

Agradecimientos

Quiero agradecer a todas las personas que han hecho posible la realización de este Trabajo de Fin de Grado. En primer lugar, agradezco a mis tutores, José Flich y Pedro Juan López, por su orientación y apoyo durante todo el proceso. Su experiencia y conocimientos han sido fundamentales para el desarrollo de este proyecto.

Agradezco también a mi familia y en especial a mis padres, por su apoyo incondicional y su aliento constante durante todo el proceso de formación académica. Su sacrificio y dedicación han sido fundamentales para alcanzar esta meta.

Resum

Aquest Treball de Fi de Grau aborda el repte de la detecció automàtica de defectes en objectes que es desplacen en entorns industrials, com ara les línies de producció. Es presenta el disseny i la implementació d'un sistema complet de visió artificial basat en tècniques d'aprenentatge profund.

El nucli del sistema són els models de Xarxes Neuronals Convolucionals (CNN), específicament unes variants de l'arquitectura YOLO, entrenades per identificar i localitzar amb precisió defectes en els objectes en temps real. Per fer front a les restriccions computacionals i energètiques pròpies dels sistemes embebuts, els models s'han implementat sobre acceleradors hardware de baix consum de la sèrie NVIDIA Jetson. S'han aplicat tècniques avançades d'optimització, incloent la conversió a TensorRT i l'exploració de diferents precisions numèriques, per maximitzar la velocitat d'inferència (FPS) i minimitzar el consum energètic (Watts) sense comprometre significativament la precisió de la detecció (mAP).

El treball inclou una anàlisi exhaustiva del rendiment del sistema sota diverses configuracions de hardware i software, demostrant la viabilitat d'aplicar solucions d'intel·ligència artificial d'alt rendiment en escenaris industrials amb recursos limitats.

Paraules clau: Xarxes Neuronals Convolucionals, Detecció de defectes, Visió per computador, NVIDIA Jetson, Temps real, Optimització energètica

Resumen

Este trabajo de fin de grado aborda el desafío de la detección automática de defectos en objetos que se desplazan en entornos industriales, como líneas de producción. Se presenta el diseño y la implementación de un sistema completo de visión artificial basado en técnicas de aprendizaje profundo.

El núcleo del sistema son los modelos de Redes Neuronales Convolucionales (CNN), específicamente unas variantes de la arquitectura YOLO, entrenadas para identificar y localizar con precisión defectos en los objetos en tiempo real. Para hacer frente a las restricciones computacionales y energéticas propias de los sistemas embebidos, los modelos se han implementado sobre aceleradores hardware de bajo consumo de la serie NVIDIA Jetson. Se han aplicado técnicas avanzadas de optimización, incluyendo la conversión a TensorRT y la exploración de diferentes precisiones numéricas, para maximizar la velocidad de inferencia (FPS) y minimizar el consumo energético (Watts) sin comprometer significativamente la precisión de la detección (mAP).

El trabajo incluye un análisis exhaustivo del rendimiento del sistema bajo diversas configuraciones de hardware y software, demostrando la viabilidad de aplicar soluciones de inteligencia artificial de alto rendimiento en escenarios industriales con recursos limitados.

Palabras clave: Redes Neuronales Convolucionales, Detección de defectos, Visión por computador, NVIDIA Jetson, Tiempo real, Optimización energética

Abstract

This Bachelor's thesis addresses the challenge of automatic defect detection in moving objects within industrial environments, such as production lines. It presents the design and implementation of a complete computer vision system based on deep learning techniques.

The core of the system consists of Convolutional Neural Network (CNN) models, specifically variants of the YOLO architecture, trained to accurately identify and locate defects in objects in real-time. To address the computational and energy constraints typical of embedded systems, the models have been implemented on low-power hardware accelerators from the NVIDIA Jetson series. Advanced optimization techniques, including conversion to TensorRT and exploration of different numerical precisions, have been applied to maximize inference speed (FPS) and minimize energy consumption (Watts) without significantly compromising detection accuracy (mAP).

The thesis includes a comprehensive performance analysis of the system under various hardware and software configurations, demonstrating the feasibility of applying high-performance artificial intelligence solutions in industrial scenarios with limited resources.

Key words: Convolutional Neural Networks, Defect Detection, Computer Vision, NVIDIA Jetson, Real-Time, Energy Optimization

Índice general

Agradecimientos	III
Índice general	VII
Índice de figuras	IX
Índice de tablas	x
1 Introducción	1
1.1 Motivación	3
1.2 Objetivos	4
1.3 Estructura de la memoria	4
1.4 Colaboraciones	5
2 Conceptos previos	7
2.1 Fundamentos y avances en redes neuronales para visión artificial	7
2.1.1 Fundamentos de la inteligencia artificial	7
2.1.2 Tareas fundamentales en visión por computador	8
2.1.3 Arquitectura y funcionamiento de las CNN	9
2.1.4 Entrenamiento de las CNN	13
2.1.5 Detectores de dos etapas	15
2.1.6 Detectores de una etapa	16
2.1.7 Métricas de evaluación de modelos de detección de objetos	18
2.2 Aceleradores de procesamiento gráfico	19
2.2.1 Limitaciones del hardware tradicional	20
2.2.2 Arquitectura y funcionamiento de las GPUs	21
2.2.3 Serie Jetson: dispositivos para IA de bajo consumo	22
2.2.4 TensorRT	24
2.3 Seguimiento de objetos en tiempo real	25
2.3.1 Introducción al seguimiento de objetos	26
2.3.2 BYTETrack	27
2.3.3 Métricas de evaluación en seguimiento de objetos múltiples	30
3 Diseño e implementación de la solución	33
3.1 Análisis del problema	33
3.2 Entrenamiento de los modelos	35
3.3 Descripción del sistema	38
3.4 Diseño de las etapas del sistema	39
3.4.1 Captura de imágenes	39
3.4.2 Inferencia	40
3.4.3 Seguimiento	40
3.4.4 Escritura de resultados	41
3.5 Segmentación de las etapas del sistema	44
3.5.1 Secuencial	44
3.5.2 Segmentación en hilos	45
3.5.3 Segmentación en procesos	46
3.5.4 Segmentación heterogénea	48

3.5.5 Segmentación basada en procesos con memoria compartida	49
4 Evaluación de la solución	53
4.1 Metodología de evaluación y métricas de rendimiento	53
4.2 Variación de la configuración del sistema	55
4.2.1 Cantidad de objetos	55
4.2.2 Tipo de segmentación	62
4.2.3 Modelo y talla	64
4.2.4 Precisión numérica y acelerador de inferencia	66
4.2.5 Modo de energía y cores de la CPU	69
4.2.6 Dispositivos Jetson	70
4.3 Evaluación del seguimiento de objetos	71
5 Prueba de concepto	73
5.1 Diseño y construcción del sistema físico	73
5.2 Integración del sistema de visión artificial	75
5.3 Resultados experimentales y evaluación	78
6 Conclusiones	79
6.1 Objetivos alcanzados y dificultades	79
6.2 Aprendizaje	80
6.3 Relación con los estudios cursados	81
6.4 Trabajo futuro	81
Bibliografía	83
<hr/>	
Apéndices	
A Objetivos de desarrollo sostenible	87
B Código fuente	89

Índice de figuras

1.1	Evolución del interés público en inteligencia artificial según datos de Google Trends (2020-2025)	1
1.2	Proyección del consumo eléctrico de los centros de datos en el mundo . .	2
1.3	Esquema del sistema de visión artificial propuesto	3
2.1	Estructura de un perceptrón multicapa (MLP)	8
2.2	Ejemplo de HOG aplicado a una imagen	8
2.3	Tareas fundamentales en visión por computador	9
2.4	Relación entre Machine Learning, Deep Learning, CNN, Computer Vision y Human Vision	10
2.5	Operación de convolución sobre los pixeles de una imagen	10
2.6	Proceso de convolución aplicado a una imagen de un autobús	11
2.7	Ejemplo de operación de max-pooling con una ventana de 2×2 y un <i>stride</i> de 2	12
2.8	Ejemplo de una CNN simple	13
2.9	Gráfico de entrenamiento con early stopping	14
2.10	Proceso de búsqueda selectiva aplicado a una imagen	15
2.11	Arquitectura de R-CNN	16
2.12	Ejemplo de detección de objetos utilizando YOLO	17
2.13	Arquitectura de YOLO	18
2.14	Evolución histórica de las características de los microprocesadores (1970–2020)	20
2.15	Comparativa de arquitecturas CPU y GPU	21
2.16	Módulos Jetson de NVIDIA	23
2.17	Ejemplo de optimización de grafo computacional en TensorRT	24
2.18	Ejemplo de flujo de trabajo de optimización con TensorRT	25
2.19	Diagrama de flujo del filtro de Kalman	26
2.20	Comparativa de rendimiento de BYTETrack con otros algoritmos de seguimiento	28
2.21	Ejemplo de detección y seguimiento de objetos utilizando BYTETrack . .	28
2.22	Ejemplo visual de IDF1	31
3.1	Ejemplo de entorno simulado con canicas	34
3.2	Ejemplo de anotación de imágenes utilizando CVAT	36
3.3	Curvas de pérdida de entrenamiento y validación para las distintas tallas de modelos YOLOv5, YOLOv8 y YOLO11	38
3.4	Figura del sistema propuesto	39
3.5	Ejemplo de salida del sistema	42
3.6	Modelo del objeto rastreado	43
3.7	Diagrama de flujo del sistema sin segmentar	45
3.8	Diagrama de flujo del sistema segmentado en hilos	46
3.9	Diagrama de flujo del sistema segmentado en procesos	47
3.10	Diagrama de flujo del sistema segmentado	47

3.11	Diagrama de flujo del sistema segmentado en diferentes unidades de procesamiento	49
3.12	Diagrama de flujo del sistema segmentado en procesos con memoria compartida	50
3.13	Ejemplo de buffer circular	50
4.1	Cantidad de objetos en los vídeos de prueba	56
4.2	FPS por fotograma en función de la cantidad de objetos para los cuatro vídeos de prueba	56
4.3	Ejecución temporal de las etapas del sistema durante el vídeo de prueba con carga variable	57
4.4	Ejecución temporal de las etapas del sistema durante el vídeo de prueba 1 (carga baja y constante)	58
4.5	FPS y cantidad de objetos en función del tiempo para el vídeo de carga variable	59
4.6	Tiempo de ejecución de la etapa de seguimiento en función de la cantidad de objetos	60
4.7	Tiempo de ejecución de la etapa de escritura en función de la cantidad de objetos	61
4.8	Tiempos de ejecución de la etapa de inferencia para los diferentes modelos y tallas	65
4.9	Exportación del modelo YOLO11n con TensorRT a FP16 para su ejecución en la DLA	67
4.10	Tiempos de ejecución de la etapa de inferencia para las diferentes precisiones en GPU con TensorRT	68
5.1	Planos de la cinta transportadora	73
5.2	Cinta transportadora construida para la prueba de concepto	75
5.3	Diagrama de la arquitectura del sistema de visión artificial	76
5.4	Imagen de ejemplo de la prueba de concepto	78
B.1	Diagrama de clases del objeto DetectionTrackingPipeline	89
B.2	Diagrama de clases del programa principal	104
B.3	Diagrama de clases del objeto TrackerWrapper	114
B.4	Diagrama de clases del objeto SharedCircularBuffer	116

Índice de tablas

2.1	Comparativa técnica entre diferentes modelos NVIDIA Jetson	23
3.1	Ánalisis comparativo de variantes de YOLO (v5, v8, 11) indicando tamaño, parámetros, latencias CPU/GPU y la GPU específica utilizada para la medición de latencia GPU	36
3.2	Comparativa del rendimiento de los modelos YOLOv5, YOLOv8 y YOLO11 en términos de tiempo de entrenamiento, precisión, recall y mAP	37
4.1	Resultados del experimento con distintas cantidades de objetos	59

4.2	Resultados del experimento con distintos tipos de segmentación a máxima capacidad	62
4.3	Resultados del experimento con distintos tipos de segmentación a 30 fps	63
4.4	Resultados del experimento con distintos modelos y tallas a máxima capacidad con un vídeo de carga alta y constante	64
4.5	Resultados del experimento con distintos modelos y tallas a 30 fps	65
4.6	Resultados del experimento con distintas precisiones y aceleradores a máxima capacidad con un vídeo de carga alta y constante	66
4.7	Resultados del experimento con distintas precisiones y aceleradores a 30 fps con un vídeo de carga alta y constante	68
4.8	Resultados del experimento con distintos modelos y tallas a máxima capacidad con un vídeo de carga alta y constante	69
4.9	Resultados del experimento con distintos modelos y tallas a 30 fps con un vídeo de carga alta y constante	69
4.10	Resultados del experimento con distintos dispositivos Jetson a máxima capacidad con un vídeo de carga alta y constante	70
4.11	Resultados del experimento con distintos dispositivos Jetson a 30 fps con un vídeo de carga alta y constante	71
4.12	Resultados de la evaluación de métricas de seguimiento de objetos	72

CAPÍTULO 1

Introducción

Durante los últimos años, la Inteligencia Artificial (IA) ha experimentado un crecimiento en popularidad sin precedentes, transformando nuestra capacidad tecnológica con herramientas revolucionarias. Este avance ha sido impulsado por la disponibilidad de grandes volúmenes de datos, el desarrollo de algoritmos avanzados y las mejoras significativas en el hardware de procesamiento, que han permitido a las máquinas aprender y adaptarse a situaciones complejas. Algunos campos destacados de aplicación incluyen el procesamiento del lenguaje natural, la visión por computador y la robótica. En particular, la visión por computador ha visto un auge significativo, con aplicaciones en áreas como la seguridad, la medicina y la automoción. Esta creciente popularidad por el mundo de la IA se refleja en la evolución del interés público en ella, como muestra la Figura 1.1.



Figura 1.1: Evolución del interés público en inteligencia artificial según datos de Google Trends (2020-2025).

El progreso en visión por computador ha sido posible gracias a los avances en Red Neuronal Convolucionals (CNNs), que han revolucionado la capacidad de los sistemas para detectar y clasificar objetos en imágenes y vídeos con una gran precisión y velocidad.

Estos algoritmos de visión artificial requieren una potencia computacional significativa tanto para su entrenamiento como para su ejecución. Las Unidad Central de Pro-

cesamientos (CPUs) tradicionales resultan insuficientes para estas tareas, por lo que la industria ha desarrollado arquitecturas específicas como las Unidad de Procesamiento Gráficos (GPUs)[11, cap. 3, pp. 2-7], Unidad de Procesamiento Tensorial (TPUs)[3] y Acelerador de Aprendizaje Profundo (DLAs)[29]. Estos componentes están optimizados para ejecutar operaciones de entrenamiento e inferencia de manera eficiente, permitiendo implementar sistemas de visión artificial capaces de procesar información visual en tiempo real. Sin embargo, estos aceleradores suelen presentar un consumo energético elevado, lo que plantea importantes retos de eficiencia y sostenibilidad.

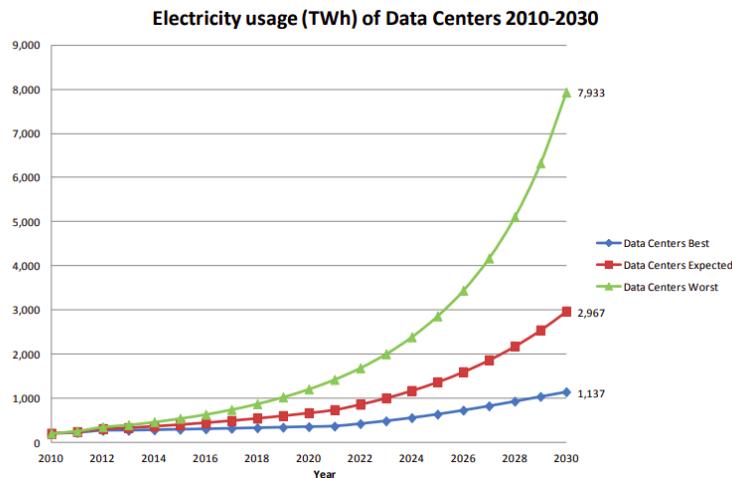


Figura 1.2: Proyección del consumo eléctrico de los centros de datos en el mundo. Extraído de [1, fig. 4, p. 17].

Como se observa en la Figura 1.2, el consumo eléctrico de los centros de datos en el mundo ha ido aumentando de forma exponencial, lo que plantea un desafío significativo para la sostenibilidad del crecimiento tecnológico [1]. En el peor escenario, esta tendencia podría llevar a un incremento insostenible en la huella de carbono del sector tecnológico, mientras que en el mejor de los casos, la adopción de tecnologías eficientes podría moderar el crecimiento. Este aumento del consumo energético no solo afecta a los centros de datos, sino también a los dispositivos embebidos y móviles, donde la eficiencia energética es crucial para prolongar la vida útil de las baterías y reducir el impacto ambiental.

Para enfrentar estos desafíos, se han desarrollado diversas técnicas de optimización y compresión que reducen el tamaño y la complejidad de los modelos neuronales manteniendo su rendimiento. Paralelamente, han surgido arquitecturas hardware específicamente diseñadas para la inferencia de modelos de aprendizaje profundo en entornos con restricciones energéticas. En este contexto, los dispositivos de la serie Jetson de NVIDIA[31] destacan por ofrecer un equilibrio entre alto rendimiento en tareas de IA y un consumo energético contenido, ideal para aplicaciones embebidas de visión artificial.

La combinación de CNNs y aceleradores hardware ha permitido la creación de sistemas de visión artificial capaces de detectar y clasificar objetos en movimiento, aplicaciones esenciales en campos como la vigilancia, la conducción autónoma y la robótica. En este contexto, el presente trabajo se centra en el desarrollo de un sistema de visión artificial para detectar y clasificar objetos con posibles defectos en movimiento, utilizando CNNs y aceleradores hardware de bajo consumo.

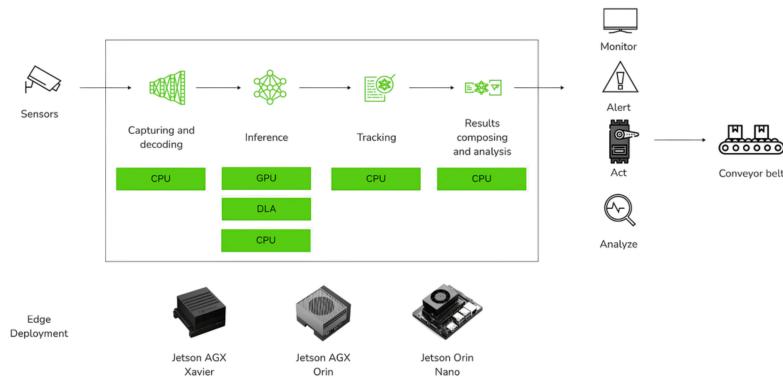


Figura 1.3: Esquema del sistema de visión artificial propuesto.

La Figura 1.3 ilustra el esquema del sistema de visión artificial propuesto. Este sistema se basa en la captura de imágenes de objetos en movimiento, que son procesadas por un modelo de CNN para detectar y clasificar posibles defectos. El modelo se ejecuta en un dispositivo NVIDIA Jetson, optimizado para ofrecer un rendimiento eficiente en términos de velocidad y consumo energético. La implementación del sistema incluye la captura y etiquetado de imágenes, el diseño y entrenamiento del modelo CNN, así como la integración con el hardware NVIDIA para su despliegue en entornos industriales.

1.1 Motivación

Los humanos somos capaces de ver y entender el mundo que nos rodea. Dada una imagen, podemos identificar objetos, reconocer patrones y tomar decisiones basadas en la información visual. Sin embargo, esta capacidad no es innata en las máquinas. La visión por computador es la ciencia que busca dotar a las máquinas de la capacidad de interpretar y comprender imágenes y videos, emulando la forma en que los humanos percibimos el entorno.

Como se mencionó anteriormente, la IA ha revolucionado la forma en que interactuamos con la tecnología. Se ha convertido en una herramienta esencial para aplicar soluciones innovadoras en una amplia gama de campos. En particular, la visión por computador ha demostrado ser un área de gran potencial. También la existencia de dispositivos de bajo consumo, como los de la serie Jetson de NVIDIA, ha permitido llevar la IA a entornos de *edge computing* (cómputo en el borde), donde se acerca el procesamiento de datos a la fuente de información. Esto reduce la latencia y el consumo energético. Con todo esto, se abre un abanico de posibilidades para la implementación de sistemas de visión artificial en aplicaciones industriales.

Centrándose en el ámbito industrial, la detección y clasificación de objetos en movimiento es crucial para optimizar procesos, mejorar la seguridad y aumentar la eficiencia. En la mayoría de los entornos productivos, la detección de defectos se realiza de forma manual, lo que puede ser ineficiente y propenso a errores. La automatización de este proceso mediante sistemas de visión artificial puede reducir costes, aumentar la precisión y mejorar la calidad del producto final.

La motivación de este trabajo radica en la necesidad de desarrollar un sistema de visión artificial capaz de detectar y clasificar objetos en movimiento en un entorno industrial, específicamente en una cinta transportadora.

1.2 Objetivos

El objetivo principal de este trabajo es desarrollar un sistema de visión artificial capaz de detectar y clasificar objetos en movimiento en una cinta transportadora utilizando CNNs y aceleradores hardware de bajo consumo. Para lograr este objetivo, se plantean los siguientes objetivos específicos:

- Realizar un estudio del estado del arte en CNNs, aceleradores hardware de bajo consumo y técnicas avanzadas de optimización para visión artificial.
- Desarrollar un conjunto de datos para el entrenamiento y evaluación del sistema, mediante la captura y etiquetado de imágenes de objetos en movimiento.
- Entrenar y validar diferentes modelos de redes neuronales convolucionales optimizados para la detección y clasificación en tiempo real de defectos en objetos en movimiento.
- Implementar un sistema completo de visión artificial que integre los modelos entrenados con los aceleradores hardware NVIDIA, enfocado en maximizar la eficiencia y minimizar la latencia.
- Analizar los cuellos de botella del sistema, y aplicar técnicas específicas de optimización para mejorar el rendimiento y la eficiencia energética.
- Cuantificar de manera exhaustiva el rendimiento del sistema mediante métricas precisas de exactitud (mean Average Precision (mAP), precisión, recall), latencia (Frames Por Segundo (FPS)) y consumo energético (W, J/inferencia).
- Realizar un análisis comparativo sistemático entre diferentes configuraciones de hardware, software y parámetros de optimización para identificar la combinación que ofrezca el mejor equilibrio entre precisión, velocidad y eficiencia energética.

1.3 Estructura de la memoria

La memoria se estructura en seis capítulos, cada uno dedicado a un aspecto fundamental del trabajo desarrollado:

El **Capítulo 1** introduce el proyecto, detallando la motivación subyacente, los objetivos específicos que se persiguen, la organización general de esta memoria y las colaboraciones para el desarrollo del trabajo.

El **Capítulo 2** sienta las bases teóricas del trabajo. Comienza explorando los fundamentos de la IA y avanza hacia los desarrollos más recientes en CNNs aplicadas a la visión por computador. A continuación, se justifica el uso de aceleradores hardware para tareas de IA, presentando los dispositivos de bajo consumo NVIDIA Jetson. El capítulo concluye con una descripción de los algoritmos de seguimiento de objetos en tiempo real, detallando el funcionamiento de ByteTrack y las métricas de evaluación pertinentes.

El **Capítulo 3** detalla el proceso de diseño e implementación del sistema de visión artificial. Se inicia con un análisis del problema, seguido de la descripción de la recolección de imágenes, su etiquetado y el entrenamiento de los modelos de detección. Concluye con el diseño modular del sistema, especificando sus etapas y las estrategias de segmentación consideradas para su implementación.

El **Capítulo 4** presenta la metodología empleada para evaluar el rendimiento del sistema. Define las métricas clave, describe la configuración de los experimentos y la recolección de datos. Posteriormente, se analizan exhaustivamente los resultados obtenidos, comparando el comportamiento del sistema bajo diversas configuraciones de hardware y software, y evaluando su precisión, velocidad y eficiencia energética. Por último, se analizan las métricas de seguimiento de objetos para evaluar la efectividad del sistema en la detección y seguimiento de objetos en movimiento.

El **Capítulo 5** describe la construcción de un prototipo de cinta transportadora, diseñado para validar el sistema de visión artificial en un entorno que simula condiciones de producción. Se presentan los resultados obtenidos durante esta prueba de concepto.

Finalmente, el **Capítulo 6** resume las principales conclusiones del trabajo, destacando los logros, las limitaciones identificadas, la conexión con los estudios realizados y las posibles líneas de investigación y desarrollo futuras.

1.4 Colaboraciones

El presente Trabajo de Fin de Grado se ha desarrollado en el marco de una beca de colaboración y un periodo de prácticas de empresa, ambos focalizados en la investigación y desarrollo que sustenta este proyecto. Estas actividades se han llevado a cabo en el Grupo de Arquitecturas Paralelas (GAP) de la Universidad Politécnica de Valencia (UPV). El GAP es un grupo de investigación dentro de la universidad, que se dedica al diseño y evaluación de redes de interconexión para sistemas de computación paralela de altas prestaciones, tales como supercomputadores, clústeres y centros de datos.

El Grupo de Arquitecturas Paralelas (GAP) de la Universidad Politécnica de Valencia cuenta con una amplia experiencia en el trabajo sobre redes de interconexión para computadoras paralelas, abarcando desde grandes supercomputadoras, pasando por clústeres de ordenadores personales, usualmente utilizados como servidores, hasta las redes en chip en procesadores multinúcleo. El GAP también ha desarrollado investigación en temas relacionados, como la microarquitectura de procesadores y los protocolos de coherencia de caché. Su investigación abarca también áreas como la optimización de arquitecturas para aplicaciones específicas y la mejora de la eficiencia energética en sistemas computacionales.

Esta colaboración se encuentra estrechamente vinculada con los contenidos y objetivos de la asignatura Sistemas Basados en Deep Learning para la Industria (SDL), una asignatura que forma parte de la mención de Ingeniería de Computadores dentro del Grado en Ingeniería Informática. La oportunidad de colaborar con el GAP ha permitido aplicar de manera práctica y en un contexto de investigación real los conocimientos teóricos y técnicos adquiridos durante la asignatura. Específicamente, ha facilitado la profundización en el diseño, implementación y optimización de sistemas de visión por computador basados en IA, enfrentando desafíos reales relacionados con el rendimiento, la eficiencia y el despliegue en hardware especializado. Esta sinergia entre la formación académica y la investigación aplicada ha enriquecido significativamente la experiencia de aprendizaje, fomentando el desarrollo de habilidades prácticas avanzadas y una comprensión más profunda de las complejidades inherentes al campo de la visión artificial y el aprendizaje profundo en entornos industriales.

CAPÍTULO 2

Conceptos previos

En este capítulo se describirán los conceptos previos que constituyen la base teórica y técnica de este trabajo. Primero, se examinarán los fundamentos en IA centrándose en las CNNs, desde sus bases hasta los modelos más recientes en detección de objetos. A continuación, se analizarán los aceleradores hardware de bajo consumo, con especial énfasis en la arquitectura y capacidades de los dispositivos NVIDIA Jetson. Finalmente, se estudiarán los algoritmos de seguimiento de objetos en tiempo real, fundamentales para aplicaciones con elementos en movimiento. Este marco teórico permitirá contextualizar adecuadamente la solución propuesta para la detección de defectos en objetos en movimiento.

