2381>
- Fathy, C., & Saleh, S. N. (2022). Integrating Deep Learning-Based IoT and Fog Computing with Software-Defined Networking for Detecting Weapons in Video Surveillance Systems. *Sensors (Basel, Switzerland)*, 22(14), 5075. <https://doi.org/10.3390/s22145075>
- Fernandez-Testa, S., & Salcedo, E. (2024). Distributed Intelligent Video Surveillance for Early Armed Robbery Detection Based on Deep Learning. *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 1–6. <https://doi.org/10.1109/SIBGRAPI62404.2024.10716299>
- Goenka, A., & Sitara, K. (2022). Weapon Detection from Surveillance Images using Deep Learning. *2022 3rd International Conference for Emerging Technology (INCET)*, 1–6. <https://doi.org/10.1109/INCET54531.2022.9824281>
- Grega, M., Matiolański, A., Guzik, P., & Leszczuk, M. (2016). Automated Detection of Firearms and Knives in a CCTV Image. *Sensors (Basel, Switzerland)*, 16(1), 47. <https://doi.org/10.3390/s16010047>
- Jaleel, A., Khurshid, S. K., Mustafa, R., Mehmood Aamir, K., Tahir, M., & Ziar, A. (2022). Towards Proactive Surveillance through CCTV Cameras under Edge-Computing and Deep Learning. *Mathematical Problems in Engineering*, 2022(1), 7001388. <https://doi.org/10.1155/2022/7001388>
- Lamas, A., Tabik, S., Cano Montes, A., Pérez-Hernández, F., García, J., Olmos, R., & Herrera, F. (2022). Human pose estimation for mitigating false negatives in weapon detection in video-surveillance. *Neurocomputing*, 489(C), 488–503. <https://doi.org/10.1016/j.neucom.2021.12.059>
- Lim, J., Jobayer, M. I. Al, Baskaran, V. M., Lim, J. M., Wong, K., & See, J. (2019). Gun Detection in Surveillance Videos using Deep Neural Networks. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1998–2002. <https://doi.org/10.1109/APSIPAASC47483.2019.9023182>

- Olmos, R., Tabik, S., & Herrera, F. (2018). Automatic handgun detection alarm in videos using deep learning. *Neurocomputing*, 275, 66–72. <https://doi.org/10.1016/j.neucom.2017.05.012>
- Pavithra, T., Thangaraj, R., Pandiyan, P., Uma Rani, M., & Vadivelu, B. (2022). Real-Time Handgun Detection in Surveillance Videos based on Deep Learning Approach. 2022 *International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 689–693. <https://doi.org/10.1109/ICAAIC53929.2022.9793288>
- Qi, D., Tan, W., Liu, Z., Yao, Q., & Liu, J. (2021). A Dataset and System for Real-Time Gun Detection in Surveillance Video Using Deep Learning. 2021 *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 667–672. <https://doi.org/10.1109/SMC52423.2021.9659207>
- Raja Ramesh, M. R., Mamatha, P., Viswanadha Pavan Varma, P., Wasim Akram, S., & Chandu, T. (2024). Harnessing deep learning for video based weapon detection. *Journal of Artificial Intelligence, Machine Learning and Neural Network*, 4(5), 30–40. <https://doi.org/10.55529/jaimlInn.45.30.40>
- Rao, D., ManiSaiDeep, C., Sai, K., Reddy, K., & Kumar, G. (2026). Smart Surveillance System for Firearm Detection Using Deep Learning Model. *International Journal of Scientific Research in Engineering and Management*, 10(3), 1–9. <https://doi.org/10.55041/IJSREM58675>
- Razzaq, F. A., Chaudary, M. A., Fareed, S., Tariq, W., Waqas, M., & Javaid, S. (2023). Enhancing Public Safety: Detection of Weapons and Violence in CCTV Videos with Deep Learning. 2023 *25th International Multitopic Conference (INMIC)*, 1–6. <https://doi.org/10.1109/INMIC60434.2023.10465800>
- Salazar González, J. L., Álvarez-García, J. A., Rendón-Segador, F. J., & Carrara, F. (2023). Conditioned Cooperative training for semi-supervised weapon detection. *Neural Networks*, 167, 489–501. <https://doi.org/10.1016/j.neunet.2023.08.043>
- Salazar González, J. L., Zaccaro, C., Álvarez-García, J. A., Soria Morillo, L. M., & Sancho Caparrini, F. (2020). Real-time gun detection in CCTV: An open problem. *Neural Networks*, 132, 297–308. <https://doi.org/10.1016/j.neunet.2020.09.013>
- Salido, J., Lomas, V., Ruiz-Santaquiteria, J., & Deniz, O. (2021). Automatic Handgun Detection with Deep Learning in Video Surveillance Images. *Applied Sciences*, 11(13), 6085. <https://doi.org/10.3390/app11136085>
- Santos, T., Oliveira, H., & Cunha, A. (2024). Systematic review on weapon detection in surveillance footage through deep learning. *Computer Science Review*, 51, 100612. <https://doi.org/10.1016/j.cosrev.2023.100612>
- Yadav, P., Gupta, N., & Sharma, P. K. (2023). A comprehensive study towards high-level approaches for weapon detection using classical machine learning and deep learning methods. *Expert Systems with Applications*, 212, 118698. <https://doi.org/10.1016/j.eswa.2022.118698>
- Yadav, P., Gupta, N., & Sharma, P. K. (2024). Enhancing Video Surveillance with Deep Learning-Based Real-Time Handgun Detection and Tracking. In H. Kaur, V. Jakhetiya, P. Goyal, P. Khanna, B. Raman, & S. Kumar (Eds.), *Computer Vision and Image Processing* (pp. 173–184). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-58181-6_15

- Yeddula, N., & Reddy, B. E. (2022). Effective Deep Learning Technique for Weapon Detection in CCTV Footage. *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNBC)*, 1–6. <https://doi.org/10.1109/ICMNBC56175.2022.10031724>
- Zahrawi, M., & Shaalan, K. (2023). Improving video surveillance systems in banks using deep learning techniques. *Scientific Reports*, 13(1), 7911. <https://doi.org/10.1038/s41598-023-35190-9>