2.1 Fundamentos y avances en redes neuronales para visión artificial

En esta sección se realizará un estudio de las redes neuronales profundas hasta las CNNs, desde sus fundamentos hasta los modelos más recientes en detección de objetos. Se explicarán los conceptos básicos de las redes neuronales y la evolución de las arquitecturas.

2.1.1. Fundamentos de la inteligencia artificial

La **Inteligencia Artificial** es un campo de estudio que busca desarrollar sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el reconocimiento de voz, la toma de decisiones y la comprensión del lenguaje natural. Dentro de este campo, existen diversas subdisciplinas, entre las cuales destacan el *Machine Learning* y el *Deep Learning*.

El **Machine Learning** o Machine Learning (ML) es una rama de la IA que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender de los datos y realizar predicciones o tomar decisiones sin ser programadas explícitamente. Este enfoque se basa en la idea de que las máquinas pueden identificar patrones y relaciones en grandes conjuntos de datos, lo que les permite generalizar y adaptarse a nuevas situaciones.

El **Deep Learning** o aprendizaje profundo es una rama del ML que utiliza redes neuronales artificiales con múltiples capas para modelar y resolver problemas complejos. Este enfoque permite aprender representaciones jerárquicas de los datos, donde cada capa extrae características cada vez más abstractas.

Una de las arquitecturas fundamentales es el Multilayer Perceptron (MLP) o perceptrón multicapa, que consiste en una red de neuronas artificiales organizadas en al menos tres capas: una de entrada, una o más capas ocultas y una capa de salida, como se muestra en la Figura 2.1. En un MLP, cada neurona recibe un conjunto de entradas ponderadas por pesos, aplica una función de activación no lineal a la suma de estas entradas ponderadas, y produce una salida que se transmite a la siguiente capa. Esta estructura permite al Deep Learning abordar tareas complejas en visión por computador, procesamiento del lenguaje natural y otros dominios con un alto grado de precisión.

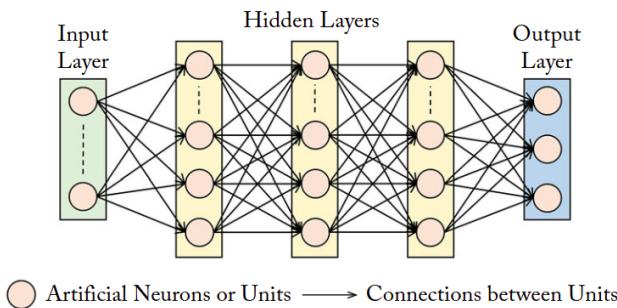


Figura 2.1: Estructura de un perceptrón multicapa (MLP). Extraído de [19, fig. 3.1, p. 32].

2.1.2. Tareas fundamentales en visión por computador

La visión por computador aborda el desafío de permitir que las máquinas interpreten y comprendan el contenido visual de imágenes y videos. Antes del auge del aprendizaje profundo, se empleaban descriptores de características diseñados manualmente, como el Histogram of Oriented Gradients (HOG). HOG captura la forma local de los objetos analizando la distribución de las orientaciones de los gradientes en pequeñas regiones de la imagen (celdas y bloques). La Figura 2.2 muestra una visualización de las características HOG extraídas.



Figura 2.2: Ejemplo de HOG aplicado a una imagen. Extraído de [27].

Aunque útiles en su momento, estos algoritmos “hechos a mano” presentan limitaciones significativas. En particular, no facilitan el aprendizaje por transferencia, es decir, la reutilización de conocimiento aprendido en tareas previas. Además, la complejidad de estas características está intrínsecamente limitada por la capacidad humana para diseñarlas explícitamente.

Estos inconvenientes son superados por los algoritmos de ML de características, como las CNN, que aprenden representaciones relevantes directamente de los datos. Por ello, las CNN han demostrado ser más efectivas en tareas complejas, logrando avances signifi-

ficitivos en precisión y eficiencia al aprender automáticamente características a partir de grandes conjuntos de datos.

En el ámbito del procesamiento de imágenes mediante técnicas de deep learning, existen diversas tareas con diferentes niveles de complejidad:

1. **Clasificación de imágenes:** Es la tarea más básica, donde la red neuronal asigna una etiqueta a toda la imagen. Por ejemplo, determinar si una imagen contiene un perro, gato o coche. El modelo genera un vector de probabilidades para cada clase posible.
2. **Clasificación con localización:** Además de clasificar el objeto principal, la red también proporciona un cuadro delimitador (bounding box) que indica dónde se encuentra ese objeto en la imagen. Es útil cuando existe un único objeto de interés.
3. **Detección de objetos:** Extiende la tarea anterior para identificar y localizar múltiples objetos en una imagen. Los algoritmos de detección se dividen principalmente en:
 - *Detectores de dos etapas:* Como R-CNN, Fast R-CNN y Faster R-CNN, primero generan propuestas de regiones que podrían contener objetos, y luego clasifican estas regiones. Son más precisos pero computacionalmente más costosos.
 - *Detectores de una etapa:* Como You Only Look Once (YOLO) y Single Shot MultiBox Detector (SSD) que predicen las cajas delimitadoras y las clases directamente en una sola pasada. Son más rápidos aunque tradicionalmente menos precisos.

Ambos enfoques proporcionan para cada objeto detectado su clasificación y cuadro delimitador.

4. **Segmentación:** Es la tarea más compleja, donde la red no solo identifica y localiza objetos, sino que también asigna una etiqueta a cada píxel de la imagen. Esto permite distinguir entre diferentes objetos y sus contornos, facilitando una comprensión más detallada de la escena.

La Figura 2.3 ilustra estas tareas fundamentales en visión por computador. Para este trabajo, nos centraremos en la tarea de detección de objetos, que es esencial para identificar y clasificar varios objetos en movimiento en un vídeo o imagen.

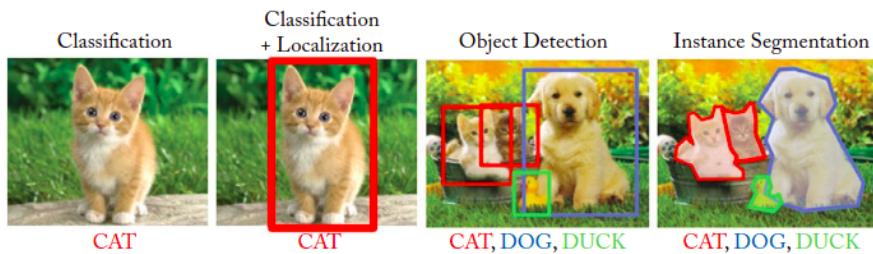


Figura 2.3: Tareas fundamentales en visión por computador. Extraído de [19, fig. 1.1, p. 2].

2.1.3. Arquitectura y funcionamiento de las CNN

Las CNN son un tipo específico de red neuronal profunda. Estas redes están diseñadas para procesar imágenes y extraer características relevantes de manera eficiente, lo que las hace especialmente adecuadas para tareas de visión por computador.

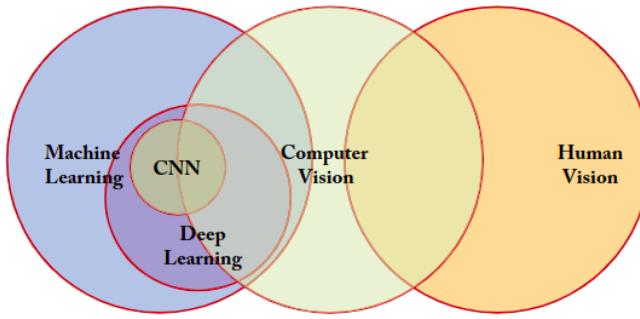


Figura 2.4: Relación entre Machine Learning, Deep Learning, CNN, Computer Vision y Human Vision. Extraído de [19, fig. 1.3, p. 7].

La Figura 2.4 ilustra la relación entre estos conceptos. Las CNN son una subcategoría del Deep Learning, que a su vez es una subcategoría del Machine Learning. Además, las CNN están estrechamente relacionadas con la visión por computador, que busca emular la capacidad de los humanos para interpretar imágenes y videos.

Las CNN se inspiran en la forma en que los humanos percibimos el mundo visual. Al igual que nuestro sistema visual, que procesa la información de manera jerárquica, las CNN utilizan capas convolucionales para extraer características de bajo nivel (como bordes y texturas) y capas más profundas para identificar patrones y objetos más complejos. Esta jerarquía de características permite a las CNN aprender representaciones ricas y abstractas de los datos visuales.

Estas redes se componen de varias capas, cada una de las cuales realiza operaciones específicas en los datos de entrada. Las capas más comunes y técnicas asociadas en una CNN, que se describen a continuación, son las capas convolucionales, las capas de activación, las capas de pooling, la normalización por lotes y las capas completamente conectadas.

Las **capas convolucionales** utilizan la operación de convolución para extraer características. Esta operación es fundamental en las CNN y consiste en aplicar un filtro (también llamado *kernel*) a una imagen para extraer características locales. El filtro se desliza sobre la imagen, multiplicando sus valores por los valores de la imagen en cada posición y sumando los resultados. Este proceso genera un mapa de activación que resalta las características relevantes de la imagen.

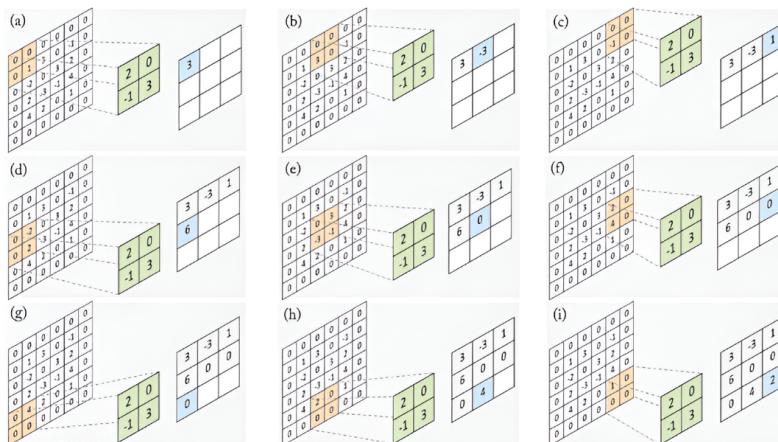
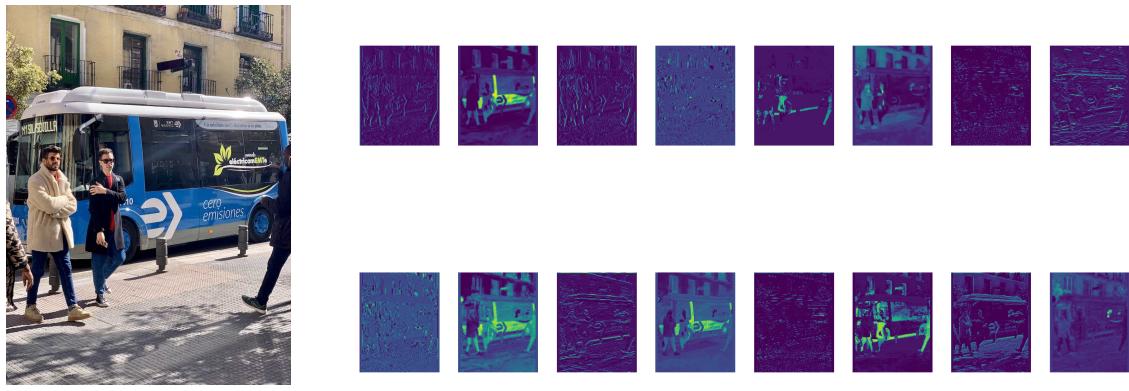


Figura 2.5: Operación de convolución sobre los pixeles de una imagen. Extraído de [19, fig. 2.5, p. 48].

La Figura 2.5 ilustra la operación de una capa de convolución. En este ejemplo, se aplica un filtro de 2×2 (mostrado en verde) a un mapa de características de entrada de 6×6 (incluyendo un relleno de ceros de 1) con un paso (stride) de 2. El filtro se desliza sobre la entrada, y en cada paso, se realiza una multiplicación elemento a elemento entre el filtro y la región correspondiente de la entrada. La suma de estos productos genera un valor en el mapa de características de salida (mostrado en azul).



(a) Imagen de un autobús.

(b) Resultado de la operación de convolución.

Figura 2.6: Proceso de convolución aplicado a una imagen de un autobús.

La Figura 2.6 ilustra el proceso de la primera convolución del modelo YOLO11n [14]. En la parte izquierda se muestra la imagen original de un autobús, mientras que en la parte derecha se presenta el resultado de aplicar la operación de convolución. En este caso, los 16 filtros de la primera capa convolucional han detectado diferentes características de la imagen, como bordes y texturas. Este proceso se repite en múltiples capas, lo que permite a la red aprender representaciones cada vez más complejas de la imagen.

Las **capas de activación** son fundamentales para introducir no-linealidad en el modelo. Estas capas aplican una función no lineal a la salida de cada neurona, lo que permite a la red aprender relaciones complejas entre los datos de entrada y salida. Sin las capas de activación, la red neuronal se comportaría como un modelo lineal, lo que limitaría su capacidad para resolver problemas complejos. Estas capas se aplican típicamente después de cada capa convolucional y completamente conectada.

Existen varias funciones de activación comunes, incluyendo:

- **Sigmoide:** Esta función mapea cualquier valor de entrada a un valor entre 0 y 1. Es útil para problemas de clasificación binaria, donde la salida representa la probabilidad de pertenecer a una clase. Sin embargo, sufre del problema de desvanecimiento del gradiente, lo que dificulta el entrenamiento de redes profundas.
- **Tangente Hiperbólica (tanh):** Similar a la sigmoide, pero mapea los valores de entrada a un rango entre -1 y 1. También sufre del problema de desvanecimiento del gradiente.
- **ReLU (Rectified Linear Unit):** Esta función transforma cada valor negativo en cero mientras mantiene los valores positivos sin cambios, lo que ayuda a mitigar el problema del desvanecimiento del gradiente y acelera el proceso de entrenamiento. Se define como:

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } x \geq 0 \end{cases} \quad (2.1)$$

- **Leaky ReLU:** Similar a ReLU, pero permite un pequeño gradiente cuando la entrada es negativa. Esto ayuda a evitar el problema de “neuronas muertas” que pueden ocurrir con ReLU cuando una neurona deja de aprender completamente.
- **ELU (Exponential Linear Unit):** Similar a ReLU, pero utiliza una función exponencial para valores negativos. Esto permite a la red aprender representaciones más robustas.

Estas capas se aplican típicamente después de cada capa convolucional y completamente conectada.

Las CNN incorporan **capas de pooling** (también llamadas submuestreo). Estas capas desempeñan un papel crucial en la reducción progresiva de la dimensión espacial (ancho y alto) de los mapas de características, lo que conlleva varios beneficios importantes: disminuyen la cantidad de parámetros y la carga computacional, ayudan a controlar el sobreajuste (*overfitting*) al reducir la complejidad del modelo, y proporcionan un cierto grado de invarianza a pequeñas traslaciones y distorsiones.

El funcionamiento del pooling implica deslizar una ventana sobre el mapa de características de entrada y aplicar una operación de agregación. Las operaciones más comunes son Max-Pooling, que selecciona el valor máximo dentro de la ventana (eficaz para capturar las características más prominentes, como se ilustra en la Figura 2.7), y Average-Pooling, que calcula el valor promedio (tiende a suavizar las características). El resultado es un mapa de características de menor tamaño pero que conserva la información esencial.

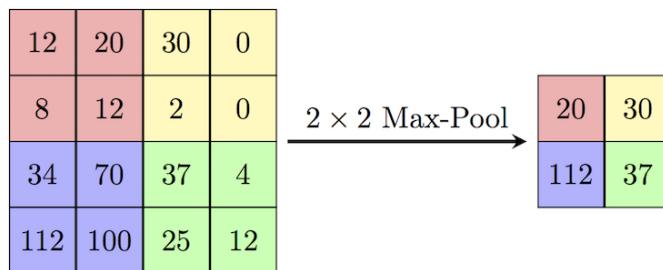


Figura 2.7: Ejemplo de operación de max-pooling con una ventana de 2×2 y un *stride* de 2. Se selecciona el valor máximo de cada región de color. Extraído de [4].

La **Normalización por Lotes** (*Batch Normalization*, BN) es una técnica fundamental para mitigar el problema del *cambio interno de covariables* (*internal covariate shift*), que es la alteración de la distribución de las activaciones intermedias durante el entrenamiento. BN opera a nivel de mini-lotes $\mathcal{B} = \{x_1, \dots, x_m\}$, calculando la media $\mu_{\mathcal{B}}$ y la varianza $\sigma_{\mathcal{B}}^2$:

$$\mu_{\mathcal{B}} = \frac{1}{m} \sum_{i=1}^m x_i \quad (2.2)$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad (2.3)$$

Luego, normaliza cada activación x_i :

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad (2.4)$$

donde ϵ es una constante pequeña para estabilidad numérica. Para preservar la capacidad expresiva, BN introduce parámetros aprendibles γ (escala) y β (desplazamiento) para realizar una transformación afín:

$$y_i = \gamma \hat{x}_i + \beta \quad (2.5)$$

La salida y_i se propaga a la siguiente operación. BN estabiliza y acelera el entrenamiento, permite tasas de aprendizaje más altas y tiene un efecto regularizador que ayuda a prevenir el sobreajuste. Se inserta típicamente después de capas convolucionales o totalmente conectadas, antes de la activación.

En el final de la red, se utilizan **capas completamente conectadas** (*fully connected*). Estas capas toman las características de alto nivel extraídas por las capas anteriores y las combinan para realizar la tarea final, como la clasificación de objetos. Cada neurona en una capa completamente conectada está conectada a todas las neuronas de la capa anterior. En la salida final, se utiliza una función de activación como Softmax para convertir las salidas en probabilidades de clase. La función Softmax se define como:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (2.6)$$

donde $z = (z_1, \dots, z_K)$ es el vector de salidas de la capa anterior para K clases. La función Softmax transforma este vector en un vector de probabilidades $\sigma(z) = (\sigma(z_1), \dots, \sigma(z_K))$, donde cada componente $\sigma(z_i)$ se calcula como se muestra. El resultado es un vector de tamaño K donde cada elemento está en el rango $[0, 1]$ y la suma de todos los elementos es igual a 1 ($\sum_{j=1}^K \sigma(z_j) = 1$). Esto permite interpretar la salida como una distribución de probabilidad sobre las K posibles clases.

En la Figura 2.8 se muestra un ejemplo de una CNN simple. Esta red incluye capas convolucionales para la extracción de características, capas de activación para introducir no linealidad, capas de pooling para reducir la dimensionalidad y capas completamente conectadas para realizar la clasificación final. La combinación de estas capas permite a las CNN aprender representaciones jerárquicas y complejas de los datos visuales, lo que las convierte en una herramienta poderosa para tareas de visión por computador.

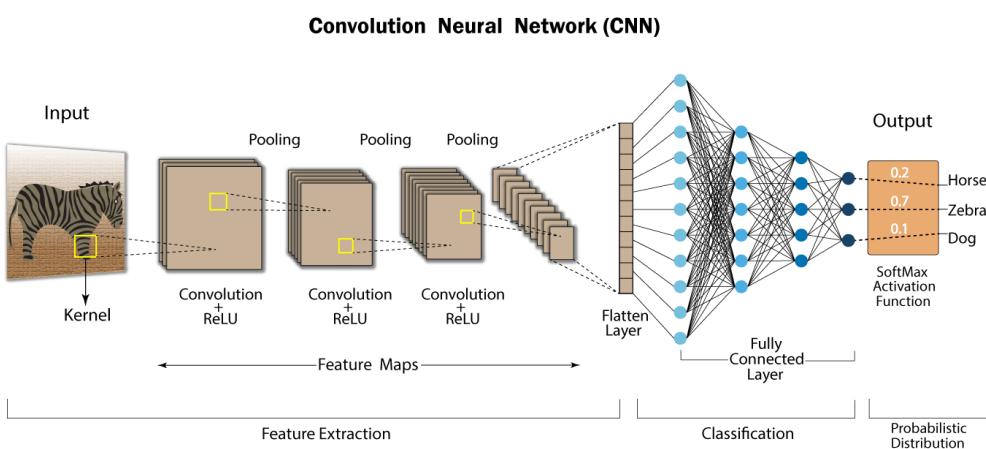


Figura 2.8: Ejemplo de una CNN simple. Extraído de [18].

2.1.4. Entrenamiento de las CNN

Tras comprender la arquitectura interna de las CNN, el siguiente paso es el entrenamiento. Este proceso se basa fundamentalmente en el uso de un conjunto de datos

etiquetado, que actúa como la verdad fundamental (*ground truth*). Este conjunto contiene una colección de imágenes junto con sus correspondientes etiquetas o anotaciones, que la red utilizará para aprender.

El objetivo principal del entrenamiento es ajustar los parámetros internos de la CNN (pesos y sesgos) para minimizar una función de pérdida predefinida. Esta función mide la discrepancia o el error entre las predicciones generadas por la red y las etiquetas reales proporcionadas en los datos de entrenamiento.

El proceso de ajuste se realiza de forma iterativa mediante un algoritmo de optimización. El Descenso de Gradiente Estocástico (SGD) y sus variantes, como Adam o RMSprop, son opciones comunes. Estos algoritmos utilizan el cálculo de gradientes para determinar cómo modificar los pesos y sesgos para reducir el error.

El mecanismo central de este ajuste implica dos fases: la propagación hacia adelante (*forward propagation*) y la retropropagación (*backpropagation*). Durante la propagación hacia adelante, los datos de entrada atraviesan las capas de la red para generar una salida. Esta salida se compara con la etiqueta real mediante la función de pérdida. En la retropropagación, el algoritmo calcula el gradiente del error con respecto a los pesos y sesgos, utilizando la regla de la cadena. Estos gradientes guían la actualización de los parámetros.

Típicamente, todo el conjunto de datos de entrenamiento se procesa varias veces en ciclos conocidos como épocas. Este refinamiento iterativo continúa hasta que el rendimiento de la red, evaluado en un conjunto de datos de validación separado (que no se utiliza para el entrenamiento directo), alcanza un nivel satisfactorio. La validación es crucial para asegurar que el modelo generaliza bien a datos no vistos previamente.

Además, para prevenir el sobreajuste (*overfitting*) —donde el modelo aprende demasiado bien los datos de entrenamiento pero falla en generalizar— se emplean técnicas de regularización. El *dropout* es una técnica común que desactiva aleatoriamente un porcentaje de neuronas durante cada iteración de entrenamiento, forzando a la red a aprender representaciones más robustas. La normalización por lotes (*batch normalization*), explicada previamente, también actúa como regularizador y ayuda a estabilizar el entrenamiento. Otra técnica utilizada es el *early stopping*, observado en la Figura 2.9, que implica detener el entrenamiento cuando el rendimiento en el conjunto de validación comienza a deteriorarse, evitando así el sobreajuste.

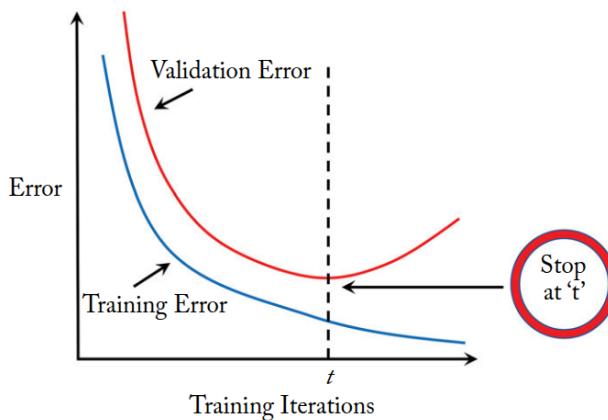


Figura 2.9: Gráfico de entrenamiento con early stopping. Extraído de [19, fig. 5.2, p. 79]

En escenarios prácticos, desarrollar y entrenar una CNN desde cero puede ser computacionalmente costoso y requerir grandes cantidades de datos etiquetados.

Por lo tanto, una estrategia común y muy eficaz consiste en aprovechar modelos preentrenados. Arquitecturas como VGG16, ResNet50 y MobileNetV2, que han sido entrenadas previamente en conjuntos de datos de referencia masivos como ImageNet[7] (que contiene millones de imágenes en miles de categorías) o COCO[23], sirven como puntos de partida potentes. Estos modelos ya han aprendido características jerárquicas ricas a partir de datos visuales diversos.

Mediante una técnica conocida como *transfer learning* o aprendizaje por transferencia, estos modelos preentrenados pueden adaptarse eficientemente a tareas nuevas y específicas, incluso con conjuntos de datos personalizados más pequeños. El aprendizaje por transferencia generalmente implica tomar las capas de extracción de características del modelo preentrenado (la base convolucional) y ajustarlas (*fine-tuning*) o añadir nuevas capas de clasificación adaptadas a la tarea objetivo.

Este enfoque acelera significativamente el proceso de entrenamiento para la nueva tarea y, a menudo, conduce a un mejor rendimiento en comparación con el entrenamiento desde cero, ya que transfiere eficazmente el conocimiento visual general adquirido durante el entrenamiento inicial a gran escala.

2.1.5. Detectores de dos etapas

Los detectores de dos etapas, funcionan mediante un proceso secuencial: primero generan propuestas de regiones (region proposals) que podrían contener objetos y posteriormente clasifican estas regiones. Este enfoque favorece la precisión, aunque generalmente a costa de un mayor tiempo de procesamiento.

La primera arquitectura exitosa de detección de objetos basada en deep learning fue R-CNN (Regions with CNN features) [10]. Este modelo introdujo un enfoque de dos etapas que revolucionó el campo. En su primera fase, R-CNN utiliza un algoritmo de búsqueda selectiva (*Selective Search*[40]) para generar aproximadamente 2,000 propuestas de regiones que podrían contener objetos. Este algoritmo de búsqueda selectiva divide la imagen en nodos y aristas, e iterativamente agrupa estas regiones en función del color, textura, tamaño y forma hasta que se obtienen las propuestas finales. En la figura 2.10 se muestra un ejemplo del resultado del algoritmo de búsqueda selectiva.



(a) Imagen original.



(b) Resultado de búsqueda selectiva.

Figura 2.10: Proceso de búsqueda selectiva aplicado a una imagen. Extraído de [22].

En la segunda fase, cada región propuesta es redimensionada y procesada individualmente por una CNN para extraer características de alto nivel. Estas características alimentan posteriormente a un clasificador SVM (Support Vector Machine) para determinar la categoría del objeto y a un regresor lineal para mejorar la localización del cuadro delimitador. Como se ilustra en la Figura 2.11, este enfoque fue innovador pero compu-

tacionalmente costoso, ya que requiere procesar cada propuesta de región de manera independiente.

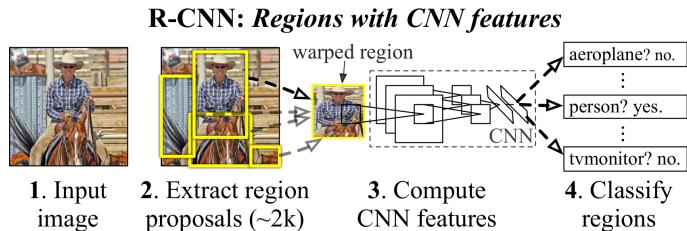


Figura 2.11: Arquitectura de R-CNN. Extraído de [10, fig. 1, p. 1].

2.1.6. Detectores de una etapa

En contraste con los detectores de dos etapas, los detectores de una etapa (one-stage detectors) adoptan un enfoque más directo y eficiente. Estos detectores realizan la localización y clasificación de objetos simultáneamente en una sola pasada a través de la red, sin necesidad de un paso intermedio de generación de propuestas.

La arquitectura de los detectores de una etapa procesa la imagen completa una única vez, típicamente mediante una red troncal o *backbone* (generalmente una CNN) para la extracción de características. Estas características son posteriormente procesadas por componentes intermedios (*neck*) y alimentadas a una cabeza de detección (*detection head*) que predice simultáneamente las coordenadas de los cuadros delimitadores (*bounding boxes*) y las probabilidades de clase.

Esta arquitectura de una etapa prioriza la velocidad de inferencia, resultando idónea para aplicaciones en tiempo real donde la latencia es un factor crítico, aunque pueda suponer una ligera concesión en la precisión máxima. Modelos representativos de este enfoque incluyen SSD (Single Shot MultiBox Detector)[24] y YOLO (You Only Look Once)[38]. Estos han demostrado un equilibrio eficaz entre rapidez y exactitud, permitiendo la detección en tiempo real incluso en dispositivos con recursos computacionales limitados.

YOLO se destaca como una de las arquitecturas de detección de objetos más populares y efectivas. Concebida específicamente para la detección en tiempo real, YOLO introdujo un enfoque unificado que procesa la imagen completa en una sola pasada, realizando la localización y clasificación de forma simultánea. Esta metodología ha sido fundamental para su adopción en aplicaciones que requieren alta velocidad de procesamiento.

YOLO en su primera versión procesa la imagen completa de una vez. Divide la imagen en una cuadrícula de $S \times S$ celdas. Cada celda es responsable de detectar los objetos cuyo centro se encuentre dentro de ella. Para cada celda, YOLO predice B cuadros delimitadores (*bounding boxes*) y puntuaciones de confianza para esos cuadros. La puntuación de confianza indica la probabilidad de que haya un objeto en el cuadro y la precisión de la predicción del cuadro. Al mismo tiempo, predice las probabilidades de clase para cada objeto detectado en la celda.

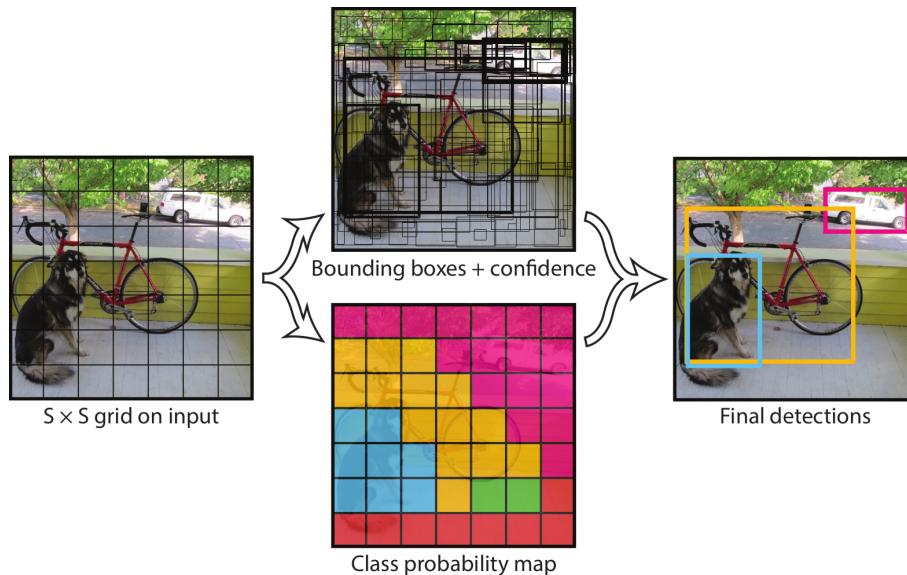


Figura 2.12: Ejemplo de detección de objetos utilizando YOLO. Extraído de [38, fig. 2, p. 2].

Como se ilustra en la Figura 2.12, YOLO modela la detección de objetos como un problema de regresión directa. El modelo divide la imagen en una cuadrícula de $S \times S$ celdas, y cada celda predice B cuadros delimitadores, codificados por sus coordenadas normalizadas (x, y, w, h), una puntuación de confianza por cada caja, y una distribución de probabilidad sobre las C clases posibles. Estas predicciones se encapsulan en un tensor de dimensiones $S \times S \times (B \cdot 5 + C)$, donde el factor 5 corresponde a las coordenadas y la puntuación de confianza de cada caja. Por ejemplo, si se utilizan $B = 2$ y $C = 20$, el tensor de salida tendrá dimensiones $S \times S \times 30$.

Una vez generadas las predicciones, YOLO implementa un post-procesamiento mediante Non-Maximum Suppression (NMS) para eliminar detecciones redundantes y conservar únicamente las más precisas. El algoritmo NMS funciona de la siguiente manera:

- Ordenamiento por Confianza:** Las cajas delimitadoras detectadas se ordenan por su puntuación de confianza, de mayor a menor.
- Selección de la Detección Más Confiable:** Se selecciona la caja delimitadora con la puntuación de confianza más alta y se añade a la lista de detecciones finales.
- Cálculo de la Intersección sobre Unión (IoU):** Se calcula la Intersection over Union (IoU) entre la caja delimitadora seleccionada y todas las demás cajas delimitadoras restantes.
- Eliminación de Detecciones Redundantes:** Se eliminan todas las cajas delimitadoras con una IoU superior a un umbral predefinido con la caja delimitadora seleccionada.
- Iteración:** Se repiten los pasos 2-4 hasta que no queden más cajas delimitadoras por procesar.

En resumen, NMS evalúa la superposición entre las cajas y elimina aquellas que exceden un cierto umbral con la caja de mayor confianza, asegurando que solo las detecciones más precisas y no redundantes se conserven, mejorando la calidad general de las detecciones.

Este enfoque unificado permite que YOLO procese imágenes a velocidades significativamente mayores que los detectores de dos etapas, mientras mantiene una precisión competitiva, lo que lo hace ideal para aplicaciones en tiempo real como las que se abordan en este trabajo.

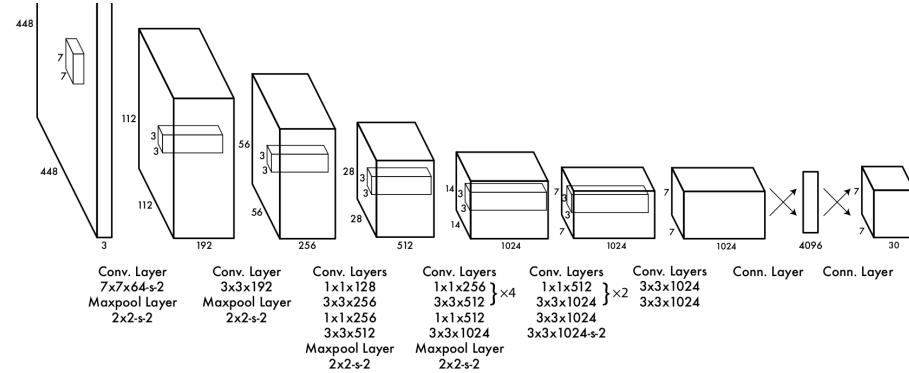


Figura 2.13: Arquitectura de YOLO. Extraído de [38, fig. 3, p. 3].

En la Figura 2.13 se presenta la arquitectura del modelo primigenio de YOLO [38]. Esta arquitectura se basa en una red neuronal convolucional que extrae características de la imagen de entrada y las procesa a través de varias capas para generar las predicciones finales. A lo largo de los años, se han desarrollado múltiples versiones y mejoras de YOLO, cada una optimizando aspectos como la precisión, la velocidad y la capacidad de detección de objetos pequeños o densamente agrupados.

En este proyecto, se han seleccionado diversas versiones de YOLOv5 [12], YOLOv8 [13] y YOLOv11 [14]. Los modelos mencionados forman parte de la librería Ultralytics [15], que proporciona una implementación eficaz y sencilla de varias variantes de YOLO. Las aportaciones de Ultralytics a la comunidad de visión artificial son notables, ya que sus herramientas y recursos facilitan el ciclo de vida completo (entrenamiento, evaluación, implementación) de los modelos YOLO.

2.1.7. Métricas de evaluación de modelos de detección de objetos

Para evaluar el rendimiento de los modelos de detección de objetos, se utilizan métricas específicas que permiten cuantificar la precisión y efectividad del sistema. Entre las métricas más comunes se encuentran:

- **Precisión (Precision):** Mide la proporción de verdaderos positivos (TP) del total de predicciones positivas realizadas por el modelo. Se calcula como:

$$\text{Precisión} = \frac{TP}{TP + FP}$$

donde FP representa los falsos positivos. Una alta precisión indica que el modelo realiza pocas predicciones incorrectas.

- **Exhaustividad (Recall):** Mide la proporción de verdaderos positivos del total de objetos relevantes en la imagen. Se calcula como:

$$\text{Exhaustividad} = \frac{TP}{TP + FN}$$

donde FN representa los falsos negativos. Una alta exhaustividad indica que el modelo es capaz de detectar la mayoría de los objetos relevantes, aunque pueda incluir algunas predicciones incorrectas.

- **IoU (Intersección sobre Unión):** Mide la superposición entre el cuadro delimitador predicho y el cuadro delimitador real. Se calcula como:

$$\text{IoU} = \frac{\text{Área de intersección}}{\text{Área de unión}}$$

Un IoU alto indica que el modelo ha localizado correctamente el objeto. Generalmente, se considera que un IoU superior a 0.5 indica una detección correcta.

- **mAP (mean Average Precision):** El cálculo de mAP se realiza en varios pasos: primero, para cada clase, se calcula la curva PR (*Precision - Recall*) y se obtiene la Precisión Promedio (AP), que representa el área bajo esta curva. Luego, el mAP se obtiene como el promedio de todos los AP de las diferentes clases. En la detección de objetos, el mAP usualmente incorpora diferentes umbrales de IoU (Intersección sobre Unión), expresado como mAP50 o mAP0.5-0.95. Esta métrica es especialmente útil cuando las clases están desbalanceadas, ya que da igual importancia a las clases minoritarias y mayoritarias. Un valor de mAP cercano a 1 indica un modelo con alta precisión y exhaustividad en todas las clases.
- **Latencia:** Mide el tiempo que tarda el modelo en procesar una imagen y generar una predicción, se mide en milisegundos (ms) y un valor bajo indica que el modelo es capaz de realizar inferencias rápidamente.
- **FPS (Fotogramas Por Segundo):** Mide la velocidad de procesamiento del modelo, indicando cuántas imágenes puede procesar por segundo; un valor de FPS alto indica que el modelo es capaz de realizar inferencias rápidamente.
- **Puntuación F1 (F1 Score):** Es la media armónica entre la *Precision* y el *Recall*, proporcionando un equilibrio entre ambas. Se utiliza para evaluar el rendimiento del modelo en situaciones donde existe un desequilibrio entre las clases. Se calcula como:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Un *F1 Score* alto indica que el modelo tiene un buen equilibrio entre la *Precision* y el *Recall*, lo cual es especialmente importante en aplicaciones donde tanto la detección correcta como la minimización de falsos positivos son críticos.

- **Pérdida (Loss):** Es una medida de la discrepancia entre las predicciones del modelo y las etiquetas reales. Durante el entrenamiento, el objetivo es minimizar esta pérdida. Existen diferentes tipos de funciones de pérdida, como la pérdida de clasificación (*Cross-Entropy Loss*) y la pérdida de localización (*Smooth L1 Loss*), que se combinan para evaluar el rendimiento del modelo.

2.2 Aceleradores de procesamiento gráfico

Un aspecto fundamental para el despliegue eficiente de modelos de IA es el hardware utilizado para su ejecución. A continuación, se analizan los principales aspectos de los aceleradores de procesamiento gráfico y su importancia en aplicaciones de visión artificial.

2.2.1. Limitaciones del hardware tradicional

La Dennard Scaling[8], formulada por Robert Dennard en 1974, establecía que a medida que los transistores se reducían de tamaño, su consumo de energía por unidad de área se mantenía constante. Esto significaba que, al reducir el tamaño de los transistores a la mitad, su área se reducía a un cuarto, pero su consumo de energía por unidad de área permanecía igual. Como resultado, el consumo total de energía se reducía a la mitad, permitiendo aumentar la frecuencia de reloj y el número de transistores sin incrementar significativamente el consumo total de energía. Este principio fue fundamental para el avance exponencial en el rendimiento de los procesadores durante décadas.

Sin embargo, a partir de 2005, la Dennard Scaling dejó de cumplirse debido a varios factores físicos fundamentales. A escalas nanométricas, los efectos cuánticos y las fugas de corriente se volvieron significativos, impidiendo que el consumo de energía por unidad de área se mantuviera constante.

Por otro lado, la ley de Moore[28], formulada por Gordon Moore en 1965, predecía que el número de transistores en un chip se duplicaría aproximadamente cada año, predicción que posteriormente se ajustó a cada dos años. Durante décadas, esta ley se cumplió con notable precisión, permitiendo un crecimiento exponencial en la capacidad de procesamiento.

No obstante, a medida que los transistores se acercan a escalas atómicas (actualmente en torno a los 3-5 nanómetros), los límites físicos y los desafíos de fabricación han ralentizado significativamente este ritmo de avance. Los costes de investigación y desarrollo para mantener esta tendencia se han disparado, y los beneficios en términos de rendimiento por transistor se han reducido.

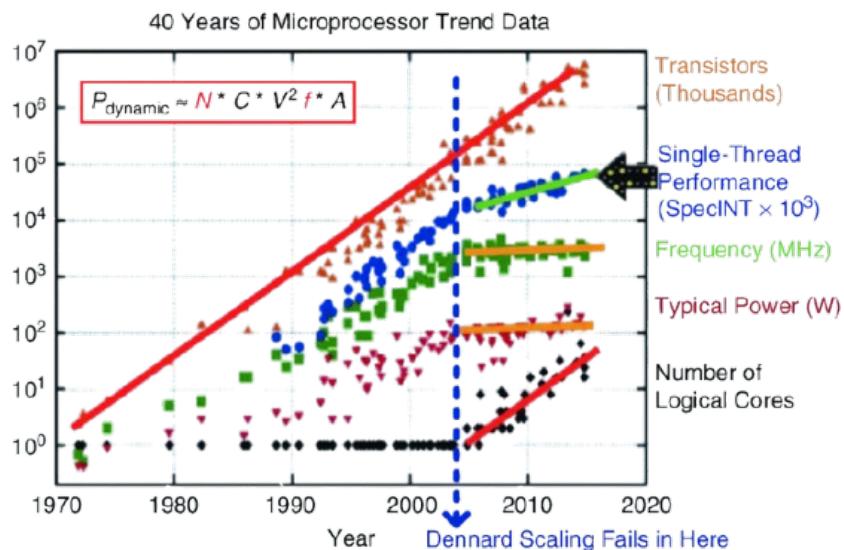


Figura 2.14: Evolución histórica de las características de los microprocesadores (1970–2020). Extraído de [5, p. 8].

En la Figura 2.14 se ilustra claramente el impacto combinado del fin de la Dennard Scaling y el estancamiento de la ley de Moore. La gráfica muestra cinco métricas fundamentales en escala logarítmica: el número de transistores (triángulos naranja) que sigue la ley de Moore, el rendimiento de un solo hilo (círculos azules), la frecuencia de reloj (cuadrados verdes), el consumo de potencia (triángulos invertidos morados) y el número de núcleos lógicos (rombos negros). La ecuación de potencia dinámica ($P_{dynamic} \approx$

$N * C * V^2 * f * A$) explica la relación entre el número de transistores (N), la capacitancia (C), el voltaje (V), la frecuencia (f) y el factor de actividad (A).

El punto de inflexión en 2005, marcado como *Dennard Scaling Fails in Here*, marca el momento en que la industria tuvo que cambiar radicalmente su estrategia. La frecuencia de reloj se estancó en torno a los 3-4 GHz, el rendimiento por núcleo comenzó a crecer más lentamente, y como respuesta, se adoptaron dos estrategias principales: el aumento del número de núcleos y la estabilización del consumo de potencia alrededor de los 100W. Este fenómeno ha llevado a la industria a buscar alternativas como las arquitecturas de dominio específico para continuar mejorando el rendimiento de los sistemas computacionales.

2.2.2. Arquitectura y funcionamiento de las GPUs

Para superar las limitaciones de las CPUs en tareas específicas, se emplean arquitecturas hardware especializadas. Entre ellas, las Field Programmable Gate Arrays (FPGAs) que ofrecen flexibilidad al ser reconfigurables post-fabricación, aunque su programación es compleja. En el extremo de la especialización se encuentran los Application-Specific Integrated Circuits (ASICs), diseñados a medida para una única función, logrando máxima eficiencia a costa de un alto coste de diseño y nula reprogramabilidad; un ejemplo son las TPUs de Google, optimizadas para ML.

Las GPUs, por su parte, son arquitecturas orientadas al paralelismo masivo. Su diseño *many-core*, con miles de núcleos de procesamiento más simples, las hace idóneas para cargas de trabajo intensivas y paralelizables, como las operaciones matriciales del aprendizaje profundo. Dada esta capacidad, las GPUs son fundamentales en este proyecto.

El modelo de programación de las GPUs está basado en la ejecución masiva de hilos, organizados en bloques y rejillas (*blocks* y *grids*), según la terminología de CUDA[30], el modelo de programación desarrollado por NVIDIA[32]. Cada hilo ejecuta la misma función, conocida como *kernel*, pero opera sobre diferentes fragmentos de datos. Este enfoque, conocido como SIMT (Single Instruction, Multiple Threads), permite aprovechar al máximo el paralelismo inherente a muchas aplicaciones de IA, como el entrenamiento y la inferencia de redes neuronales.

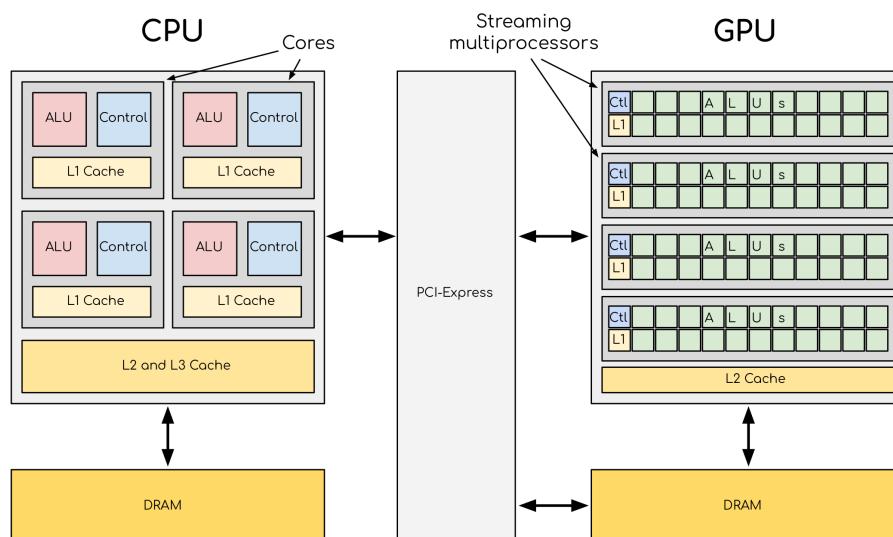


Figura 2.15: Comparativa de arquitecturas CPU y GPU. Extraído de [9].

La figura 2.15 ilustra las diferencias arquitectónicas entre CPU y GPU. Las CPUs constan de un número reducido de núcleos (generalmente entre 4 y 64) optimizados para un alto rendimiento secuencial, donde cada núcleo dispone típicamente de caché y unidad de control privadas. En cambio, las GPUs se basan en una arquitectura *many-core*, integrando miles de núcleos más simples diseñados para el paralelismo masivo, aunque con menor rendimiento individual por núcleo. Estos núcleos de GPUs suelen compartir recursos como la caché y las unidades de control, lo que posibilita empaquetar una mayor cantidad de unidades de procesamiento en el mismo chip y, consecuentemente, alcanzar una mayor densidad de cómputo.

NVIDIA ha sido una figura central en la evolución de la aceleración de IA, realizando contribuciones clave que han modelado el campo. Fue pionera en la computación de propósito general en GPU con la introducción de CUDA en 2006. Esta plataforma permitió utilizar la masiva capacidad de procesamiento paralelo de las GPU para tareas computacionales generales, extendiendo su uso más allá de los gráficos tradicionales. La programación de GPU se realiza utilizando lenguajes y APIs especializadas como CUDA y OpenCL, que otorgan al desarrollador un control explícito sobre la distribución de datos y tareas entre los miles de núcleos disponibles.

El desarrollo eficaz en este paradigma requiere identificar y explotar el paralelismo inherente a los algoritmos, así como gestionar eficientemente la compleja jerarquía de memoria de la GPU. Es crucial considerar la latencia significativamente mayor del acceso a la memoria global en comparación con memorias más rápidas pero de menor capacidad, como la memoria compartida local a un bloque de hilos.

Dada la complejidad de esta programación a bajo nivel, NVIDIA ha desarrollado un ecosistema de software robusto y optimizado para facilitar el desarrollo en IA. Este incluye bibliotecas fundamentales como cuDNN, específica para acelerar primitivas de redes neuronales profundas, y cuBLAS, para operaciones de álgebra lineal básica, ambas esenciales para el rendimiento. Además, es común recurrir a *frameworks* de alto nivel como PyTorch y TensorFlow. Estas librerías abstraen muchos detalles de la implementación en GPU, utilizando las bibliotecas de NVIDIA subyacentes y simplificando enormemente el desarrollo de aplicaciones de IA.

En el frente del hardware, la compañía ha impulsado continuamente la innovación con el desarrollo de componentes especializados como los Tensor Cores, introducidos en la arquitectura Volta. Estos núcleos están diseñados específicamente para acelerar las operaciones matriciales intensivas (como multiplicaciones de matrices mixtas) que son omnipresentes en el entrenamiento e inferencia de modelos de IA. Como resultado de estas continuas innovaciones en hardware y la creación de un ecosistema de software integral, NVIDIA ha logrado establecer estándares de facto para la aceleración de IA, consolidándose como la plataforma preferida en una amplia gama de entornos, desde grandes centros de datos hasta sistemas embebidos con recursos limitados.

2.2.3. Serie Jetson: dispositivos para IA de bajo consumo

La serie Jetson de NVIDIA constituye una familia de módulos computacionales diseñados específicamente para habilitar la IA y el aprendizaje profundo en dispositivos de borde (*edge devices*). Estos sistemas compactos y de bajo consumo son fundamentales para aplicaciones que requieren procesamiento local de datos con alta capacidad de cómputo.

El enfoque principal de la serie Jetson es la computación en el borde (*edge computing*), un paradigma que acerca el procesamiento de datos y la IA a la fuente donde se generan. Esto resulta crucial en aplicaciones donde la latencia, el ancho de banda limitado o la privacidad son factores críticos, ya que evita la necesidad de enviar grandes volúmenes

de datos a la nube para su análisis. Los dispositivos Jetson están optimizados para operar bajo restricciones significativas de energía, tamaño y coste, características típicas de los entornos embebidos y de borde.

Cada módulo Jetson se basa en una arquitectura System-on-Chip (SoC), que integra múltiples componentes de procesamiento en un único circuito integrado. Esto incluye núcleos de CPU basados en la arquitectura ARM y potentes núcleos de GPU NVIDIA con arquitecturas modernas. Además, algunos modelos de gama alta, como los Jetson AGX Orin y AGX Xavier utilizados en este trabajo, incorporan aceleradores de hardware dedicados conocidos como DLAs. Estas unidades especializadas están diseñadas para ejecutar operaciones de inferencia de redes neuronales de manera altamente eficiente en términos de rendimiento y consumo energético, liberando así la GPU y la CPU para otras tareas. Esta integración heterogénea permite una alta eficiencia computacional y energética, reduce la huella física del sistema y simplifica el diseño de la placa portadora, resultando en soluciones más compactas y eficientes.

Un pilar fundamental del diseño de la serie Jetson es la optimización de la eficiencia energética. Estos dispositivos están diseñados para ofrecer un alto rendimiento computacional por cada vatio de energía consumido (TOPS/W), esencial para aplicaciones embebidas con fuentes de alimentación limitadas o restricciones térmicas. La capacidad de configurar diferentes perfiles de energía permite ajustar dinámicamente el equilibrio entre rendimiento y consumo.

Modelo	AI Performance	GPU	GPU Max Freq.	CPU	Memoria	DLAs	Precio (€)
Jetson AGX Orin	275 TOPS	2048-core Ampere, 64 Tensor Cores	1.3 GHz	12-core Cortex-A78AE, 3MB L2 + 6MB L3, 2.2 GHz	64GB LPDDR5, 204.8GB/s	2x NVDA v2	2400
Jetson Orin Nano	67 TOPS	1024-core Ampere, 32 Tensor Cores	1020 MHz	6-core Cortex-A78AE, 1.5MB L2 + 4MB L3, 1.7 GHz	8GB LPDDR5, 102 GB/s	-	300
Jetson AGX Xavier	32 TOPS	512-core Volta, 64 Tensor Cores	1377 MHz	8-core Carmel v8.2, 8MB L2 + 4MB L3, 2.2 GHz	32GB LPDDR4x, 136.5GB/s	2x NVDA v1	1000

Tabla 2.1: Comparativa técnica entre diferentes modelos NVIDIA Jetson.

La tabla 2.1[31] presenta una comparativa técnica entre diferentes modelos de la serie Jetson disponibles para la realización de este trabajo, incluyendo la presencia y tipo de DLAs. Cada modelo está diseñado para satisfacer diferentes necesidades y requisitos de rendimiento, lo que permite a los desarrolladores seleccionar el módulo más adecuado para su aplicación específica. Para este trabajo, se han utilizado los tres modelos de la tabla y se compararán sus resultados en la subsección 4.2.6.

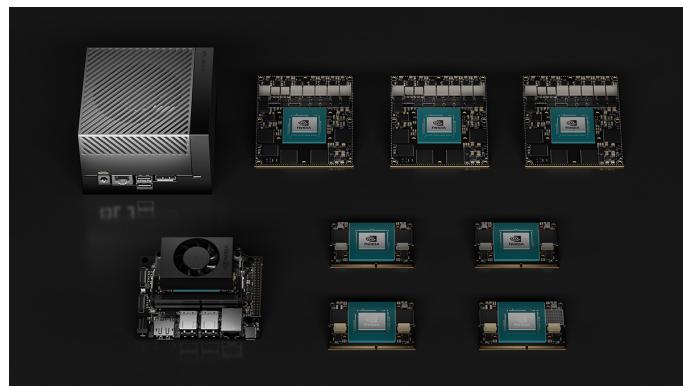


Figura 2.16: Módulos Jetson de NVIDIA

Más allá del hardware, la fortaleza de la plataforma Jetson reside en su completo ecosistema de software, proporcionado a través del SDK NVIDIA JetPack. Esto incluye un sistema operativo Linux optimizado (L4T), controladores, CUDA, cuDNN, TensorRT, herramientas de desarrollo y documentación. Esta plataforma unificada agiliza el desarrollo, desde la creación del modelo hasta la implementación optimizada.

Para este trabajo, todos los programas se ejecutaron utilizando imágenes de contenedores Docker proporcionadas por Ultralytics [41]. Construidas sobre las imágenes base de NVIDIA y optimizadas para la serie Jetson, estas imágenes preconfiguradas simplifican la implementación al incluir todas las dependencias necesarias como CUDA, cuDNN, TensorRT y las bibliotecas de Python requeridas, asegurando la reproducibilidad del entorno.

2.2.4. TensorRT

NVIDIA TensorRT es un kit de desarrollo de software (SDK) integral diseñado específicamente para la optimización de modelos de aprendizaje profundo y la consecución de una inferencia de muy alto rendimiento en la amplia gama de hardware de NVIDIA, desde centros de datos hasta sistemas embebidos como la serie Jetson. Actúa como un potente compilador y motor de ejecución en tiempo real que transforma modelos previamente entrenados en versiones altamente eficientes, optimizadas para el despliegue en producción.

El objetivo principal de TensorRT es cerrar la brecha entre los *frameworks* de entrenamiento (como TensorFlow o PyTorch), que priorizan la flexibilidad y la facilidad de desarrollo, y los requisitos estrictos de las aplicaciones de inferencia en el mundo real, que demandan baja latencia, alto rendimiento (*throughput*) y eficiencia energética. Para lograr esto, TensorRT aplica una serie de optimizaciones sofisticadas durante una fase de compilación offline:

- **Optimización del Grafo Computacional:** TensorRT analiza la estructura del modelo y realiza transformaciones significativas para mejorar la eficiencia. Esto incluye:
 - **Fusión de Capas (Layer Fusion):** Combina múltiples capas secuenciales (fusión vertical) o paralelas (fusión horizontal) en un único kernel optimizado. Por ejemplo, una secuencia de convolución, sesgo (bias) y activación (ReLU) puede fusionarse en una sola operación, reduciendo la sobrecarga de lanzamiento de kernels y el movimiento de datos en memoria.
 - **Eliminación de Capas:** Identifica y elimina capas que no son necesarias para la inferencia, como las capas de dropout.
 - **Fusión de Tensores (Tensor Fusion):** Combina operaciones que acceden a los mismos tensores para mejorar la localidad de los datos.

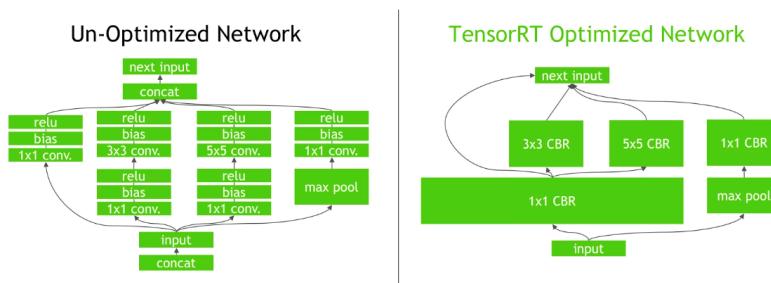


Figura 2.17: Ejemplo de optimización de grafo computacional en TensorRT.

- **Calibración y Cuantización de Precisión:** TensorRT ofrece un soporte robusto para reducir la precisión numérica de los pesos y activaciones del modelo. Puede convertir modelos de precisión completa (FP32) a precisiones más bajas como FP16 (media precisión), INT8 (enteros de 8 bits), o incluso formatos más recientes como FP8 o FP4 en hardware compatible. Esta reducción disminuye drásticamente el tamaño del modelo, el ancho de banda de memoria requerido y acelera el cómputo (especialmente en hardware con soporte nativo como los Tensor Cores), a menudo con una pérdida mínima o nula de precisión.
- **Selección Automática de Kernels (Kernel Auto-Tuning):** Durante la fase de construcción, TensorRT evalúa múltiples implementaciones (kernels) para cada operación en el hardware de destino específico y selecciona la más rápida disponible, considerando factores como el tamaño de los tensores y la precisión utilizada.
- **Gestión Dinámica de Memoria (Dynamic Tensor Memory):** Optimiza la asignación de memoria para los tensores intermedios, reutilizando la memoria siempre que sea posible para minimizar la huella de memoria global.
- **Ejecución Multi-Stream:** Facilita la ejecución concurrente de múltiples flujos de inferencia en la misma GPU, mejorando el rendimiento general del sistema.

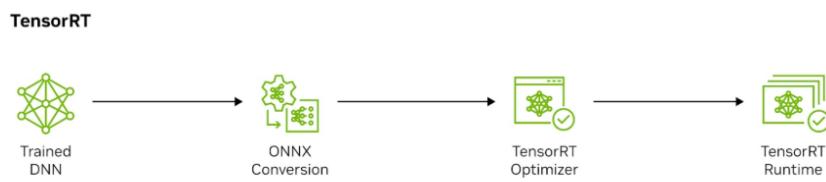


Figura 2.18: Ejemplo de flujo de trabajo de optimización con TensorRT

El proceso típico de uso de TensorRT, como se ilustra esquemáticamente en la Figura 2.18, implica tomar un modelo entrenado (generalmente exportado a un formato intermedio como Open Neural Network Exchange (ONNX)[34]), utilizar el **constructor (builder)** de TensorRT para aplicar las optimizaciones y generar un **motor (engine)** de inferencia serializado y optimizado. Este motor es específico para el modelo, la precisión deseada y la plataforma hardware de destino (p. ej., un Jetson Orin Nano específico). Finalmente, el motor se carga en el **tiempo de ejecución (runtime)** de TensorRT para realizar inferencias rápidas y eficientes.

TensorRT se integra de forma nativa con los principales *frameworks* de aprendizaje profundo (TensorFlow, PyTorch) y soporta el formato de intercambio ONNX, lo que facilita la importación de modelos desde prácticamente cualquier *framework* de entrenamiento. Su capacidad para reducir significativamente la latencia y aumentar el rendimiento lo convierte en una herramienta esencial para desplegar modelos de IA en aplicaciones sensibles al tiempo real y en dispositivos con recursos limitados como los de la serie Jetson.

2.3 Seguimiento de objetos en tiempo real

En esta sección se presentarán los conceptos básicos del seguimiento de objetos en tiempo real, se explicará el problema del Multiple Object Tracking (MOT) y se presentarán los algoritmos más relevantes en este campo.

2.3.1. Introducción al seguimiento de objetos

El seguimiento de objetos es un proceso que complementa la salida de los modelos de detección. Mientras que un modelo de detección opera sobre cada fotograma de forma independiente, identificando y localizando objetos como si fueran imágenes estáticas, el MOT actúa sobre estas detecciones para establecer una correspondencia temporal. La función esencial del MOT es asignar un identificador único a cada objeto detectado en un fotograma y mantener dicho identificador de manera consistente a lo largo de la secuencia de vídeo. Esto permite reconstruir las trayectorias de los objetos y analizar su comportamiento dinámico en la escena.

Para realizar este seguimiento, el MOT se basa en la información temporal y espacial de las detecciones. Utiliza técnicas de predicción y asociación para determinar la continuidad de los objetos a lo largo del tiempo, teniendo en cuenta factores como la posición, velocidad y apariencia de los objetos. El algoritmo utilizado para el seguimiento puede variar en complejidad, desde enfoques simples que utilizan el filtro de Kalman[16] para predecir la posición futura de un objeto, hasta métodos más avanzados que incorporan redes neuronales profundas para aprender características de apariencia y mejorar la robustez del seguimiento.

El filtro de Kalman es un algoritmo de estimación recursivo fundamental en el seguimiento de objetos. Funciona como un estimador óptimo para sistemas dinámicos lineales, permitiendo predecir el estado futuro de un objeto (como su posición y velocidad) a partir de una serie de mediciones ruidosas o incompletas a lo largo del tiempo. Su proceso se basa en dos etapas cíclicas:

- **Predicción:** Utiliza un modelo dinámico del movimiento esperado del objeto para estimar su estado en el siguiente instante de tiempo.
- **Actualización:** Incorpora la nueva medición (detección) obtenida en ese instante para corregir la predicción inicial, ponderando la información del modelo y la medición según su incertidumbre asociada.

Este ciclo permite al filtro refinarse continuamente la estimación del estado del objeto, suavizar las trayectorias y manejar eficazmente el ruido inherente a las mediciones del detector. Es una herramienta clave para mantener la identidad de los objetos entre fotogramas, especialmente cuando las detecciones son intermitentes o imprecisas.

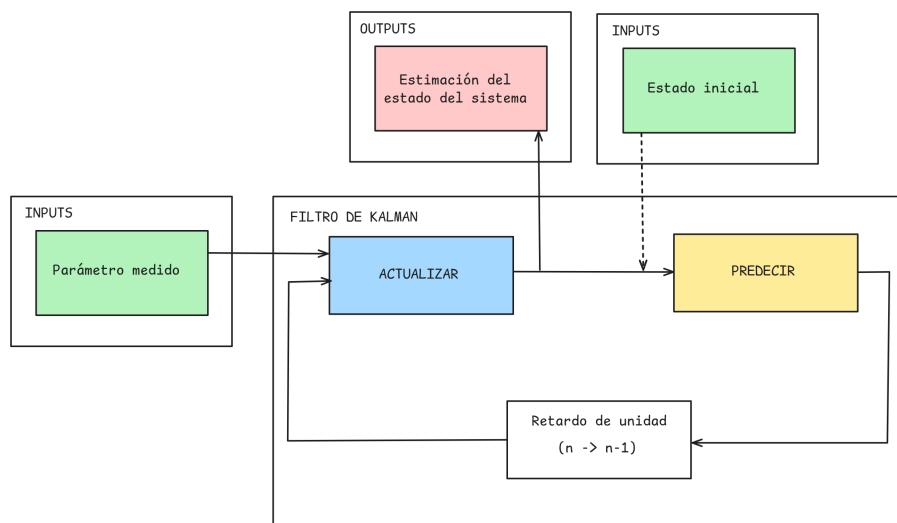


Figura 2.19: Diagrama de flujo del filtro de Kalman.

La Figura 2.19 representa el funcionamiento de un filtro de Kalman, un algoritmo muy utilizado para estimar el estado de un sistema dinámico en presencia de ruido e incertidumbre. A continuación se explica cada bloque:

- **Estado Inicial (Entrada):** Proporciona una estimación inicial del estado del sistema y su incertidumbre asociada, utilizada para inicializar el filtro de Kalman.
- **Mediciones (Entrada):** Este bloque representa las mediciones obtenidas del sistema, que contienen ruido. Estas mediciones se pasan al filtro para su procesamiento.
- **Actualización:** Este paso combina la predicción previa con la medición actual para obtener una estimación actualizada del estado del sistema, ajustando la incertidumbre.
- **Predicción:** Este paso utiliza un modelo del sistema para predecir el siguiente estado y su incertidumbre, basándose en la estimación anterior.
- **Retardo de Unidad ($n \rightarrow n - 1$):** Almacena el estado estimado en el instante anterior para su uso en la siguiente predicción.

Este es el resultado final del filtro: una estimación refinada del estado actual del sistema, que es más precisa que la simple medición directa, permitiendo suavizar las trayectorias de los objetos, predecir su posición futura y manejar la incertidumbre en las mediciones, lo que es esencial para el seguimiento efectivo de objetos en secuencias de vídeo. En conjunto, el filtro de Kalman realiza un ciclo continuo de predicción y corrección, usando tanto el modelo del sistema como las mediciones reales, para obtener una estimación óptima del estado.

2.3.2. BYTETrack

El algoritmo BYTETrack [43] es un algoritmo avanzado de MOT que se inscribe dentro del paradigma de seguimiento por detección (*tracking-by-detection*). Este paradigma, en el que se basan algoritmos como SORT [2] y DeepSORT [42], consiste en apoyarse de las detecciones de objetos en cada fotograma que generan los modelos y luego asociar estas detecciones a lo largo del tiempo para construir trayectorias coherentes, asignando un identificador único consistente a cada objeto a lo largo del tiempo. Para la predicción y suavizado de estas trayectorias, BYTETrack se apoya en los resultados que le ofrece el filtro de Kalman. La innovación fundamental de BYTETrack reside en su novedoso método de asociación de datos, denominado BYTE, que aborda explícitamente un problema común en MOT: el manejo de detecciones con baja puntuación de confianza.

Mientras que la mayoría de los algoritmos de seguimiento descartan las detecciones por debajo de un cierto umbral de confianza para evitar la introducción de falsos positivos, BYTETrack reconoce que estas detecciones de baja confianza a menudo corresponden a objetos reales que están parcialmente ocluidos o cuya apariencia ha cambiado temporalmente. Descartarlas puede llevar a la pérdida de trayectorias y a una menor precisión general del seguimiento. Esta estrategia de asociación en dos pasos permite a BYTETrack recuperar objetos reales incluso cuando la confianza del detector disminuye debido a occlusiones o desenfoque, manteniendo la continuidad de las trayectorias.

Al separar claramente las detecciones de alta y baja confianza y utilizarlas de manera diferenciada en el proceso de asociación, BYTETrack logra una notable mejora en la robustez del seguimiento, reduce significativamente la fragmentación de las trayectorias (medida por métricas como IDF1) y maneja eficazmente las variaciones en la calidad de las detecciones, todo ello manteniendo una alta eficiencia computacional.

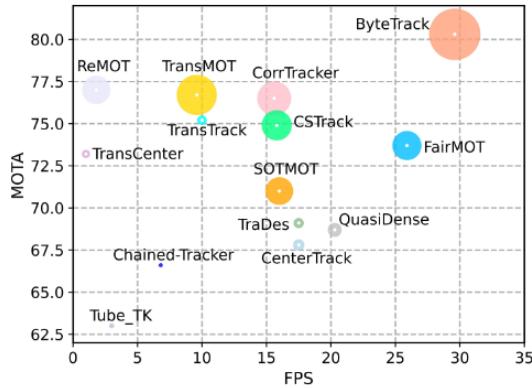


Figura 2.20: Comparativa de rendimiento de BYTETrack con otros algoritmos de seguimiento. Extraído de [43, fig. 1, p. 1].

La Figura 2.20 presenta una comparativa de rendimiento que evidencia la superioridad de BYTETrack frente a otros algoritmos de seguimiento, según los resultados publicados en [43]. Como se observa, BYTETrack no solo alcanza una mayor precisión, medida por la métrica MOTA (Multiple Object Tracking Accuracy), sino que también demuestra una velocidad de procesamiento superior. Estas características lo posicionan como una solución particularmente eficaz y atractiva para aplicaciones que demandan seguimiento de objetos en tiempo real.

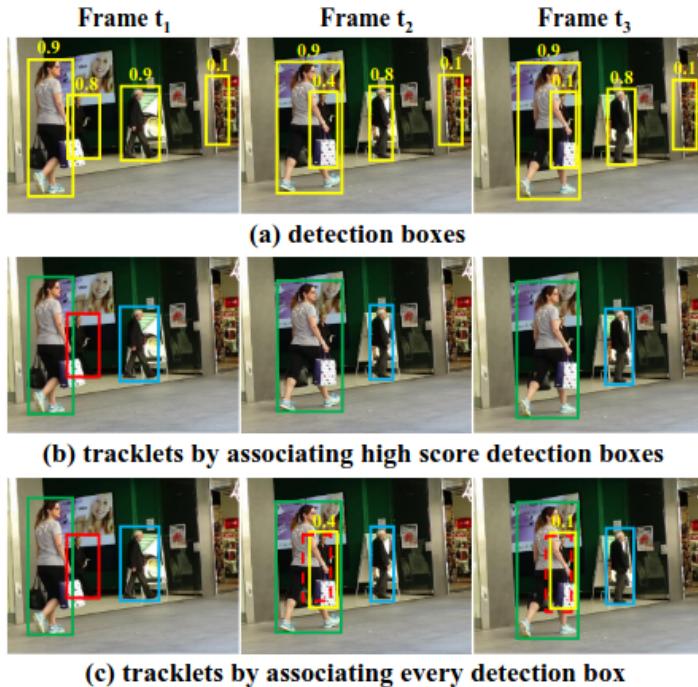


Figura 2.21: Ejemplo de detección y seguimiento de objetos utilizando BYTETrack. Extraído de [43, fig. 2, p. 2].

La Figura 2.21 ilustra el proceso de BYTETrack a través de tres fotogramas consecutivos (τ_1, τ_2, τ_3) de una secuencia de vídeo. En (a), se observan las detecciones iniciales que superan un umbral de confianza (p. ej., 0.5). La sección (b) muestra las trayectorias generadas al asociar exclusivamente las detecciones de alta confianza. Por el contrario, (c) presenta el resultado final de BYTETrack, que integra también las detecciones de baja confianza en el proceso de asociación. Esta comparación evidencia cómo la estrategia de

BYTETrack permite mantener la continuidad de las trayectorias frente a desafíos como occlusiones parciales y variaciones en la confianza de las detecciones, logrando así una representación más precisa y robusta del movimiento de los objetos en el tiempo.

El funcionamiento del algoritmo BYTETrack es el siguiente:

Input: Una secuencia de vídeo V , un detector de objetos Det , y un umbral de confianza de detección τ .

Output: Las trayectorias \mathcal{T} de los objetos detectados en el vídeo.

1. **Inicialización:** Se inicializa el conjunto de trayectorias \mathcal{T} como vacío.
2. **Procesamiento por fotograma:** Para cada fotograma f_k en la secuencia de vídeo V :
 - 2.1. **Detección:** Se utiliza el detector Det para obtener las cajas delimitadoras y sus puntuaciones de confianza para el fotograma f_k , resultando en un conjunto de detecciones \mathcal{D}_k .
 - 2.2. **Separación de detecciones:** Se inicializan dos conjuntos vacíos: \mathcal{D}_{high} para detecciones de alta confianza y \mathcal{D}_{low} para detecciones de baja confianza. Se itera sobre cada detección d en \mathcal{D}_k :
 - Si la puntuación $d.score$ es mayor que el umbral τ , la detección d se añade a \mathcal{D}_{high} .
 - En caso contrario, la detección d se añade a \mathcal{D}_{low} .
 - 2.3. **Predicción de trayectorias:** Para cada trayectoria existente t en \mathcal{T} , se predice su nueva ubicación utilizando un Filtro de Kalman.
 - 2.4. **Primera asociación:** Se asocian las trayectorias \mathcal{T} con las detecciones de alta confianza \mathcal{D}_{high} utilizando el *Hungarian algorithm*[21] con la métrica de similitud IoU (Intersection over Union). Las detecciones no asociadas se guardan en \mathcal{D}_{remain} y las trayectorias no asociadas en \mathcal{T}_{remain} .
 - 2.5. **Segunda asociación:** Se asocian las trayectorias restantes \mathcal{T}_{remain} con las detecciones de baja confianza \mathcal{D}_{low} utilizando otra métrica de similitud (Similarity#2, usualmente IoU). Las trayectorias que siguen sin asociarse se guardan en $\mathcal{T}_{re-remain}$. Solo se asocian detecciones de baja confianza a trayectorias que no pudieron ser asociadas con detecciones de alta confianza.
 - 2.6. **Eliminación de trayectorias no asociadas:** Se eliminan de \mathcal{T} las trayectorias que quedaron en $\mathcal{T}_{re-remain}$ (aquellas que no se pudieron asociar ni en la primera ni en la segunda etapa) si han permanecido sin asociar durante un número determinado de fotogramas (definido por el parámetro `track_buffer`).
 - 2.7. **Inicialización de nuevas trayectorias:** Se itera sobre las detecciones de alta confianza que no fueron asociadas (\mathcal{D}_{remain}). Cada una de estas detecciones se considera el inicio de una nueva trayectoria y se añade al conjunto \mathcal{T} .
3. **Retorno:** Una vez procesados todos los fotogramas, se devuelve el conjunto final de trayectorias \mathcal{T} .

Con todo esto, BYTETrack logra un seguimiento robusto y preciso de múltiples objetos en movimiento, incluso en condiciones desafiantes como occlusiones parciales y cambios de apariencia. Su enfoque innovador para manejar detecciones de baja confianza lo distingue de otros algoritmos de seguimiento y lo convierte en una opción atractiva para aplicaciones en tiempo real.

2.3.3. Métricas de evaluación en seguimiento de objetos múltiples

Las métricas de evaluación son fundamentales para medir el rendimiento de los algoritmos de MOT. Estas métricas permiten cuantificar la precisión y la robustez del seguimiento, facilitando la comparación entre diferentes enfoques y configuraciones.

En el contexto de MOT, los errores pueden clasificarse en tres categorías principales:

- **Errores de detección:** Ocurren cuando el sistema predice detecciones que no existen en la verdad de referencia, o cuando falla en predecir detecciones que sí están presentes en la verdad fundamental. Estos corresponden a los Falsos Positivos (FP) y Falsos Negativos (FN) respectivamente.
- **Errores de asociación (IDSW):** Ocurren cuando la identidad de un objeto se intercambia con la de otro, o cuando un objeto recibe un nuevo ID incorrectamente. Esto se manifiesta como un cambio de identificador (IDSW - *IDentity SWitch*) en la trayectoria del objeto, indicando una falla en el mantenimiento de la coherencia del seguimiento.
- **Errores de localización:** Ocurren cuando las detecciones predichas (prDets) no están perfectamente alineadas espacialmente con las detecciones de la verdad de referencia (gtDets). La calidad de esta alineación se mide típicamente con la métrica IoU (Intersection over Union).

A continuación se presentan las métricas más relevantes en este campo, que cuantifican estos errores de diversas maneras:

- **Multiple Object Tracking Accuracy (MOTA)**[17]: Es una de las métricas más consolidadas y ampliamente utilizadas para evaluar el rendimiento general de un algoritmo MOT. MOTA agrega tres tipos de errores principales que pueden ocurrir durante el seguimiento:
 - Falsos Negativos (FN): Objetos reales presentes en la escena que el algoritmo de seguimiento no detecta o no sigue.
 - Falsos Positivos (FP): Detecciones o trayectorias generadas por el algoritmo de seguimiento que no corresponden a ningún objeto real.
 - Cambios de Identidad (IDSW - *IDentity SWitches*): Ocurren cuando un objeto que ya está siendo seguido se le asigna incorrectamente un nuevo identificador, o cuando se intercambian los identificadores entre dos objetos seguidos.

La fórmula es:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDS\!W_t)}{\sum_t GT_t} \quad (2.7)$$

donde FN_t , FP_t , e $IDS\!W_t$ son el número de falsos negativos, falsos positivos y cambios de identidad en el fotograma t , respectivamente. GT_t es el número total de objetos reales (verdad de referencia) en el fotograma t . El sumatorio se realiza sobre todos los fotogramas de la secuencia. Un valor de MOTA más alto indica un mejor rendimiento, con un máximo teórico de 1 (o 100 %). Sin embargo, MOTA puede ser negativo si el número de errores supera el número de objetos reales. Aunque es una métrica integral, tiende a dar más peso a la precisión de la detección que a la consistencia de la identidad a largo plazo.

- **Multiple Object Tracking Precision (MOTP)[17]:** Esta métrica evalúa la precisión de la localización de los objetos a lo largo del tiempo. Se calcula como la media de las distancias entre las cajas delimitadoras predichas y las cajas delimitadoras reales (verdad de referencia) para todas las asociaciones correctas. Matemáticamente, se expresa como:

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \quad (2.8)$$

donde $d_{t,i}$ es la distancia (generalmente IoU) entre la predicción y la verdad de referencia para el objeto i en el fotograma t , y c_t es el número total de asociaciones correctas en el fotograma t . Un valor de MOTP más alto indica una mayor precisión en la localización de los objetos. A diferencia de MOTA, MOTP no penaliza los errores de identidad, sino que se centra exclusivamente en qué tan precisas son las localizaciones de los objetos cuando se asocian correctamente.

- **ID F1 Score (IDF1)[39]:** Esta métrica se centra específicamente en la capacidad del algoritmo de seguimiento para mantener correctamente la identidad de los objetos a lo largo del tiempo. Es la media armónica de la Precisión de ID (IDP) y el Recall de ID (IDR).
- IDP (ID Precision): Proporción de detecciones correctamente asignadas a una trayectoria (ID) respecto al total de detecciones asignadas.
 - IDR (ID Recall): Proporción de objetos reales correctamente identificados y seguidos a lo largo de su trayectoria respecto al total de objetos reales.

La fórmula es:

$$IDF1 = \frac{2 \cdot IDTP}{2 \cdot IDTP + IDFP + IDFN} \quad (2.9)$$

donde $IDTP$ (ID True Positives) son los verdaderos positivos en términos de asignación de identidad correcta a lo largo de las trayectorias, $IDFP$ (ID False Positives) son las asignaciones de identidad incorrectas, y $IDFN$ (ID False Negatives) son las identidades de objetos reales que no fueron correctamente mantenidas. Un valor de IDF1 más alto (hasta 1 o 100 %) indica una mejor consistencia en el seguimiento de la identidad. Es particularmente útil para evaluar el rendimiento en escenarios con occlusiones prolongadas o interacciones complejas entre objetos. En la figura 2.22 se muestra un ejemplo visual de cómo funciona IDF1.

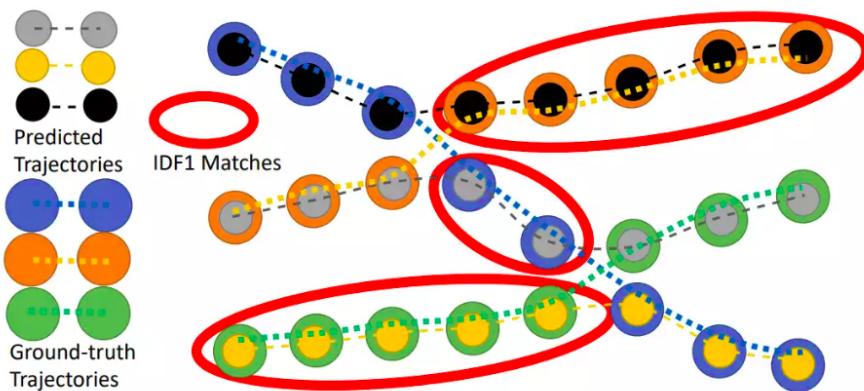


Figura 2.22: Ejemplo visual de IDF1. Extraído de [25, fig. 10, p. 20].

- **Higher Order Tracking Accuracy (HOTA)[25]:** Es una métrica más reciente diseñada para proporcionar una evaluación más equilibrada y completa del rendimiento

del MOT. HOTA descompone explícitamente el rendimiento en precisión de detección, precisión de asociación y precisión de localización. Se calcula como la media geométrica de la Precisión de Detección (DetA) y la Precisión de Asociación (AssA), donde cada una de estas componentes considera la precisión de localización (IoU).

$$HOTA = \sqrt{DetA \cdot AssA} \quad (2.10)$$

- DetA (Detection Accuracy): Mide qué tan bien se detectan los objetos, promediado sobre diferentes umbrales de IoU.
- AssA (Association Accuracy): Mide qué tan bien se asocian las detecciones correctas para formar trayectorias consistentes, también promediado sobre umbrales de IoU.

HOTA varía entre 0 y 1 (o 0 % y 100 %), donde valores más altos indican un mejor rendimiento. Se considera que HOTA ofrece una visión más matizada que MOTA, ya que penaliza de forma más equilibrada los diferentes tipos de errores y es sensible a la calidad de la localización.

CAPÍTULO 3

Diseño e implementación de la solución

En este capítulo se abordará en profundidad el diseño y la implementación del sistema de visión artificial propuesto. Se iniciará con un análisis detallado del problema a resolver, definiendo los desafíos inherentes a la detección y seguimiento de objetos en movimiento y los requisitos específicos del sistema, como la operación en tiempo real.

A continuación, se describirá el proceso de entrenamiento de los modelos de detección de objetos, la metodología para la creación y anotación del conjunto de datos de canicas, y los parámetros de entrenamiento utilizados.

Posteriormente, se presentará una descripción global del sistema, detallando su arquitectura modular y el flujo de datos entre sus componentes. Se profundizará en el diseño específico de cada una de las etapas que conforman el sistema: la captura de imágenes, la inferencia del modelo de detección, el seguimiento de los objetos mediante el algoritmo BYTETrack y la escritura de los resultados.

Finalmente, se explorarán y justificarán las diversas estrategias de segmentación del sistema que se han considerado e implementado con el objetivo de optimizar el rendimiento y la eficiencia del procesamiento en la plataforma NVIDIA Jetson.

3.1 Análisis del problema

El desafío central abordado en este trabajo consiste en el desarrollo y la implementación de un sistema de visión artificial capaz de realizar el seguimiento de múltiples objetos en movimiento en tiempo real y poder detectar sus posibles defectos.

Este sistema se fundamenta en la utilización de la plataforma de hardware NVIDIA Jetson, reconocida por su capacidad para ejecutar tareas de IA de manera eficiente en términos energéticos y computacionales. La tarea principal del sistema es procesar una secuencia de vídeo, identificar los objetos presentes en cada fotograma mediante un modelo de detección de objetos basado en redes neuronales profundas, y posteriormente, aplicar un algoritmo de seguimiento para mantener la identidad de cada objeto a lo largo del tiempo, reconstruyendo así sus trayectorias.

Un requisito fundamental y crítico para la viabilidad del sistema es su capacidad para operar en tiempo real. Esto impone una restricción estricta sobre la velocidad de procesamiento: el tiempo total necesario para analizar un fotograma individual, incluyendo tanto la detección como el seguimiento de los objetos, debe ser inferior al intervalo de tiempo que transcurre entre fotogramas consecutivos en la secuencia de vídeo. Por ejemplo, para

un vídeo a 30 fotogramas por segundo, el procesamiento completo de cada fotograma debe completarse en menos de 33.3 milisegundos. Cumplir con esta exigencia es particularmente desafiante dadas las limitaciones inherentes de los dispositivos embebidos como los de la serie Jetson, que, aunque potentes, disponen de recursos computacionales y memoria significativamente menores en comparación con sistemas de escritorio o servidores.

Dada la dificultad de acceder a un entorno industrial real para llevar a cabo las pruebas experimentales —como podría ser una línea de producción activa en una fábrica de conservas, una planta de ensamblaje de componentes electrónicos o una instalación de procesamiento de alimentos—, se optó por utilizar un entorno simulado controlado.

Este trabajo utiliza un entorno simulado debido a la dificultad de acceder a un entorno industrial real. Este entorno sustituye los objetos industriales por canicas de diversos colores, que al moverse sobre una cinta transportadora improvisada, simulan el flujo de objetos que se encontraría en una línea de producción.

La utilización de este entorno simulado ofrece ventajas significativas para la fase de desarrollo y evaluación: permite realizar pruebas de manera sistemática y repetible, facilita la variación controlada de parámetros (como la velocidad de los objetos, la iluminación o la densidad de objetos) y posibilita la obtención de datos cuantitativos precisos sobre el rendimiento del sistema en diferentes condiciones. Aunque este entorno simplifica la complejidad del mundo real, los resultados y las conclusiones obtenidas proporcionan una base sólida y pueden ser extrapolados, con las debidas consideraciones, para predecir el comportamiento y la eficacia del sistema en escenarios industriales auténticos.



Figura 3.1: Ejemplo de entorno simulado con canicas.

Para las pruebas experimentales, se configuró un entorno simulado como el ilustrado en la Figura 3.1. Este entorno utiliza canicas de cuatro colores distintos (blanco, negro, azul y verde). Con el objetivo de simular la detección de anomalías, se consideraron tanto canicas sin defectos como canicas con defectos para cada uno de los colores. Como defecto, se añadió una mancha de un color diferente al de la canica, presentando diversas formas. Esta distinción duplica el número total de clases que el sistema debe identificar, alcanzando un total de ocho categorías (cuatro colores sin defecto y cuatro colores con defecto).

3.2 Entrenamiento de los modelos

La selección y el entrenamiento de los modelos de detección de objetos constituyen una fase crítica en el desarrollo del sistema propuesto, ya que de ello dependen directamente la precisión y la robustez del sistema final. Tras un análisis comparativo, se optó por los detectores de una etapa frente a los de dos etapas, dada su superior eficiencia para aplicaciones que exigen procesamiento en tiempo real. Dentro de esta categoría, se seleccionó la familia de modelos YOLO. Esta elección se fundamenta en su reconocido equilibrio entre velocidad de inferencia y precisión, así como en la disponibilidad de un robusto ecosistema de herramientas que facilitan tanto el entrenamiento como el despliegue. Alternativas como SSD[24], si bien competentes, no ofrecían el mismo conjunto de ventajas en términos de comunidad de soporte y facilidad de integración para los fines de este proyecto.

Como modelos pertenecientes a la familia de arquitecturas YOLO, se han seleccionado diversas variantes para una evaluación exhaustiva: YOLOv5 (específicamente las versiones "n" y "m"), YOLOv8 (también en sus variantes "n" y "s") y YOLOv11 (en sus versiones "n", "s", "m" y "l"). Esta elección se basa en varios factores clave. En primer lugar, estas familias de modelos YOLO son conocidas por su excelente equilibrio entre velocidad de inferencia y precisión de detección, lo que las hace particularmente adecuadas para aplicaciones que requieren procesamiento en tiempo real, como es el caso de este proyecto. En segundo lugar, estos modelos, especialmente a través de la implementación proporcionada por la biblioteca Ultralytics, ofrecen una gran flexibilidad y facilidad de uso para el entrenamiento, la validación y el despliegue.

Una vez seleccionado los modelos, el siguiente paso crítico es la creación y preparación del conjunto de datos (*dataset*). Este conjunto de datos es la base sobre la cual los modelos aprenderán a identificar y localizar los objetos de interés (en este caso, las canicas de diferentes colores y con/sin defectos).

La calidad y representatividad del *dataset* tienen un impacto directo y significativo en el rendimiento final de los modelos. Un *dataset* bien construido debe incluir una variedad suficiente de ejemplos que cubran las diferentes condiciones que el sistema podría encontrar en el entorno real, como variaciones en la iluminación, ángulos de visión, occlusiones parciales y la diversidad intrínseca de los objetos mismos.

El proceso de creación del *dataset* implica la captura de imágenes o videos del entorno simulado, seguido de una meticulosa fase de etiquetado (anotación), donde se marcan manualmente las cajas delimitadoras (*bounding boxes*) alrededor de cada objeto de interés y se les asigna la etiqueta de clase correspondiente (p. ej., "canica_azul_defecto"). Este proceso, aunque laborioso, es indispensable para proporcionar a los modelos la "verdad fundamental" (*ground truth*) necesaria para su aprendizaje supervisado.

Para el etiquetado de las imágenes, se utilizó la herramienta Computer Vision Annotation Tool (CVAT)[6] de código abierto, que permite realizar anotaciones precisas y eficientes en imágenes, como se muestra en la Figura 3.2. Permite la creación de diferentes tipos de anotaciones, como cajas delimitadoras, polígonos y puntos clave además de la exportación de los datos anotados en varios formatos compatibles con diferentes frameworks de aprendizaje profundo. En este caso, se optó por el formato YOLO, que es ampliamente utilizado y compatible con la implementación de Ultralytics.

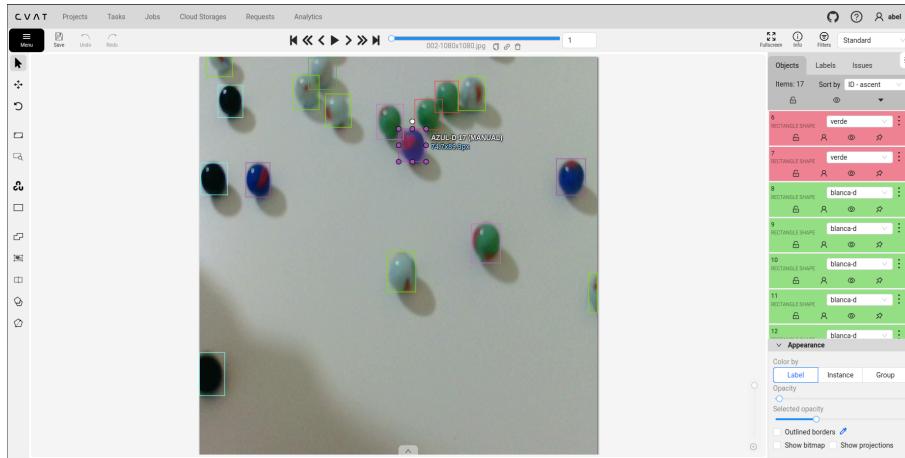


Figura 3.2: Ejemplo de anotación de imágenes utilizando CVAT.

Este formato sigue una estructura de texto simple, donde cada línea representa una anotación para un objeto en la imagen. Cada línea contiene cinco valores: el índice de la clase (0 para canica blanca, 1 para canica negra, 2 para canica azul, 3 para canica verde, 4 para canica blanca con defecto, 5 para canica negra con defecto, 6 para canica azul con defecto y 7 para canica verde con defecto), seguido de las coordenadas normalizadas del centro de la caja delimitadora (x, y) y su ancho y alto (w, h), todos ellos en relación a las dimensiones de la imagen.

Volviendo al dataset, se capturaron un total de 600 imágenes, de las cuales el 80 % se utilizaron para el entrenamiento, el 10 % para la validación y el 10 % restante para el testeo del modelo. Esta división es crucial para garantizar que el modelo no solo aprenda a detectar los objetos en las imágenes de entrenamiento, sino que también generalice bien a nuevas imágenes que no ha visto antes. La validación se utiliza para ajustar los hiperparámetros del modelo y evitar el sobreajuste (*overfitting*), mientras que el conjunto de testeo proporciona una evaluación final del rendimiento del modelo.

Como se ha mencionado en la sección 2.1.6 y se detalla en la Tabla 3.1, los modelos YOLO presentan diversas variantes que difieren en tamaño, complejidad, latencia y precisión. Esta tabla comparativa es fundamental para seleccionar el modelo que mejor equilibre la velocidad de procesamiento y la precisión requerida.

Familia	Variante	Tamaño (px)	Parámetros (M)	Latencia CPU ONNX (ms)	Latencia GPU (ms)	GPU (para Latencia)
YOLOv5	nu	640	2.6	73.6	1.06	A100 TensorRT
	mu	640	25.1	233.9	1.86	
YOLOv8	n	640	3.2	80.4	0.99	A100 TensorRT
	s	640	11.2	128.4	1.20	
YOLO11	n	640	2.6	56.1 ± 0.8	1.5 ± 0.0	
	s	640	9.4	90.0 ± 1.2	2.5 ± 0.0	
	m	640	20.1	183.2 ± 2.0	4.7 ± 0.1	T4 TRT10
	l	640	25.3	238.6 ± 1.4	6.2 ± 0.1	
	x	640	56.9	462.8 ± 6.7	11.3 ± 0.2	

Tabla 3.1: Análisis comparativo de variantes de YOLO (v5, v8, 11) indicando tamaño, parámetros, latencias CPU/GPU y la GPU específica utilizada para la medición de latencia GPU. Datos extraídos de [14].

Para el entrenamiento de los modelos se han seleccionado los siguientes hiperparámetros:

- **Tasa de aprendizaje (*learning rate*):** Se ha utilizado un valor inicial de 0.01, con un ajuste posterior basado en la tasa de convergencia observada durante el entrenamiento.
- **Número de épocas (*epochs*):** Se han realizado 30 épocas, para permitir que el modelo aprenda de manera efectiva sin caer en el sobreajuste.
- **Tamaño del lote (*batch size*):** Se ha establecido en 16, lo que permite un equilibrio entre la velocidad de entrenamiento y la utilización de memoria.
- **Tamaño de imagen (*image size*):** Se ha utilizado una resolución de 640x640 píxeles, que es un tamaño estándar para los modelos YOLO y proporciona un buen compromiso entre precisión y velocidad.
- **Optimizador (*optimizer*):** Se ha utilizado el optimizador AdamW, que es conocido por su eficacia en el entrenamiento de modelos de aprendizaje profundo.
- **Tasa de aumento de datos (*data augmentation*):** Se han aplicado técnicas de aumento de datos como rotación, cambio de brillo y contraste, y recortes aleatorios para mejorar la generalización del modelo.
- **Dispositivo (*device*):** Se ha utilizado una Jetson AGX Xavier, que proporciona un entorno de hardware optimizado para el entrenamiento.

Los resultados del entrenamiento se han evaluado utilizando el conjunto de validación, y se han registrado métricas como la **precisión**, el **recall** y el **mAP**. Estas métricas son fundamentales para entender el rendimiento del modelo y su capacidad para generalizar a nuevos datos.

Modelo	Tiempo (h) ↓	Precisión ↑	Recall ↑	mAP50 ↑	mAP50-95 ↑
YOLOv5nu	0.128	0.898	0.922	0.939	0.759
YOLOv5mu	0.333	0.952	0.928	0.955	0.790
YOLOv8n	0.137	0.934	0.905	0.950	0.770
YOLOv8s	0.202	0.943	0.928	0.955	0.786
YOLO11n	0.140	0.950	0.882	0.939	0.761
YOLO11s	0.192	0.941	0.936	0.963	0.796
YOLO11m	0.377	0.941	0.936	0.963	0.796
YOLO11l	0.485	0.954	0.947	0.967	0.801

Tabla 3.2: Comparativa del rendimiento de los modelos YOLOv5, YOLOv8 y YOLO11 en términos de tiempo de entrenamiento, precisión, recall y mAP.

La Tabla 3.2 resume los resultados del entrenamiento para las variantes de YOLOv5, YOLOv8 y YOLO11. Se observa una tendencia general: a medida que aumenta la complejidad del modelo (de las versiones más pequeñas a las más grandes dentro de cada familia), tanto la precisión como el recall experimentan una mejora constante. Esto sugiere que los modelos de mayor tamaño poseen una capacidad superior para aprender representaciones de características más ricas y discriminativas, resultando en una detección de objetos más efectiva. No obstante, esta mejora en el rendimiento se acompaña de un incremento en el tiempo de entrenamiento y la demanda de recursos computacionales, lo que refleja el inherente compromiso entre la capacidad del modelo y la eficiencia del proceso de aprendizaje.

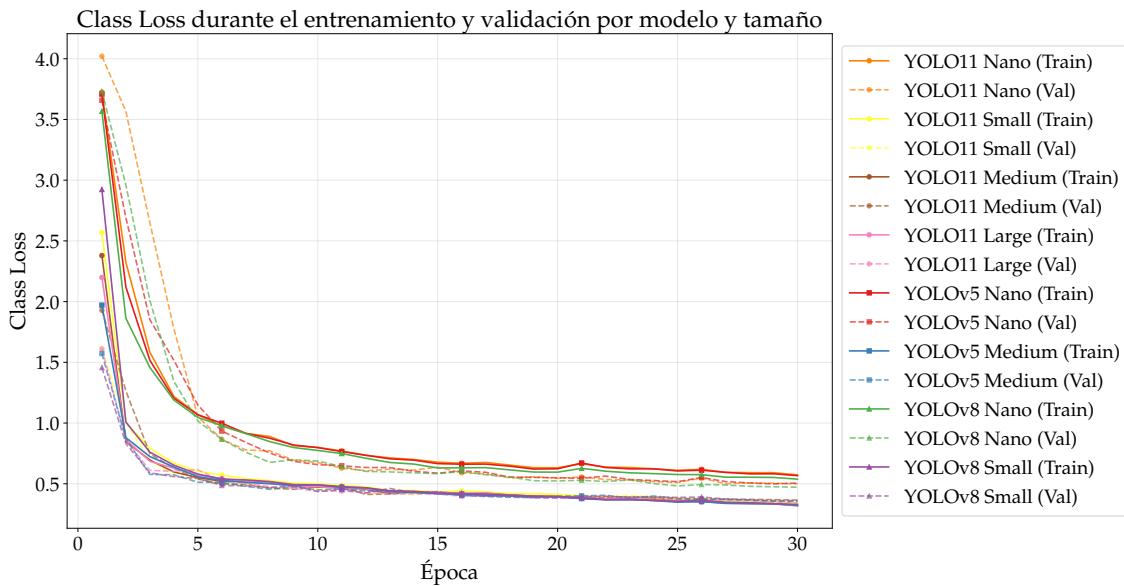


Figura 3.3: Curvas de pérdida de entrenamiento y validación para las distintas tallas de modelos YOLOv5, YOLOv8 y YOLO11.

La Figura 3.3 presenta las curvas de pérdida (*loss*) durante el entrenamiento y la validación para las diferentes variantes de YOLOv5, YOLOv8 y YOLO11. Estas gráficas evidencian una progresiva disminución de la función de pérdida a lo largo de las épocas, lo cual es un indicador clave de que cada modelo está aprendiendo a generalizar a partir de los datos para la tarea de detección de objetos.

Es notable que, en todas las familias, la pérdida de validación sigue una tendencia descendente similar a la de entrenamiento, sugiriendo una buena capacidad de generalización y la ausencia de un sobreajuste significativo.

Se aprecia un descenso pronunciado de la pérdida en las épocas iniciales, indicativo de una rápida asimilación de patrones, y conforme avanza el entrenamiento, la tasa de reducción disminuye gradualmente hasta alcanzar una meseta. Este comportamiento es característico en el entrenamiento de modelos de aprendizaje profundo y señala la convergencia hacia un mínimo local en la función de pérdida, donde las mejoras adicionales resultan marginales.

3.3 Descripción del sistema

El sistema está organizado como una serie de etapas de procesamiento secuenciales que trabajan de forma coordinada para lograr la detección y seguimiento de objetos en tiempo real. Como se muestra en la Figura 3.4, el sistema recibe como entrada imágenes de una cámara y consta de cuatro componentes principales: un módulo de captura de imágenes que obtiene los fotogramas del vídeo, un módulo de inferencia que ejecuta el modelo de detección de objetos, un módulo de seguimiento que implementa el algoritmo BYTETrack para mantener la identidad de los objetos detectados, y un módulo de escritura que gestiona la salida del sistema.

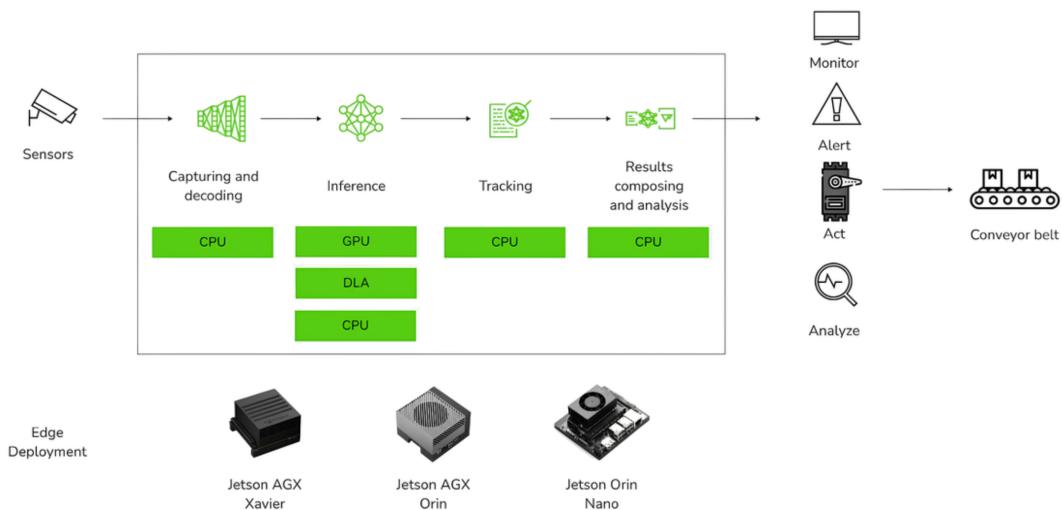


Figura 3.4: Figura del sistema propuesto.

Los resultados del sistema pueden utilizarse de diversas formas, monitorización en tiempo real a través de visualización de las detecciones, generación de alertas basadas en reglas predefinidas, interacción con sistemas de control automatizado para realizar acciones específicas, y almacenamiento de datos para su posterior análisis y extracción de métricas que permitan optimizar procesos industriales.

La arquitectura del sistema se ha diseñado estratégicamente para maximizar el rendimiento computacional y la eficiencia en el uso de recursos, asegurando una operación coordinada entre sus componentes modulares. Los datos fluyen secuencialmente a través de las distintas etapas, permitiendo un procesamiento continuo y eficaz de la información visual.

Esta estructura modular intrínseca no solo simplifica el mantenimiento y las futuras actualizaciones, sino que también dota al sistema de una gran flexibilidad para adaptarse a diferentes requisitos operativos y escenarios de despliegue. Aunque optimizado específicamente para la plataforma NVIDIA Jetson, su diseño modular facilita su potencial portabilidad a otras plataformas hardware de NVIDIA, como GPUs de escritorio o servidores de alto rendimiento, siempre que se cumplan los requisitos de hardware y software.

3.4 Diseño de las etapas del sistema

El sistema propuesto se compone de cuatro etapas principales, cada una de las cuales desempeña un papel crucial en el procesamiento y análisis de las imágenes capturadas. A continuación, se describen en detalle cada una de estas etapas.

3.4.1. Captura de imágenes

La etapa de captura de imágenes es la responsable de adquirir los fotogramas del flujo de vídeo en tiempo real. Sus funciones abarcan la configuración y el control de la cámara, la adquisición de las imágenes y su posible preprocesamiento inicial antes de ser transferidas al módulo de inferencia. Para la implementación de este módulo se ha empleado la biblioteca OpenCV, la cual ofrece una interfaz robusta y eficiente para la interacción con dispositivos de captura y el procesamiento básico de imágenes. La cámara

se configura para operar a una resolución y tasa de fotogramas adecuadas a los requisitos de la aplicación. El preprocessamiento puede incluir diversas operaciones, tales como la conversión a escala de grises, la normalización de píxeles o el redimensionamiento, adaptándose a las especificaciones del modelo de detección de objetos utilizado. En el sistema desarrollado, se empleó una cámara de alta definición capaz de alcanzar una resolución máxima de 1920x1080 píxeles y una tasa de 30 fps. Si bien se experimentó con diversas configuraciones de resolución y tasa de fotogramas según las necesidades de cada prueba, la configuración estándar para los experimentos realizados fue de 640x640 píxeles a 30 fps. Esta etapa se ejecuta íntegramente en la CPU.

3.4.2. Inferencia

La etapa de inferencia constituye el núcleo computacional del sistema, siendo responsable de ejecutar el modelo de detección de objetos preentrenado sobre cada fotograma adquirido por la etapa de captura.

El proceso de inferencia comienza con la adquisición de un fotograma de vídeo, que se convierte en un tensor adecuado para la entrada del modelo. Este tensor es una representación numérica de la imagen, donde cada píxel se traduce en un valor que el modelo puede procesar. La transformación del fotograma a tensor incluye operaciones como la normalización y el redimensionamiento, asegurando que los datos estén en el formato correcto para el modelo. En el caso de los modelos YOLO, este tensor suele tener una forma de $(batch_size, channels, height, width)$, donde "channels" representa los canales de color (típicamente 3 para RGB), y "height" y "width" son las dimensiones de la imagen redimensionada. Este paso es crucial para garantizar que el modelo pueda interpretar correctamente la información visual contenida en el fotograma.

La ejecución de la inferencia propiamente dicha implica cargar el modelo YOLO (potencialmente optimizado mediante NVIDIA TensorRT para maximizar el rendimiento en la plataforma Jetson) y pasarle el tensor de entrada. Aprovechando la aceleración por hardware (GPU y/o DLAs disponibles en los módulos Jetson), el modelo procesa la imagen y genera un conjunto de predicciones. Estas predicciones iniciales suelen ser numerosas y requieren un postprocesamiento para refinar los resultados. Este postprocesamiento, a menudo gestionado internamente por el framework Ultralytics o aplicado explícitamente, incluye la aplicación de un umbral de confianza para descartar detecciones poco fiables y la ejecución del algoritmo de NMS para eliminar cajas delimitadoras redundantes que correspondan al mismo objeto.

El resultado final de la etapa de inferencia, que se transfiere a la etapa de seguimiento, es una lista estructurada de las detecciones finales para el fotograma actual. Cada detección en esta lista contiene información esencial: las coordenadas de la caja delimitadora que localiza al objeto (comúnmente en formato (x, y, w, h) , donde (x, y) representan las coordenadas del centro de la caja, y (w, h) su anchura y altura), una puntuación de confianza que cuantifica la fiabilidad de la detección, y la etiqueta de la clase predicha para el objeto (p. ej., "canica_azul", "canica_verde_defecto"). La eficiencia y rapidez de esta etapa son críticas para mantener la capacidad de procesamiento en tiempo real del sistema global.

3.4.3. Seguimiento

La etapa de seguimiento es la encargada de mantener la identidad de los objetos detectados a lo largo del tiempo, asegurando que cada objeto en el flujo de vídeo conserve su etiqueta y trayectoria a pesar de las variaciones en su posición, apariencia o posibles

occlusiones. Para lograr esto, se ha utilizado la implementación del algoritmo BYTETrack en el *framework* de Ultralytics[15], que se basa en un enfoque de seguimiento por detección explicado en la subsección 2.3.2. Este algoritmo se encarga de asociar las detecciones generadas por la etapa de inferencia con las trayectorias existentes, utilizando tanto las detecciones de alta confianza como las de baja confianza para mejorar la robustez del seguimiento.

Para la configuración del algoritmo BYTETrack existen varios parámetros ajustables que permiten optimizar su rendimiento según las características específicas del entorno y los objetos a seguir. Estos parámetros son:

- **track_high_thresh**: Umbral de confianza para considerar una detección como de alta confianza (valor típico: 0.6). Las detecciones por encima de este umbral se utilizan en la primera etapa de asociación.
- **track_low_thresh**: Umbral de confianza para considerar una detección como de baja confianza (valor típico: 0.15). Las detecciones que se encuentren entre este umbral y **track_high_thresh** se utilizan en la segunda etapa de asociación para recuperar objetos ocluidos.
- **new_track_thresh**: Umbral de confianza mínimo para iniciar una nueva trayectoria a partir de una detección de alta confianza no asociada (valor típico: 0.6). Debe ser al menos tan alto como **track_high_thresh**.
- **track_buffer**: Número máximo de fotogramas que una trayectoria puede permanecer sin asociar antes de ser eliminada. Define la “edad” máxima de una pista perdida. Un valor típico es 30 fotogramas.
- **match_thresh**: Umbral de IoU (Intersection over Union) para la asociación entre las predicciones del Filtro de Kalman y las detecciones (valor típico: 0.8). Si la IoU es mayor que este umbral, se considera una coincidencia potencial.

Este algoritmo se basa en un ciclo continuo de predicción y corrección, donde el Filtro de Kalman se utiliza para predecir la posición futura de los objetos y suavizar las trayectorias, manejando la incertidumbre en las mediciones. La asociación de detecciones y trayectorias se realiza mediante un algoritmo de asignación, como el algoritmo Húngaro[21], que busca minimizar el costo total de la asociación entre las detecciones y las trayectorias existentes. Se ejecuta íntegramente en la CPU debido a la implementación del algoritmo en la biblioteca de Ultralytics.

Tras aplicar la asociación de detecciones y trayectorias, el algoritmo ofrece como salida una lista con el identificador único de cada objeto, su clase, la puntuación de confianza y las coordenadas de la caja delimitadora. Esta información es fundamental para la siguiente etapa del sistema.

3.4.4. Escritura de resultados

La etapa de escritura de resultados es responsable de gestionar la salida del sistema, que puede adoptar diversas formas según los requisitos específicos de la aplicación. Esta etapa se encarga de presentar los resultados de manera comprensible y útil, permitiendo su interpretación y análisis posterior. Las principales funciones de esta etapa incluyen la visualización de los resultados en tiempo real, la generación de alertas basadas en reglas predefinidas y el almacenamiento de datos para su posterior análisis. También se encarga de los posibles conexiones a sistemas de control automatizado, permitiendo la interacción con otros sistemas o dispositivos.



Figura 3.5: Ejemplo de salida del sistema.

La Figura 3.5 muestra un ejemplo de la salida del sistema, donde se visualizan las detecciones y trayectorias de los objetos en el flujo de vídeo. Cada objeto detectado está representado por una caja delimitadora que incluye su etiqueta de clase y un identificador único. Esta representación gráfica permite una rápida identificación y seguimiento de los objetos en movimiento, facilitando la monitorización en tiempo real del sistema.

Un problema potencial en esta etapa es que un objeto, como una canica defectuosa, haya sido detectado como tal en fotogramas anteriores, pero que, debido a su movimiento, el defecto se haya ocultado temporalmente. En este caso, el sistema podría no detectarlo como defectuoso, llevando a una interpretación errónea. Para mitigar esto, se implementa un algoritmo de memoria que, utilizando el identificador único de cada objeto, recuerda su estado anterior y aplica una lógica de corrección para mantener la coherencia en la clasificación a lo largo del tiempo.

El siguiente algoritmo, se encarga de actualizar una estructura de memoria que almacena información sobre los objetos detectados y rastreados a lo largo del tiempo en una secuencia de vídeo. Su propósito principal es recordar si un objeto ha sido clasificado como defectuoso, incluso cuando el defecto deja de ser visible temporalmente en fotogramas posteriores.

La memoria es un diccionario donde cada clave es el identificador único de un objeto (`track_id`), y el valor asociado es un objeto con la siguiente estructura:

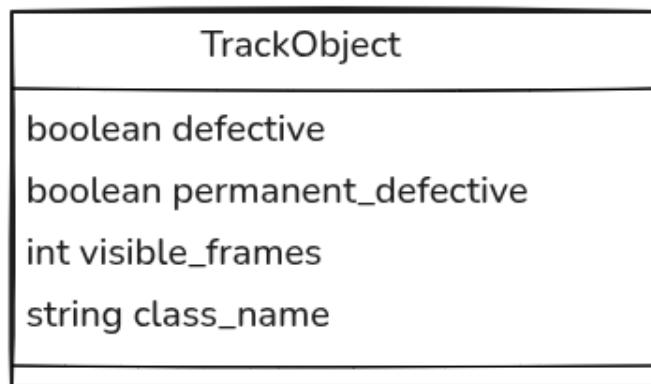


Figura 3.6: Modelo del objeto rastreado.

En el objeto, `defective` indica si el objeto en cuestión es defectuoso en el fotograma actual, `permanent_defective` indica si el objeto ha sido clasificado como defectuoso de forma permanente (es decir, ha sido detectado como defectuoso en varios fotogramas consecutivos), `visible_frames` es un contador que indica cuántos fotogramas han pasado desde la última vez que el objeto fue visto, y `class` es la clase del objeto detectado.

```

{
  1: {
    "defective": False,
    "permanent_defective": False,
    "visible_frames": 30,
    "class": "canica_azul"
  },
  2: {
    "defective": True,
    "permanent_defective": True,
    "visible_frames": 5,
    "class": "canica_roja_defecto"
  },
  ...
}
  
```

Listing 3.1: Ejemplo de memoria de seguimiento

■ Parámetros de entrada:

- `tracked_objects`: Lista de objetos rastreados en el fotograma actual. Cada objeto contiene, entre otros atributos, un identificador único de seguimiento (`track_id`) y la clase detectada.
- `memory`: Diccionario que representa la memoria persistente de los objetos rastreados. La clave es el identificador del objeto, y el valor es un diccionario con información como si el objeto es defectuoso, cuántos fotogramas ha sido visible, su clase, etc.
- `classes`: Lista de nombres de clase, utilizada para traducir el índice de clase numérico a su representación textual.

■ Funcionamiento:

1. Se define una constante `FRAME_AGE = 60`, que representa cuántos fotogramas puede permanecer un objeto en la memoria sin ser visto antes de ser eliminado.

2. También se define un umbral `PERMANENT_DEFECT_THRESHOLD = 3`, que indica cuántos fotogramas consecutivos un objeto debe ser detectado como defectuoso para marcarse como defectuoso de forma permanente.
3. Para cada objeto en `tracked_objects`:
 - Se extrae el identificador de seguimiento y la clase correspondiente.
 - Se determina si la clase es defectuosa o no, basándose en un conjunto predefinido de clases defectuosas.
 - Si el objeto ya existe en la memoria:
 - Si ya estaba marcado como defectuoso permanente, simplemente se actualiza su contador de visibilidad.
 - Si se detecta nuevamente como defectuoso, se incrementa un contador. Si este contador alcanza el umbral establecido, el objeto se marca como defectuoso permanente.
 - Si el objeto no es defectuoso en este fotograma, se reinicia el contador.
 - Se actualizan los campos `defective`, `visible_frames` y `class`.
 - Si el objeto no existía previamente en la memoria, se crea una nueva entrada con sus atributos iniciales.
4. Finalmente, se recorre toda la memoria decrementando `visible_frames` de cada objeto. Si este contador llega a cero, se elimina el objeto de la memoria.

Este procedimiento permite mantener un historial consistente de los objetos rastreados, y ayuda a preservar la información de defectos incluso cuando estos no son visibles temporalmente, mejorando la robustez del sistema frente a interrupciones breves en la detección.

3.5 Segmentación de las etapas del sistema ---

Tras la implementación de las etapas del sistema, se ha considerado la posibilidad de segmentar el sistema mediante estas etapas para mejorar el rendimiento y la eficiencia del procesamiento. La segmentación permite distribuir la carga de trabajo entre diferentes unidades de procesamiento, optimizando así el uso de los recursos disponibles en la plataforma.

La segmentación de las etapas del sistema se puede realizar de varias maneras, dependiendo de los requisitos específicos de la aplicación y de los recursos disponibles. A continuación, se describen las diferentes opciones de segmentación que se han considerado, implementado y evaluado en el sistema propuesto.

3.5.1. Secuencial

La primera y más trivial opción es la ejecución secuencial de las etapas del sistema. En este enfoque, cada etapa se ejecuta de forma consecutiva, donde la salida de una etapa se convierte en la entrada de la siguiente. Este método es el más sencillo de implementar y no requiere una configuración adicional para la comunicación entre etapas. Sin embargo, presenta limitaciones significativas en términos de rendimiento y eficiencia, ya que no aprovecha al máximo los recursos disponibles. Si se hace la analogía con un procesador, este enfoque se asemeja a un procesador no segmentado.

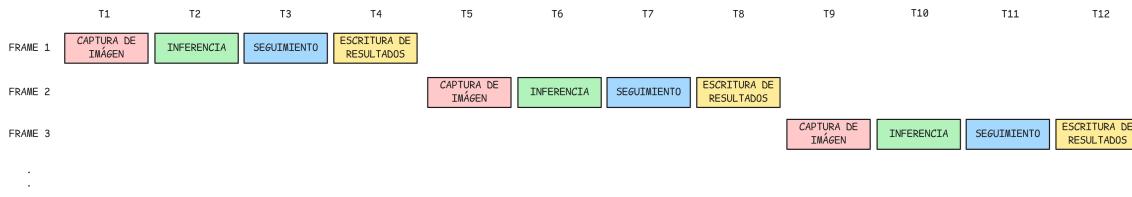


Figura 3.7: Diagrama de flujo del sistema sin segmentar.

La Figura 3.7 ilustra el flujo de datos en un sistema secuencial. En este diagrama, cada etapa del sistema se ejecuta de forma lineal, donde la salida de una etapa se convierte en la entrada de la siguiente. Este enfoque es fácil de entender y de implementar, pero no aprovecha al máximo los recursos disponibles.

3.5.2. Segmentación en hilos

La segunda opción es la segmentación del sistema en diferentes hilos. En este enfoque, cada etapa principal del sistema (captura, inferencia, seguimiento, escritura) se ejecuta en un hilo (*thread*) independiente. A diferencia del enfoque secuencial donde cada etapa debe esperar a que la anterior finalice, la segmentación por hilos permite que las etapas operen de forma concurrente, solapando sus ejecuciones. Esto puede mejorar significativamente el rendimiento (*throughput*) y reducir la latencia, ya que mientras una etapa espera por una operación (p. ej., E/S de la cámara), otra etapa puede estar procesando datos (p. ej., inferencia en GPU o seguimiento en CPU).

Los hilos operan dentro del mismo proceso, compartiendo el mismo espacio de memoria. La comunicación y transferencia de datos (fotogramas, detecciones) entre estas etapas/hilos se gestiona mediante colas de mensajes (*Queue[?]*) de la librería estándar de Python, que aseguran una transferencia eficiente y segura entre hilos (*thread-safe*).

En Python, un hilo es una unidad básica de ejecución. Sin embargo, al trabajar con hilos en Python, es crucial entender el impacto del Global Interpreter Lock (GIL). El GIL es un mecanismo de bloqueo mutuo (*mutex*) que protege el acceso al intérprete de Python. Su función principal es permitir que solo un hilo ejecute código de bytes Python (*Python bytecode*) a la vez dentro de un único proceso, incluso si el sistema dispone de múltiples núcleos de CPU. Es decir, en Python, debido al GIL, dos hilos de un mismo proceso no pueden ejecutar código Python de manera simultánea en núcleos de CPU diferentes. Esta restricción se implementó originalmente para simplificar la gestión de memoria (específicamente, el conteo de referencias) y prevenir condiciones de carrera en el acceso a objetos Python.

La principal consecuencia del GIL es que impide el verdadero paralelismo para tareas que son intensivas en CPU (*CPU-bound*) y están escritas puramente en Python. Aunque se creen múltiples hilos, solo uno podrá ejecutar código Python en un instante dado. En aplicaciones con muchos hilos compitiendo por la CPU, el GIL puede incluso introducir sobrecarga y contención, llevando a un rendimiento inferior al de una ejecución secuencial.

No obstante, el GIL no bloquea la ejecución en todas las circunstancias. Se libera automáticamente durante operaciones que no ejecutan código Python directamente, como:

- Operaciones de entrada/salida (E/S): Lectura/escritura de archivos, operaciones de red, interacción con dispositivos como cámaras.

- Llamadas a código nativo compilado: Cuando se utilizan bibliotecas como NumPy, SciPy, o las bibliotecas específicas para la ejecución en GPU (como las de CUDA/-TensorRT), que realizan el cómputo fuera del intérprete Python.
- Llamadas explícitas de espera: Como ‘time.sleep()’.

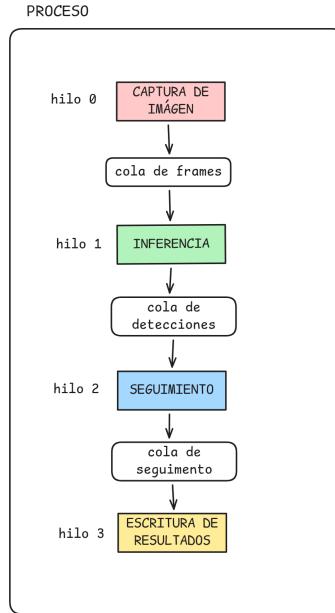


Figura 3.8: Diagrama de flujo del sistema segmentado en hilos.

La Figura 3.8 ilustra este enfoque, donde cada etapa opera en su propio hilo y se comunica mediante colas.

Considerando estas características, la segmentación basada en hilos puede ser beneficiosa para nuestro sistema. La etapa de captura de imágenes es intensiva en E/S. La etapa de inferencia, especialmente cuando se ejecuta en la GPU o DLA utilizando bibliotecas optimizadas como TensorRT, realiza la mayor parte de su trabajo en código nativo, liberando el GIL. La etapa de escritura también implica operaciones de E/S. Durante los momentos en que estas etapas liberan el GIL, otros hilos pueden progresar.

Aunque más complejo de implementar que el enfoque secuencial debido a la necesidad de sincronización y comunicación entre hilos, este modelo permite una mejor utilización de los recursos y mejora la capacidad de respuesta y el rendimiento general del sistema al solapar operaciones de E/S y cómputo intensivo (en GPU/DLA) con otras tareas.

3.5.3. Segmentación en procesos

La tercera opción es la segmentación del sistema en diferentes procesos. En este enfoque, cada etapa principal del sistema (captura, inferencia, seguimiento, escritura) se ejecuta en un proceso independiente. La comunicación y transferencia de datos entre estas etapas/procesos se gestiona mediante colas de mensajes (*Queue*). Estas colas provienen del módulo `multiprocessing`[36] de la librería estándar de Python y comparten la misma interfaz que las colas estándar del módulo `queue`, lo que asegura una transferencia eficiente y segura entre procesos (*process-safe*).

La principal ventaja de este enfoque es que cada etapa del sistema se ejecuta en su propio proceso independiente, lo que permite un mejor aprovechamiento de los recursos disponibles. Además, al estar cada etapa en su propio proceso, se evita el problema

del GIL (Global Interpreter Lock) presente en Python. Esto se debe a que cada proceso tiene su propio intérprete de Python y, por lo tanto, su propio GIL independiente. Como resultado, múltiples procesos pueden ejecutar código Python simultáneamente en diferentes núcleos de CPU, logrando un paralelismo real, a diferencia de los hilos dentro de un mismo proceso. Esto es especialmente beneficioso en sistemas con múltiples núcleos de CPU, donde cada proceso puede ejecutarse en un núcleo diferente, maximizando así el rendimiento del sistema.

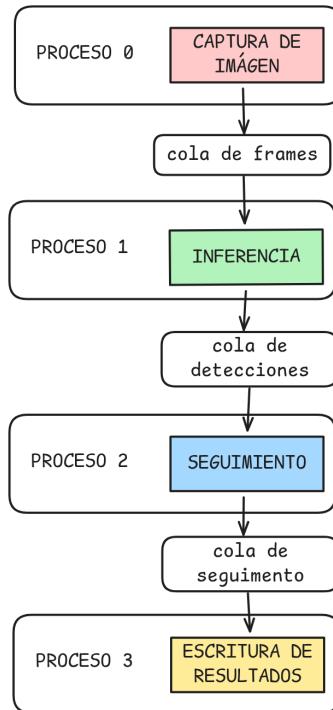


Figura 3.9: Diagrama de flujo del sistema segmentado en procesos.

La Figura 3.9 ilustra este enfoque, donde cada etapa opera en su propio proceso y se comunica mediante colas.

Sin embargo, este enfoque también presenta desventajas. La comunicación entre procesos es más costosa en términos de tiempo y recursos que la comunicación entre hilos dentro de un mismo proceso. Además, la gestión de memoria y el intercambio de datos entre procesos pueden ser más complejos, lo que puede aumentar la dificultad de implementación y mantenimiento del sistema.

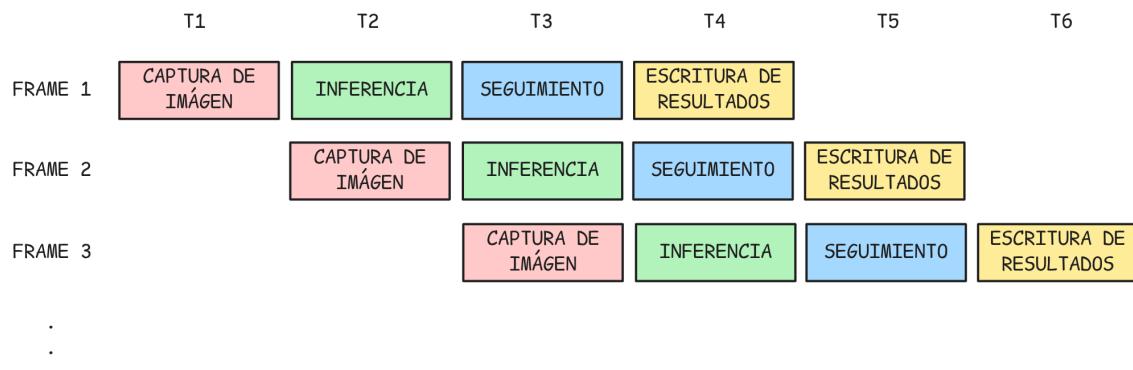


Figura 3.10: Diagrama de flujo del sistema segmentado.

La Figura 3.10 ilustra el flujo de datos en un sistema segmentado mediante procesos independientes, contrastando con el enfoque secuencial mostrado en la Figura 3.7. En este diseño segmentado, cada etapa principal (captura, inferencia, seguimiento, escritura) opera en su propio proceso, permitiendo la ejecución concurrente en diferentes núcleos de CPU si están disponibles.

Siguiendo la analogía con la arquitectura de un procesador, este enfoque se asemeja a un procesador segmentado (*pipelined*), donde diferentes instrucciones se encuentran en distintas fases de ejecución simultáneamente. Sin embargo, existe una diferencia fundamental: mientras que en un procesador segmentado todas las etapas avanzan sincronizadas por un ciclo de reloj común, determinado por la duración de la etapa más lenta, en nuestro sistema las etapas operan de forma asíncrona.

Cada etapa del sistema (captura, inferencia, seguimiento, escritura) tiene una duración variable y no necesariamente igual a las demás. Por ejemplo, la inferencia en la GPU puede ser mucho más rápida o lenta que la captura de imágenes o el seguimiento en la CPU. Las colas de mensajes actúan como buffers intermedios que desacoplan las etapas, permitiendo que cada una procese datos a su propio ritmo. Una etapa más rápida puede producir resultados que se acumulan en la cola de salida, mientras que una etapa más lenta consumirá datos de su cola de entrada cuando estén disponibles, esperando si la cola está vacía.

Esta asincronía, gestionada mediante colas, permite un mayor rendimiento (*throughput*) en comparación con el modelo estrictamente secuencial (Figura 3.7), donde cada etapa debe esperar a que la anterior finalice completamente. No obstante, si una etapa es significativamente más lenta que las demás, puede convertirse en un cuello de botella, haciendo que las colas anteriores se llenen y las posteriores permanezcan vacías, limitando el rendimiento general del sistema al ritmo de la etapa más lenta.

3.5.4. Segmentación heterogénea

La cuarta opción es la segmentación basada en computación heterogénea. Aprovechando la arquitectura heterogénea de la plataforma NVIDIA Jetson, esta opción de segmentación distribuye las tareas entre las diferentes unidades de procesamiento disponibles. La etapa de inferencia, supuestamente la más exigente computacionalmente, se descarga específicamente a los aceleradores de hardware: la GPU o uno de los DLAs (DLA0, DLA1) si están presentes en el módulo Jetson. Las demás etapas (captura, seguimiento y escritura) permanecen asignadas a la CPU.

Este enfoque permite una ejecución paralela real, donde la CPU gestiona el flujo de datos y la lógica de seguimiento mientras la GPU y/o las DLAs procesan simultáneamente los fotogramas para la detección de objetos. Para manejar los resultados que llegan de forma asíncrona desde estos aceleradores, a cada fotograma capturado se le asigna un identificador único. Esto garantiza que las detecciones se asocien correctamente con el fotograma original antes de pasar a la etapa de seguimiento, preservando así el orden temporal de la secuencia. La comunicación entre los procesos que se ejecutan en la CPU y aquellos que gestionan la inferencia en los aceleradores se realiza mediante las colas inter-proceso seguras (*process-safe queues*) descritas anteriormente.

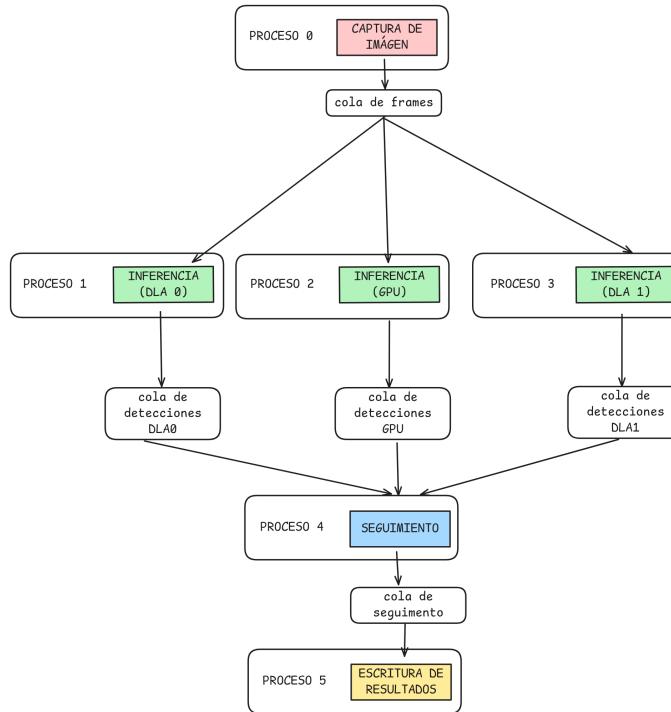


Figura 3.11: Diagrama de flujo del sistema segmentado en diferentes unidades de procesamiento.

La Figura 3.11 ilustra cómo la etapa de inferencia puede ejecutarse en paralelo en la GPU o DLA, comunicándose con la etapa de seguimiento (en CPU) a través de colas. Esta distribución optimiza el uso de los recursos especializados, acelerando significativamente el rendimiento general del sistema.

3.5.5. Segmentación basada en procesos con memoria compartida

La quinta opción de segmentación emplea procesos independientes como en la sección 3.5.3, pero busca optimizar la comunicación entre ellos utilizando memoria compartida como alternativa a las colas estándar del módulo `multiprocessing`. La transferencia de grandes volúmenes de datos, como los fotogramas de vídeo, puede volverse ineficiente con `multiprocessing`. Queue debido a la sobrecarga asociada a la serialización (*pickling*) y deserialización de objetos, así como a la posible copia de datos entre los espacios de memoria de los procesos a través de mecanismos subyacentes como pipes.

Para superar estas limitaciones, se hace uso de la librería estándar que ofrece Python `multiprocessing.shared_memory`[37]. Esta permite a múltiples procesos acceder directamente a la misma región de memoria.

Un proceso crea un bloque de memoria compartida, y otros procesos pueden adjuntarse a él usando su nombre único. Ambos pueden leer y escribir directamente en el *buffer* de memoria (`shm.buf`). Este acceso directo elimina los pasos de serialización/deserialización y las copias intermedias, resultando en una latencia mucho menor y un mayor ancho de banda, lo cual es especialmente beneficioso para datos grandes y estructurados como imágenes o *arrays* NumPy.

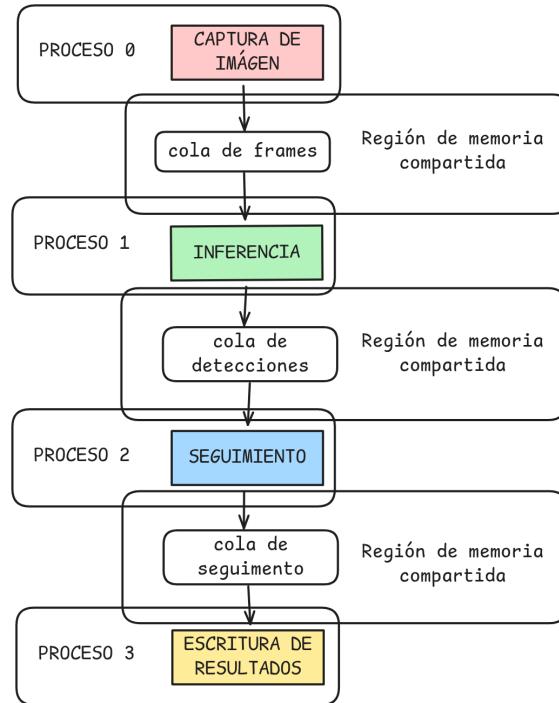


Figura 3.12: Diagrama de flujo del sistema segmentado en procesos con memoria compartida.

La Figura 3.12 ilustra este enfoque, donde cada etapa opera en su propio proceso y se comunica mediante memoria compartida. Este método permite una transferencia de datos más eficiente entre procesos, eliminando la necesidad de serialización y deserialización, lo que resulta en una latencia reducida y un mayor rendimiento general del sistema.

Sin embargo, la gestión directa de la memoria compartida requiere una implementación cuidadosa. Para facilitar su uso y gestionar el flujo de datos de manera estructurada, se ha implementado una capa de abstracción: un buffer circular. Este buffer opera sobre un bloque de memoria compartida preasignado y funciona como una cola de capacidad fija. Los datos se escriben en una posición (*tail*) y se leen desde otra (*head*). Cuando los índices alcanzan el final del *buffer*, vuelven al principio, permitiendo un uso continuo del espacio de memoria. La Figura 3.13 ilustra esta estructura conceptual.

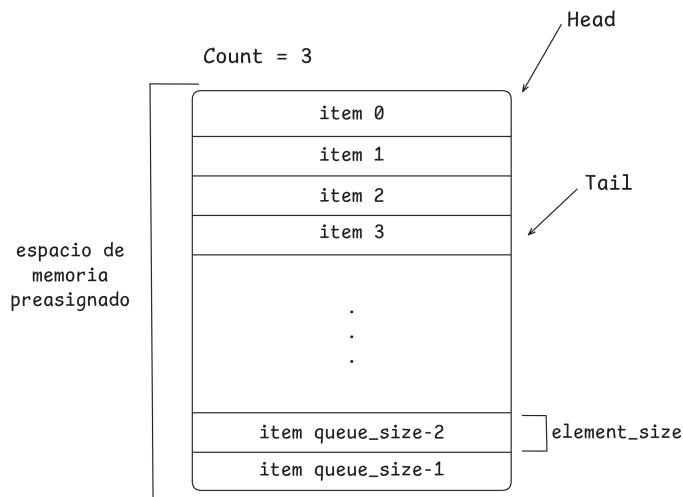


Figura 3.13: Ejemplo de buffer circular.

Dado que múltiples procesos acceden concurrentemente a la misma memoria a través del *buffer* circular, es crucial garantizar la coherencia de los datos y prevenir condiciones de carrera. A diferencia de `multiprocessing.Queue`, que gestiona la sincronización internamente, `shared_memory` no la proporciona automáticamente. Por lo tanto, el *buffer* circular implementado integra mecanismos de sincronización explícitos, como bloqueos (*Lock*) y variables de condición (*Semaphore*) del módulo `multiprocessing`. Estos controlan el acceso: un proceso productor que intente añadir datos a un buffer lleno se bloqueará hasta que un consumidor libere espacio, y viceversa, un consumidor que intente leer de un buffer vacío esperará. Este comportamiento asegura que no se pierdan datos y que las operaciones se realicen de forma segura.

La configuración de cada buffer circular requiere definir estáticamente su capacidad, el tamaño de memoria para cada elemento y un nombre único para la región de memoria compartida. Una limitación clave es la necesidad de preasignar la memoria. Una asignación incorrecta (demasiado grande o demasiado pequeña) puede llevar a desperdicio de recursos o a bloqueos frecuentes que limiten el rendimiento. Por ello, se requiere una calibración experimental para determinar los tamaños óptimos de buffer para cada enlace entre etapas, buscando el equilibrio entre uso eficiente de memoria y fluidez en el procesamiento. Además, el programador debe gestionar manualmente el ciclo de vida del bloque de memoria compartida (creación, cierre con `close()` y liberación final con `unlink()`).

En resumen, la segmentación basada en procesos con memoria compartida mediante un buffer circular busca ofrecer un rendimiento superior para la transferencia de grandes bloques de datos como fotogramas de vídeo, aunque esto implicaría una mayor complejidad en la implementación y gestión de la memoria.

CAPÍTULO 4

Evaluación de la solución

En este capítulo se presenta la evaluación del sistema propuesto, analizando su rendimiento y eficiencia en diferentes configuraciones. Se describen las metodologías de evaluación empleadas, las métricas de rendimiento consideradas y los resultados obtenidos en cada caso. El objetivo es proporcionar una visión clara de cómo el sistema se comporta bajo diversas condiciones y configuraciones, permitiendo identificar sus fortalezas y áreas de mejora.

4.1 Metodología de evaluación y métricas de rendimiento

Tras el desarrollo de la solución propuesta, se procede a analizar su rendimiento. La evaluación se basa en diferentes vídeos de prueba de 80 segundos (2400 fotogramas) que muestran un flujo de canicas, grabados a 640x640 píxeles y 30 fotogramas por segundo. Estos vídeos sirven como entrada estándar para el sistema, permitiendo medir diversas métricas de rendimiento según la configuración.

Para evaluar el rendimiento del sistema, se han implementado dos metodologías de ejecución:

1. **Ejecución a máxima capacidad:** Los fotogramas se entregan al sistema consecutivamente, tan pronto como el procesamiento del fotograma anterior ha concluido, utilizando colas intermedias que no descartan fotogramas. En este escenario, se mide el tiempo total que el sistema tarda en procesar la secuencia completa, lo que permite evaluar su rendimiento máximo (*throughput*) y compararlo con la duración real del vídeo.
2. **Simulación de entrada en tiempo real:** Los fotogramas del vídeo se suministran al sistema a una tasa fija de 30 fps (un fotograma cada 33.3 ms), emulando la entrada de una cámara. Si el sistema no logra procesar un fotograma dentro de este intervalo, dicho fotograma se descarta. Esta metodología permite cuantificar la capacidad del sistema para operar en tiempo real, midiendo el número de fotogramas procesados y perdidos.

Como métricas de rendimiento, se han considerado las siguientes:

- **Tasa de fotogramas por segundo (FPS):** Número de fotogramas completamente procesados por segundo. Esta métrica es fundamental para evaluar la capacidad del sistema para operar en tiempo real, ya que cuantifica directamente su velocidad de procesamiento. Un valor mayor o igual a 30 FPS generalmente indica que el sistema puede funcionar en tiempo real para vídeos estándar.

- **Tasa de fotogramas perdidos (LFPS):** Número de fotogramas descartados por segundo debido a la incapacidad del sistema para procesarlos dentro del intervalo temporal requerido. Esta métrica refleja la robustez del sistema bajo restricciones de tiempo real y su eficacia para mantener sincronización con la fuente de entrada. Un valor cercano a cero indica un rendimiento óptimo.
- **Potencia media consumida (W):** Potencia eléctrica media requerida por el sistema durante la ejecución, medida en vatios. Se obtiene promediando las lecturas de potencia instantánea registradas por las herramientas de monitorización hardware. Esta métrica es crucial para evaluar la viabilidad del sistema en entornos con restricciones energéticas, especialmente en aplicaciones *embedded* o en el *edge*.
- **Energía consumida (J):** Cantidad total de energía eléctrica utilizada por el sistema para procesar la secuencia completa, medida en julios. Se calcula integrando la potencia instantánea a lo largo del tiempo de ejecución. Esta métrica proporciona una perspectiva integral de la eficiencia energética y es fundamental para estimar costes operativos y requisitos de refrigeración en implementaciones a largo plazo.
- **Fotogramas por vatio (Frames/W):** Métrica compuesta que indica la eficiencia energética del procesamiento, cuantificando cuántos fotogramas se procesan por cada vatio de potencia consumida. Permite comparar directamente configuraciones con diferentes compromisos entre velocidad y consumo. Se calcula mediante la siguiente relación:

$$\text{FPS}/\text{W} = \frac{\text{Frames procesados}}{\text{Potencia (W)} \cdot \text{Tiempo (s)}} \quad (4.1)$$

donde

$$\text{Energía consumida (J)} = \text{Potencia (W)} \cdot \text{Tiempo (s)} \quad (4.2)$$

por lo tanto

$$\text{FPS}/\text{W} = \frac{\text{Frames procesados}}{\text{Energía consumida (J)}} \quad (4.3)$$

- **Speedup:** Factor de aceleración que cuantifica la mejora relativa en velocidad de procesamiento. Se calcula como el cociente entre el tiempo de ejecución de la configuración de referencia (generalmente la más lenta) y el tiempo de la configuración evaluada. Un valor de 2.0 indicaría que la configuración actual es exactamente dos veces más rápida que la referencia. Esta métrica permite evaluar objetivamente los beneficios de optimizaciones específicas.
- **Uso medio de la GPU (*GPU Average Utilization*):** Porcentaje medio de utilización de los recursos de la GPU durante el tiempo de ejecución. Este valor, obtenido mediante las herramientas de monitorización de NVIDIA, representa qué fracción de la capacidad computacional total de la GPU se aprovecha efectivamente. Una utilización cercana al 100 % sugiere un aprovechamiento óptimo del hardware, mientras que valores bajos podrían indicar cuellos de botella en otras partes del sistema o inefficiencias en la implementación.
- **Uso medio de la CPU (*CPU Average Utilization*):** Porcentaje medio de utilización de los núcleos de la CPU durante el tiempo de ejecución. Esta métrica es la media aritmética de todos los núcleos disponibles (entre 1 y 12, dependiendo de la configuración específica del dispositivo Jetson) y lo promedia. Permite identificar si las etapas del sistema que se ejecutan en CPU están equilibradas respecto a las que utilizan aceleradores hardware, y detectar posibles desbalances en la carga de trabajo entre las diferentes unidades de procesamiento.

4.2 Variación de la configuración del sistema

Para evaluar el rendimiento del sistema propuesto, se han realizado pruebas variando la configuración de las etapas del sistema. Estas pruebas se han llevado a cabo utilizando los vídeos de prueba descritos en la sección 4.2.1, con el objetivo de analizar cómo diferentes configuraciones afectan al rendimiento y a las métricas de eficiencia energética.

Para obtener estas métricas, se han utilizado herramientas de medición de potencia y energía, como el comando `tegrastats`[33] de NVIDIA, que proporciona información sobre el consumo de energía y la carga de la CPU y GPU. Esta herramienta permite monitorizar el rendimiento del sistema en tiempo real, proporcionando datos precisos sobre el uso de recursos y el consumo energético.

4.2.1. Cantidad de objetos

La cantidad de objetos presentes en el flujo de vídeo es un factor determinante en la evaluación del rendimiento del sistema, ya que influye directamente en la carga computacional de la etapa de seguimiento y escritura. Para analizar esta influencia de manera sistemática, se han realizado pruebas utilizando cuatro vídeos de prueba distintos, cada uno diseñado para representar un escenario de carga diferente:

1. **Vídeo 1: Carga baja y constante:** Un vídeo que mantiene una cantidad constante baja de 17 objetos en cada fotograma. Este escenario permite evaluar el rendimiento base del sistema bajo una carga predecible y reducida.
2. **Vídeo 2: Carga media y constante:** Un vídeo que presenta una cantidad constante media de 43 objetos por fotograma. Este escenario permite observar el rendimiento del sistema bajo una carga moderada.
3. **Vídeo 3: Carga alta y constante:** Un vídeo que presenta una cantidad constante media de 84 objetos por fotograma. Este escenario somete al sistema a una carga significativamente mayor, permitiendo identificar posibles cuellos de botella bajo condiciones de alta densidad de objetos.
4. **Vídeo 4: Carga variable:** Un vídeo donde la cantidad de objetos fluctúa dinámicamente a lo largo de su duración, variando entre un mínimo de 0 y un máximo de 180 objetos. Este escenario simula condiciones más realistas y complejas, donde el sistema debe adaptarse a cambios abruptos en la carga de trabajo.

Todos los vídeos de prueba se grabaron con una resolución de 640x640 píxeles y a una tasa de 30 fotogramas por segundo, sirviendo como entrada estándar para el sistema, lo que asegura la consistencia en las condiciones de captura.

Para el resto de la configuración del sistema durante estas pruebas específicas sobre la cantidad de objetos, se ha empleado la segmentación por procesos con memoria compartida. El modelo de detección de objetos utilizado fue YOLO11n, optimizado mediante NVIDIA TensorRT para su ejecución en la GPU.

Se configuró para operar con precisión FP16 y con el perfil de energía del dispositivo Jetson ajustado al modo de máxima potencia (MAXN). Esta configuración se seleccionó con el objetivo de maximizar el rendimiento del sistema y evaluar su capacidad de respuesta y robustez bajo condiciones exigentes impuestas por la variación en la densidad de objetos.

Las pruebas que se presentan a continuación se realizaron siguiendo la metodología de ejecución a máxima capacidad descrita en la sección 4.1, donde los fotogramas se

entregan al sistema consecutivamente, tan pronto como el procesamiento del fotograma anterior ha concluido, utilizando colas intermedias que no descartan fotogramas. En este escenario, se mide el tiempo total que el sistema tarda en procesar la secuencia completa, lo que permite evaluar su rendimiento máximo (*throughput*) y compararlo con la duración real del vídeo.

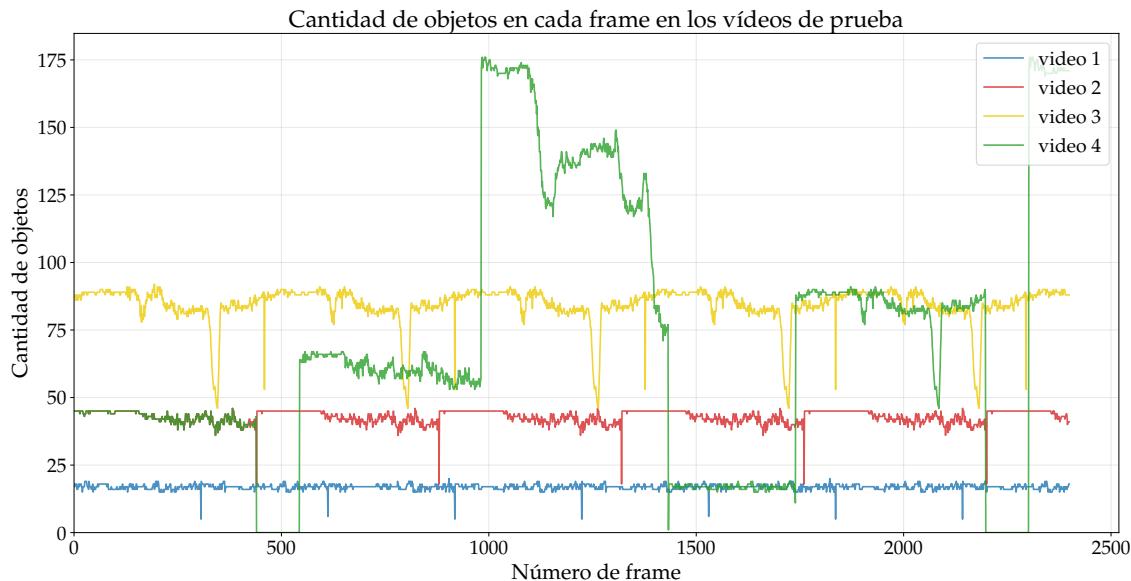


Figura 4.1: Cantidad de objetos en los vídeos de prueba.

La Figura 4.1 ilustra la cantidad de objetos presentes en cada fotograma para los cuatro vídeos de prueba utilizados: carga baja constante (17 objetos), carga media constante (43 objetos), carga alta constante (84 objetos) y carga variable (entre 0 y 180 objetos).

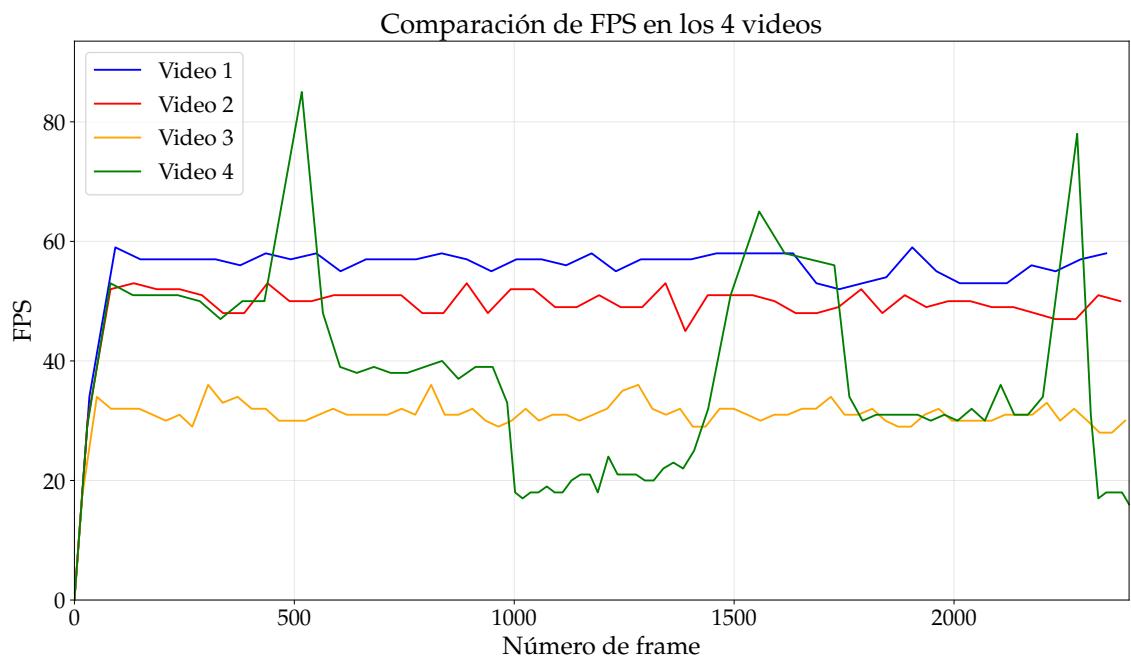


Figura 4.2: FPS por fotograma en función de la cantidad de objetos para los cuatro vídeos de prueba.

Al examinar la Figura 4.2, se evidencia una clara correlación entre el rendimiento del sistema, cuantificado en fotogramas por segundo (FPS), y la densidad de objetos presentes en el flujo de vídeo.

En el escenario de carga baja y constante (línea azul), el sistema mantiene un rendimiento consistente y elevado, oscilando entre 50-60 FPS. Para el vídeo de carga media y constante (línea roja), se observa una ligera degradación del rendimiento, que se estabiliza en el rango de 45-50 FPS. En condiciones de carga alta y constante (línea amarilla), el rendimiento experimenta una reducción más significativa, estableciéndose en un promedio de 30-40 FPS.

Esta progresiva disminución del rendimiento se alinea con las expectativas teóricas, dado que el incremento en la cantidad de objetos aumenta proporcionalmente la carga computacional en las etapas de seguimiento y escritura.

El escenario de carga variable (línea verde) revela un comportamiento particularmente informativo: el sistema alcanza picos superiores a 60 FPS durante intervalos con menos de 20 objetos, pero experimenta una pronunciada caída hasta aproximadamente 20 FPS cuando la densidad supera los 100 objetos. Esta marcada fluctuación en el rendimiento confirma la sensibilidad de determinadas etapas del sistema al volumen de objetos procesados simultáneamente, lo que resulta crucial para comprender sus limitaciones operativas en entornos dinámicos.

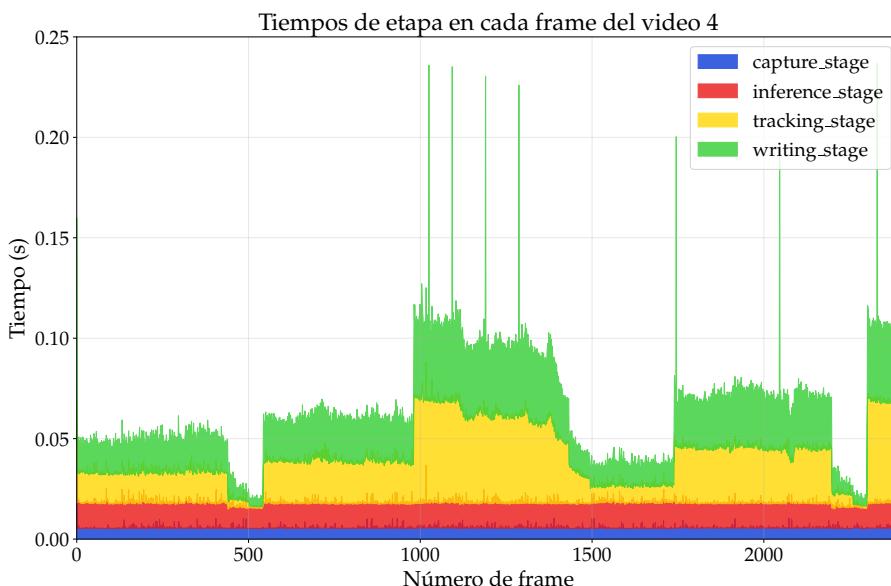


Figura 4.3: Ejecución temporal de las etapas del sistema durante el vídeo de prueba con carga variable.

Para comprender a fondo el impacto de esta variación en la cantidad de objetos, es esencial analizar cómo afecta a cada etapa dentro de la arquitectura segmentada del sistema. La Figura 4.3 presenta el tiempo de ejecución de cada etapa del sistema a lo largo del vídeo de prueba con carga variable. Se observa que la etapa de captura (línea azul) se mantiene constante, ya que sus operaciones (adquisición de fotogramas) no dependen del contenido de la imagen. De manera similar, la etapa de inferencia (línea roja), ejecutada en la GPU, muestra un tiempo de procesamiento relativamente estable. Aunque podría esperarse una ligera variación, el modelo YOLO procesa la imagen completa y su carga principal no escala linealmente con el número de objetos detectados una vez que la imagen está en la GPU.

Por el contrario, la etapa de seguimiento (línea amarilla), que se ejecuta en la CPU, muestra una clara correlación entre su tiempo de ejecución y la cantidad de objetos. A medida que aumenta el número de objetos (como se ve en la curva de carga variable de la Figura 4.1), el tiempo que tarda la etapa de seguimiento también aumenta significativamente. Esto se debe a que el algoritmo BYTETrack debe gestionar más trayectorias, realizar más comparaciones para la asociación de datos y actualizar más filtros de Kalman. Este comportamiento indica que la etapa de seguimiento puede convertirse en un cuello de botella cuando la densidad de objetos es alta.

Finalmente, la etapa de escritura (línea verde), también dependiente de la CPU, muestra un ligero incremento en su tiempo de ejecución a medida que aumenta el número de objetos. Esto es esperable, ya que debe procesar y registrar la información de un mayor número de detecciones y trayectorias.

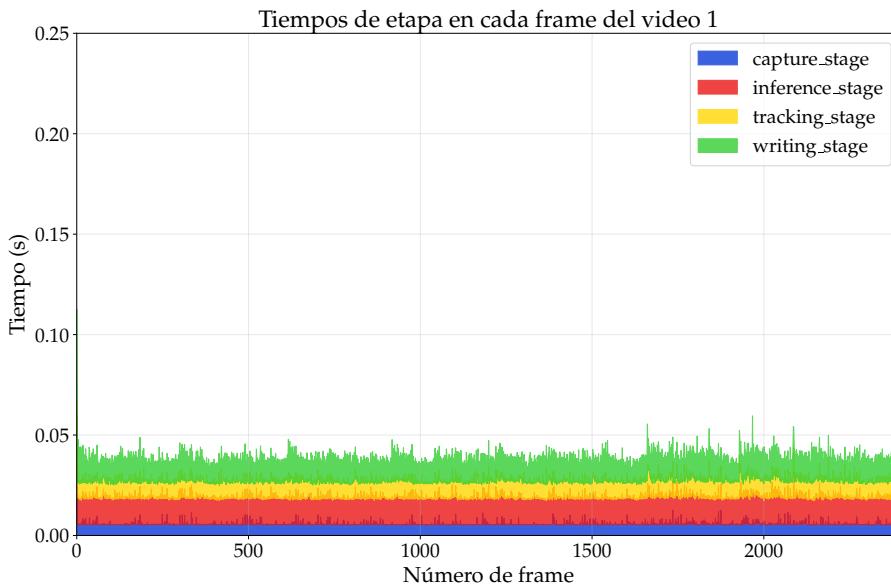


Figura 4.4: Ejecución temporal de las etapas del sistema durante el vídeo de prueba 1 (carga baja y constante).

Por otro lado, la Figura 4.4 muestra el tiempo de ejecución de cada etapa del sistema a lo largo del vídeo de prueba 1 (carga baja y constante).

En este caso, la etapa de seguimiento (línea amarilla) presenta un tiempo de ejecución bajo y estable, reflejando la menor carga de trabajo al procesar un número constante de objetos. Esto indica que el sistema gestiona correctamente escenarios de carga baja, sin generar cuellos de botella significativos en las etapas de seguimiento o escritura.

De forma similar, la etapa de escritura (línea verde) muestra tiempos de ejecución bajos y uniformes, lo que confirma la capacidad del sistema para mantener un rendimiento constante en condiciones de carga moderada.

La etapa de captura (línea azul) y la etapa de inferencia (línea roja) mantienen un rendimiento estable, similar al observado en el vídeo de carga variable, lo que indica que su rendimiento no se ve afectado por la cantidad de objetos presentes.

Ahora vamos a analizar el rendimiento del sistema con una simulación de entrada en tiempo real. En este caso, el sistema debe procesar cada fotograma en un tiempo máximo de 33.3 ms (30 fps). Si no logra hacerlo, el fotograma se descarta. Esta metodología permite evaluar la capacidad del sistema para operar en tiempo real y medir la tasa de fotogramas perdidos (LFPS).

Cant. Objetos	Frames Procesados	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	GPU Avg. (%) ↓	CPU Avg. (%) ↓
17	2,397	80,43	29,8	0,04	14,05	1,129	2,12	16,63	22,32
43	2,398	80,46	29,8	0,02	14,66	1,179	2,03	16,61	27,92
84	2,383	80,52	29,6	0,21	15,97	1,285	1,85	16,73	37,61
variable	2,208	80,64	27,38	2,38	14,8	1,192	2	15,05	30,7

Tabla 4.1: Resultados del experimento con distintas cantidades de objetos.

Los datos de la Tabla 4.1 revelan que el sistema logra procesar una cantidad considerable de fotogramas en tiempo real bajo distintas condiciones de carga, dando lugar a pensar que el sistema es capaz de manejar cargas altas de manera efectiva. Sin embargo, la Figura 4.5 muestra una relación más compleja; cuando la cantidad de objetos excede el umbral de 100, el rendimiento del sistema disminuye notablemente, llegando a generar tasas de fotogramas perdidos (LFPS) de hasta 10 frames por segundo en estos segmentos de alta densidad. Esta observación pone de manifiesto que las métricas promedio pueden resultar engañosas, ya que tienden a ocultar intervalos críticos de bajo rendimiento durante los picos de carga, precisamente cuando el sistema sería más necesario en aplicaciones industriales reales.

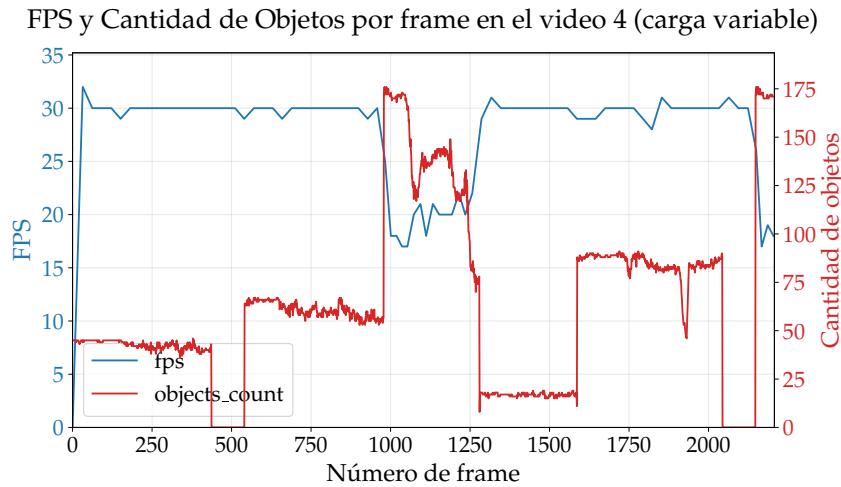


Figura 4.5: FPS y cantidad de objetos en función del tiempo para el vídeo de carga variable.

El análisis de la Tabla 4.1 también refleja una clara correlación inversa entre la cantidad de objetos y la eficiencia energética del sistema, expresada en fotogramas por vatio (Frames/W). A mayor densidad de objetos, menor es esta eficiencia, lo que indica que el sistema requiere proporcionalmente más energía para procesar escenas complejas. Este comportamiento se alinea con la expectativa de que una mayor carga de trabajo implica un incremento en el consumo de recursos computacionales y energéticos. Esta tendencia se confirma al examinar los niveles de utilización de hardware: el aumento en la cantidad de objetos se corresponde directamente con una mayor utilización media de CPU evidenciando el incremento gradual en la demanda de recursos del sistema conforme crece la complejidad de la escena analizada.

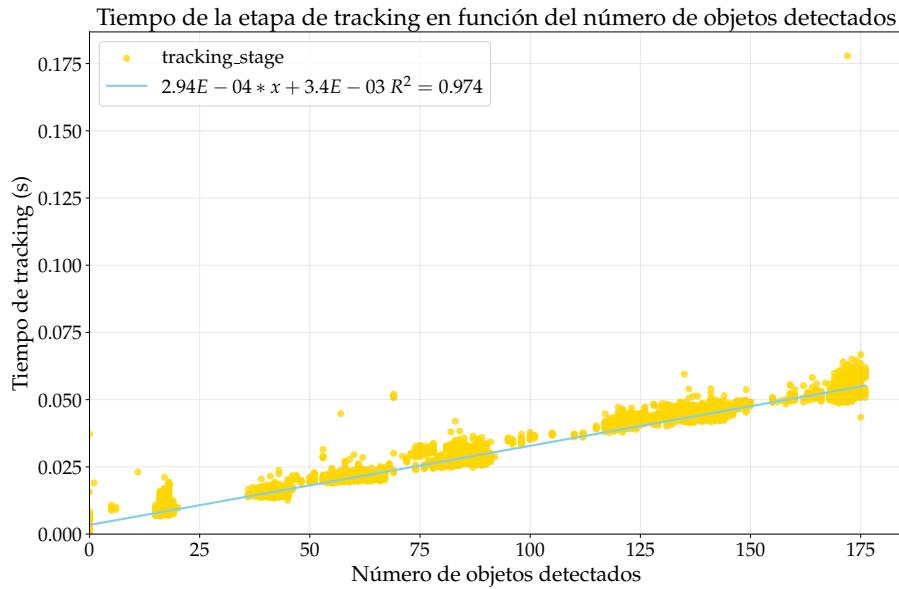


Figura 4.6: Tiempo de ejecución de la etapa de seguimiento en función de la cantidad de objetos.

Con los datos obtenidos, se ha creado una gráfica que muestra el tiempo de ejecución de la etapa de seguimiento en función de la cantidad de objetos presentes en el flujo de vídeo. La Figura 4.6 ilustra la relación empírica que describe esta dependencia. Mediante un análisis de regresión lineal, se ha obtenido la siguiente ecuación:

$$t_{seguimiento} = 2,94 \times 10^{-4} \times n_{objetos} + 3,4 \times 10^{-3} \quad (4.4)$$

donde $t_{seguimiento}$ es el tiempo de ejecución de la etapa de seguimiento en segundos y $n_{objetos}$ es la cantidad de objetos. Esta fórmula presenta un coeficiente de determinación (R^2) de 0.974, lo que indica un ajuste excelente al comportamiento observado.

Para determinar el umbral teórico donde el sistema dejaría de cumplir los requisitos de tiempo real (33.3 ms por fotograma), podemos despejar $n_{objetos}$ de la ecuación anterior:

$$2,94 \times 10^{-4} \times n_{objetos} + 3,4 \times 10^{-3} \leq 33,3 \times 10^{-3} \quad (4.5)$$

$$2,94 \times 10^{-4} \times n_{objetos} \leq 33,3 \times 10^{-3} - 3,4 \times 10^{-3} \quad (4.6)$$

$$2,94 \times 10^{-4} \times n_{objetos} \leq 29,9 \times 10^{-3} \quad (4.7)$$

$$n_{objetos} \leq \frac{29,9 \times 10^{-3}}{2,94 \times 10^{-4}} \quad (4.8)$$

$$n_{objetos} \leq 101,70 \quad (4.9)$$

Este cálculo teórico revela que el sistema alcanzaría su límite de procesamiento en tiempo real aproximadamente con 102 objetos, lo que coincide notablemente con los resultados experimentales mostrados en la Figura 4.5, donde se observaba una degradación significativa del rendimiento alrededor de los 100 objetos.

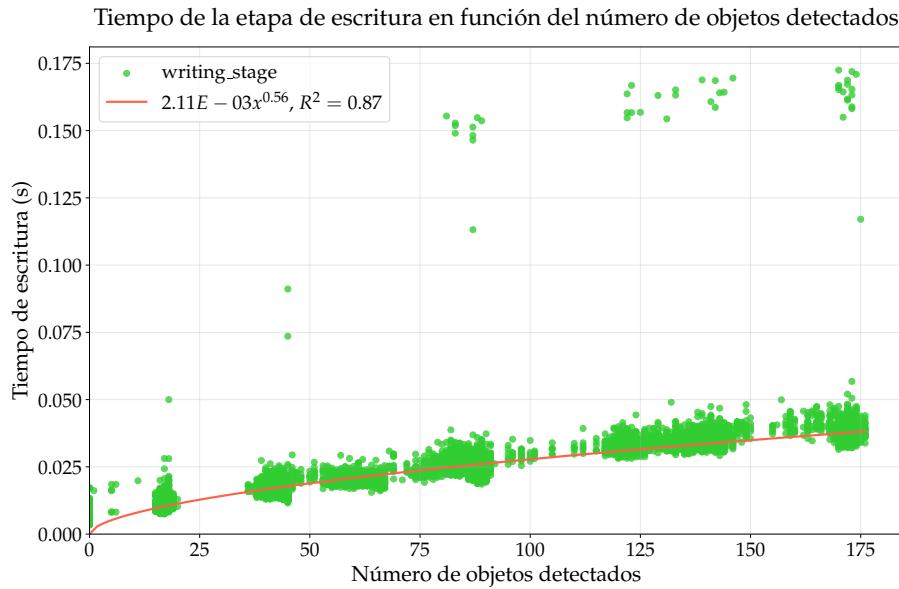


Figura 4.7: Tiempo de ejecución de la etapa de escritura en función de la cantidad de objetos.

Con los datos obtenidos, se ha creado una gráfica que muestra el tiempo de ejecución de la etapa de escritura en función de la cantidad de objetos presentes en el flujo de vídeo. La Figura 4.7 ilustra la relación empírica que describe esta dependencia. Mediante un análisis de regresión polinómica, se ha obtenido la siguiente ecuación:

$$t_{\text{escritura}} = 2,11 \times 10^{-3} \times n_{\text{objetos}}^{0,56} \quad (4.10)$$

donde $t_{\text{escritura}}$ es el tiempo de ejecución de la etapa de escritura en segundos y n_{objetos} es la cantidad de objetos. Esta fórmula presenta un coeficiente de determinación (R^2) de 0.874, lo que indica un ajuste aceptable al comportamiento observado. Para determinar el umbral teórico donde el sistema dejaría de cumplir los requisitos de tiempo real (33.3 ms por fotograma), podemos despejar n_{objetos} de la ecuación anterior:

$$2,11 \times 10^{-3} \times n_{\text{objetos}}^{0,56} \leq 33,3 \times 10^{-3} \quad (4.11)$$

$$n_{\text{objetos}}^{0,56} \leq \frac{33,3 \times 10^{-3}}{2,11 \times 10^{-3}} \quad (4.12)$$

$$n_{\text{objetos}} \leq \left(\frac{33,3}{2,11} \right)^{\frac{1}{0,56}} \quad (4.13)$$

$$n_{\text{objetos}} \leq 137,90 \quad (4.14)$$

Este cálculo teórico revela que el sistema alcanzaría su límite de procesamiento en tiempo real para la etapa de escritura aproximadamente con 138 objetos. No obstante, y aunque puedan existir algunos valores atípicos (*outliers*), la etapa de seguimiento (*tracking*) es la que marcará el cuello de botella del sistema frente a la de escritura.

Como conclusiones, se puede afirmar que el rendimiento del sistema es altamente sensible a la cantidad de objetos presentes en el flujo de vídeo. El umbral de aproximadamente 100 objetos constituye un límite crítico, donde la etapa de seguimiento comienza a experimentar un aumento significativo en su tiempo de ejecución, haciendo que el sistema no pueda cumplir con los requisitos de tiempo real.

4.2.2. Tipo de segmentación

Como se ha comentado en la sección 3.5, el sistema se puede segmentar de diferentes maneras. En esta sección se analizará el rendimiento de la solución propuesta variando el tipo de segmentación.

Repasando la sección 3.5, se han implementado cinco tipos de modos de segmentación:

1. **Ejecución secuencial** (subsección 3.5.1): Cada etapa del sistema se ejecuta de forma consecutiva. Este es el enfoque base y no se considera una segmentación propiamente dicha, sino el punto de partida para la comparación.
2. **Segmentación por hilos** (subsección 3.5.2): Cada etapa del sistema se ejecuta en un hilo independiente.
3. **Segmentación por procesos** (subsección 3.5.3): Cada etapa del sistema se ejecuta en un proceso independiente.
4. **Segmentación heterogénea** (subsección 3.5.4): La etapa de inferencia se descarga a los aceleradores de hardware (GPU o DLA).
5. **Segmentación por procesos con memoria compartida** (subsección 3.5.5): Cada etapa del sistema se ejecuta en un proceso independiente y se comunica mediante memoria compartida.

Para evaluar el rendimiento de cada tipo de segmentación, se han realizado pruebas utilizando el vídeo de 84 objetos (carga alta y constante) como entrada estándar. Este vídeo presenta una carga computacional significativa, lo que permite observar las diferencias de rendimiento entre los distintos tipos de segmentación. Para todas las pruebas se ha utilizado el modelo de detección de objetos YOLO11n, optimizado mediante NVIDIA TensorRT para su ejecución en la GPU (DLA0 y DLA1 en el caso de la segmentación heterogénea). Se configuró para operar con precisión FP16 y con el perfil de energía del dispositivo Jetson ajustado al modo de máxima potencia (MAXN).

Primero se va analizar el rendimiento de cada tipo de segmentación utilizando la metodología de ejecución a máxima capacidad, donde los fotogramas se entregan al sistema consecutivamente, tan pronto como el procesamiento del fotograma anterior ha concluido, utilizando colas intermedias que no descartan fotogramas. En este escenario, se mide el tiempo total que el sistema tarda en procesar la secuencia completa, lo que permite evaluar su rendimiento máximo (*throughput*) y compararlo con la duración real del vídeo.

Modo de Segmentación	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	Speedup ↑	GPU Avg. ¹↓	CPU Avg. ↓
secuencial	2,400	198,57	12,09	0	9,49	1,884,03	1,27	1	24,62	15,46
hilos	2,400	131,4	18,26	0	12,99	1,703,12	1,41	1,51	10,35	25,53
multiprocesos	2,400	105,16	22,82	0	14,27	1,500,58	1,6	1,89	12,84	33,65
multihardware	2,400	84,36	28,45	0	17,812	1,502,39	1,6	2,35	10	44,25
multiprocesos memoria compartida	2,400	80,27	29,9	0	16,55	1,320,25	1,82	2,47	17,12	40,1

Tabla 4.2: Resultados del experimento con distintos tipos de segmentación a máxima capacidad.

¹En la configuración de segmentación heterogénea, se utilizan la GPU y los dos núcleos DLA; sin embargo, solo se muestra la utilización de la GPU, ya que la herramienta de monitorización de NVIDIA no reporta la utilización media de la DLA.

A partir de los datos de la Tabla 4.2, se pueden extraer las siguientes conclusiones sobre el rendimiento de los diferentes enfoques de segmentación:

La ejecución secuencial, como era de esperar, ofrece el rendimiento más bajo, sirviendo como línea base para las comparaciones de speedup. Al no solapar ninguna operación, su capacidad de procesamiento es la más limitada.

La segmentación por hilos es la que ofrece el peor rendimiento después de la secuencial. Esto se debe fundamentalmente al Global Interpreter Lock (GIL) de Python, que impide el paralelismo real entre hilos dentro de un mismo proceso, limitando la capacidad de aprovechar múltiples núcleos de CPU.

La segmentación por procesos mejora esta situación al introducir paralelismo real, ya que cada proceso tiene su propio intérprete Python y, por ende, su propio GIL. Sin embargo, la comunicación estándar entre procesos mediante colas introduce una sobrecarga que impide extraer el máximo rendimiento.

La segmentación heterogénea, que asigna la inferencia a la GPU o DLA, ofrece un rendimiento notable y es capaz de procesar el vídeo en tiempo real. Si bien su velocidad podría ser comparable a la segmentación por procesos con memoria compartida, su consumo energético es superior debido al uso simultáneo de la GPU y, en su caso, la DLA. El rendimiento se ve limitado porque el modelo no se ejecuta en su totalidad en la DLA, lo que resulta en que una porción de la carga de trabajo recaiga sobre la GPU. Esto impide de que las unidades de procesamiento operen de forma completamente independiente, generando competencia por los recursos de la GPU.

Finalmente, la segmentación por procesos con memoria compartida es la que ofrece el mejor rendimiento global. Este enfoque combina el paralelismo real de los procesos con una comunicación entre ellos mucho más eficiente gracias al uso de memoria compartida, lo que minimiza la sobrecarga de la transferencia de datos. Esta combinación la posiciona como la opción más rápida, siendo también capaz de procesar el vídeo en tiempo real (30 fps).

Ahora se va a analizar el rendimiento de cada tipo de segmentación utilizando la metodología de ejecución en tiempo real, donde el sistema debe procesar cada fotograma en un tiempo máximo de 33.3 ms (30 fps). Si no logra hacerlo, el fotograma se descarta. Esta metodología permite evaluar la capacidad del sistema para operar en tiempo real y medir la tasa de fotogramas perdidos (LFPS).

Modo de Segmentación	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑
secuencial	800	80,13	9,98	19,97	9,95	796,97	1
hilos	1,468	80,98	18,13	11,51	13,27	1,106,05	1,33
multiprocesos	1,813	80,57	22,5	7,29	14,35	1,156,09	1,57
multihardware	2,353	80,81	29,12	0,58	17,36	1,402,1	1,68
multiprocesos memoria compartida	2,327	80,63	28,86	0,91	16,33	1,316,14	1,77

Tabla 4.3: Resultados del experimento con distintos tipos de segmentación a 30 fps.

Como se puede observar en la Tabla 4.3, el tipo de segmentación influye significativamente en el rendimiento del sistema. La ejecución secuencial es inviable para tiempo real, presentando la tasa más alta de fotogramas perdidos (LFPS) con 19.87 fps.

La segmentación por hilos, aunque mejora ligeramente, sigue siendo inviable para tiempo real, con una LFPS de 11.51, lo que confirma los resultados de la ejecución a máxima capacidad. Aunque la segmentación por procesos mejora este valor a 7.29 LFPS, tampoco alcanza los requisitos de tiempo real.

Por su parte, la segmentación heterogénea logra operar en tiempo real, pero a costa de un mayor consumo energético. Finalmente, la segmentación por procesos con memoria compartida destaca como la opción más eficiente, logrando el mejor rendimiento en tiempo real (LFPS de 0.91) con el menor consumo energético.

Como conclusión, la segmentación por procesos con memoria compartida es la más adecuada para aplicaciones en tiempo real, ya que combina un rendimiento óptimo con un consumo energético eficiente. Este enfoque permite al sistema operar de manera efectiva bajo condiciones de alta carga computacional, cumpliendo con los requisitos de tiempo real sin comprometer la eficiencia energética. La segmentación heterogénea, no resulta tan eficiente debido a la falta de independencia total entre la GPU y la DLA, lo que limita su capacidad de procesamiento en paralelo pero resulta una aproximación viable en un futuro si NVIDIA mejora la implementación de la DLA para que pueda ejecutar todo el modelo dentro de ella y así liberar a la GPU de esta carga de trabajo.

4.2.3. Modelo y talla

En esta sección se analizará el rendimiento de la solución propuesta variando el modelo de detección de objetos y sus diferentes tallas.

Repasando la tabla 3.1, se han empleado tres modelos de detección de objetos: YOLOv5, YOLOv8 y YOLOv11. Cada uno de estos modelos tiene diferentes tallas, que son versiones optimizadas para diferentes capacidades computacionales.

Para la evaluación del rendimiento de cada modelo y sus diferentes tallas, se estableció una configuración base común: los modelos de detección de objetos se optimizaron con NVIDIA TensorRT para su ejecución en la GPU, utilizando precisión FP16 y el perfil de energía MAXN del dispositivo Jetson AGX Xavier. Adicionalmente, se empleó la segmentación por procesos con memoria compartida, identificada en análisis previos como la más eficiente en rendimiento y consumo energético. Sobre esta configuración, las pruebas se realizaron utilizando como entrada estándar el vídeo de 84 objetos (carga alta y constante), lo que permite someter al sistema a una carga computacional elevada y así comparar objetivamente las distintas variantes de modelos y tallas.

Modelo	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	GPU Avg. (%) ↓	CPU Avg. (%) ↓	T. Prepro. (ms) ↓	T. Infer. (ms) ↓	T. Postpro. (ms) ↓
YOLOv5nu	2,400	84,44	28,42	0	14,847	1,253,41	1,91	42,73	43,73	2,8	13,84	4,11
YOLOv5mu	2,400	84	28,57	0	25,428	2,135,61	1,12	52,45	37,83	2	18,88	3,92
YOLOv8n	2,400	83,74	28,66	0	13,925	1,165,72	2,06	43,12	37,35	2,59	13,46	3,83
YOLOv8s	2,400	84,46	28,42	0	16,404	1,385,11	1,73	49,12	37,5	2,44	17,01	3,92
YOLO11n	2,400	82,28	29,17	0	14,112	1,160,85	2,07	44,4	37,9	2,55	13,94	3,81
YOLO11s	2,400	86,34	27,8	0	15,75	1,359,56	1,77	49,09	36,09	2,45	16,98	3,85
YOLO11m	2,400	84,66	28,35	0	26,125	2,211,51	1,09	55,3	37,61	1,99	19,33	3,8
YOLO11l	2,400	85,56	28,05	0	30,396	2,600,28	0,92	67,09	37,4	2,11	24,51	3,72

Tabla 4.4: Resultados del experimento con distintos modelos y tallas a máxima capacidad con un vídeo de carga alta y constante.

Al analizar los resultados presentados en la Tabla 4.4, se observa una notable similitud en el rendimiento de fotogramas por segundo (FPS) entre todos los modelos y sus respectivas tallas, manteniéndose en un rango aproximado de 28 ± 1 FPS. Esta homogeneidad en la velocidad de procesamiento se explica por la naturaleza del vídeo de prueba utilizado, que presenta una carga alta y constante de objetos. En este escenario, el rendimiento global del sistema se ve limitado predominantemente por la capacidad de las etapas posteriores a la inferencia (como el seguimiento y la escritura) para procesar la gran cantidad de objetos detectados, más que por la complejidad intrínseca o la velocidad de la etapa de inferencia en sí misma. Esto enmascara las diferencias potenciales en la velocidad de inferencia pura que podrían observarse bajo condiciones de menor carga de objetos.

No obstante, al considerar la eficiencia energética, medida en fotogramas por vatio (Frames/W), emergen diferencias significativas. Los modelos YOLO11n y YOLOv8n se destacan como las opciones más eficientes, logrando un mayor número de fotogramas procesados por cada vatio de energía consumida. En contraste, el modelo YOLO11l, a pesar de ofrecer una tasa de FPS comparable, presenta un consumo energético superior, lo que resulta en una menor eficiencia. Esta diferencia se atribuye al mayor número de parámetros del modelo YOLO11l, que inherentemente demanda más recursos computacionales y, por ende, energéticos para su ejecución.

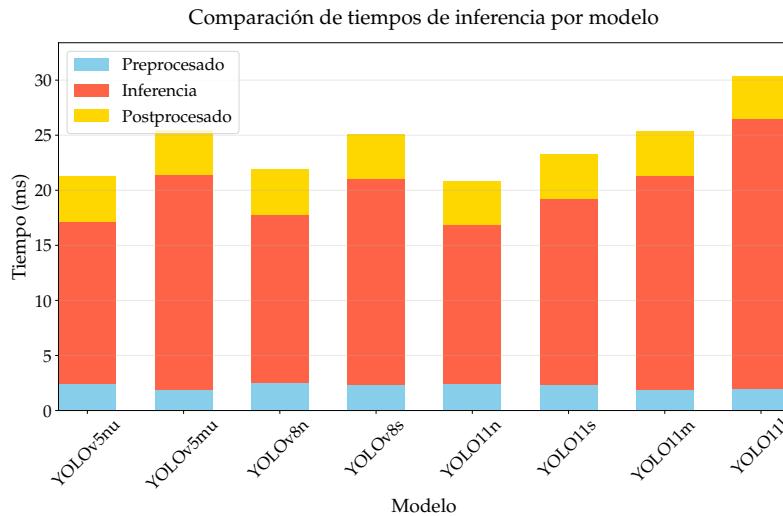


Figura 4.8: Tiempos de ejecución de la etapa de inferencia para los diferentes modelos y tallas.

Un análisis detallado de los tiempos de la etapa de inferencia, presentado en la Figura 4.8, revela que los modelos YOLOv8n y YOLO11n destacan por su menor tiempo de ejecución. En contraste, YOLOv5mu y YOLO11l exhiben tiempos más elevados, una tendencia que se alinea con el mayor consumo energético reportado en la Tabla 4.4.

Ahora se va a analizar el rendimiento de cada modelo y sus diferentes tallas utilizando la metodología de ejecución en tiempo real, donde el sistema debe procesar cada fotograma en un tiempo máximo de 33.3 ms (30 fps). Si no logra hacerlo, el fotograma se descarta. Esta metodología permite evaluar la capacidad del sistema para operar en tiempo real y medir la tasa de fotogramas perdidos (LFPS).

Modelo	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	mAP _{50–95} ↑
YOLOv5nu	2,303	80,58	28,58	1,2	14,018	1,129,33	2,04	0,759
YOLOv5mu	2,277	80,62	28,24	1,53	24,464	1,971,89	1,15	0,79
YOLOv8n	2,273	80,59	28,2	1,58	13,941	1,123,15	2,02	0,771
YOLOv8s	2,207	80,62	27,38	2,39	16,03	1,292,14	1,71	0,787
YOLO11n	2,297	80,63	28,49	1,28	12,883	1,118,99	2,05	0,761
YOLO11s	2,284	80,62	28,33	1,44	15,823	1,275,37	1,79	0,797
YOLO11m	2,296	80,6	28,49	1,29	26,085	2,102,2	1,09	0,797
YOLO11l	2,247	80,57	27,89	1,9	30,216	2,434,21	0,92	0,801

Tabla 4.5: Resultados del experimento con distintos modelos y tallas a 30 fps.

Los resultados de la Tabla 4.5 muestran una tendencia general similar a la observada en la Tabla 4.4 en términos de rendimiento. Sin embargo, la Tabla 4.5 permite un análisis más detallado de la tasa de fotogramas perdidos (LFPS) para cada combinación de modelo y talla. Se observa que el modelo YOLOv8s presenta la tasa de LFPS más alta,

alcanzando 2.39 LFPS. A pesar de esto, todas las variantes de modelos y tallas evaluadas mantienen una tasa de LFPS que podría considerarse aceptable para un sistema de visión artificial operando en tiempo real a 30 fps, ya que el sistema podría continuar funcionando de manera efectiva. El modelo YOLOv8n registra la tasa de LFPS más baja (1.2 LFPS), aunque las diferencias entre los modelos no son drásticas en este aspecto.

Para seleccionar el modelo y la talla más adecuados, es crucial considerar un equilibrio entre el consumo energético y la precisión del modelo. En el contexto de la tarea actual, que consiste en clasificar canicas de diferentes colores y con/sin defectos, la complejidad es relativamente baja. Como resultado, todos los modelos y tallas evaluados alcanzan niveles de precisión (mAP50-95) similares, tal como se evidencia en la última columna de la Tabla 4.5. Dada esta similitud en precisión, para esta aplicación específica, se podría priorizar la elección del modelo y talla que ofrezca el menor consumo energético. No obstante, en aplicaciones donde la precisión de detección sea un factor crítico, la selección debería inclinarse hacia el modelo y talla que demuestre el mayor rendimiento en dicha métrica, incluso si esto implica un mayor consumo energético.

4.2.4. Precisión numérica y acelerador de inferencia

En esta sección se analizará el rendimiento de la solución propuesta variando la precisión numérica (FP32, FP16, INT8) del modelo de detección de objetos y el acelerador de inferencia (CPU, GPU, DLA).

Para la realización de las pruebas, se utilizó el modelo YOLOv11n como base. Para las ejecuciones en GPU y DLA, dicho modelo fue optimizando mediante NVIDIA TensorRT, permitiendo así evaluar el rendimiento en distintas precisiones numéricas: FP32, FP16 e INT8.

En el caso de la ejecución sobre CPU, se empleó el modelo YOLOv11n sin optimización mediante TensorRT, ya que esta herramienta no es aplicable a dicho dispositivo. Es importante señalar que la conversión del modelo de PyTorch a TensorRT con precisión INT8 requiere un proceso de calibración utilizando un conjunto de datos específico. Para ello, se seleccionó un subconjunto de 100 imágenes extraídas del conjunto de entrenamiento original, con el objetivo de calibrar adecuadamente el modelo y maximizar su rendimiento durante la inferencia en INT8.

Acelerador	Precisión	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	GPU Avg. (%) ↓	CPU Avg. (%) ↓	mAP ₅₀₋₉₅ ↑
CPU	FP32	2,400	2,569,87	0,93	0	11,694	30,042,76	0,08	0	33,9	0,7757
GPU	FP32(sin exportar a TensorRT)	2,400	183,29	13,09	0	15,731	2,882,77	0,83	23,74	23,74	0,7757
GPU	FP32(exportado TensorRT)	2,400	84,44	28,42	0	19,371	1,635,31	1,47	29,08	37,81	0,7836
GPU	FP16	2,400	83,55	28,73	0	16,01	1,337,34	1,79	16,28	37,79	0,7831
GPU	INT8	2,400	84,4	28,44	0	14,469	1,222,37	1,96	12,86	37,75	0,7792
DLA0	FP16	2,400	86,77	27,66	0	15,597	1,439,73	1,67	6,2	35,93	0,7827
DLA0	INT8	2,400	89,11	26,93	0	14,346	1,278,04	1,88	10,71	37,46	0,7727

Tabla 4.6: Resultados del experimento con distintas precisiones y aceleradores a máxima capacidad con un vídeo de carga alta y constante.

Basándose en los resultados de la Tabla 4.6, se pueden extraer varias conclusiones sobre el impacto de la precisión numérica y el acelerador de inferencia.

En primer lugar, la ejecución del modelo en CPU resulta completamente inviable para aplicaciones en tiempo real. Como se aprecia en la tabla, la tasa de procesamiento es de tan solo 0.93 FPS, lo que demuestra la necesidad de aceleradores hardware para esta tarea.

Al comparar el rendimiento de la GPU y los DLAs, se observa que la DLA, en teoría diseñada para manejar cargas de trabajo específicas de inferencia de manera eficiente (como se repasó en la sección 2.2.3), ofrece un rendimiento similar al de la GPU en este caso.

Sin embargo, es crucial señalar que el modelo YOLO11n, tanto en precisión FP16 como en INT8, no se ejecuta completamente en la DLA, ya que algunas de sus operaciones no son compatibles con este dispositivo. Esto implica que la GPU también participa en el procesamiento. Este fenómeno se hace particularmente evidente en la precisión INT8, donde ninguna capa del modelo se ejecuta en la DLA, recayendo toda la carga en la GPU.

```
[TRT] [I] ----- Layers Running on DLA -----
[TRT] [I] [DlaLayer] {ForeignNode./model.0/conv/Conv.../model.10/m/m.0/attn/qkv/conv/Conv}
[TRT] [I] [DlaLayer] {ForeignNode./model.10/m/m.0/attn/Split.../model.10/m/m.0/attn/Transpose}
[TRT] [I] [DlaLayer] {ForeignNode./model.10/m/m.0/attn/Constant_1_output_0 + (Unnamed Layer* 146) [Shuffle] + /model.10/m/m.0/attn/Mul}
[TRT] [I] [DlaLayer] {ForeignNode./model.10/m/m.0/attn/Split_20.../Unnamed Layer* 149) [Shuffle] + /model.10/m/m.0/attn/Transpose_1]
[TRT] [I] [DlaLayer] {ForeignNode./model.10/m/m.0/attn/PeSplit}
[TRT] [I] [DlaLayer] {ForeignNode./model.10/m/m.0/attn/pe/conv/Conv.../model.23/Concat_2]
[TRT] [I] ----- Layers Running on GPU -----
[TRT] [I] [GpuLayer] SHUFFLE: /model.10/m/m.0/attn/Reshape
[TRT] [I] [GpuLayer] MATRIX_MULTIPLY: /model.10/m/m.0/attn/MatMul
[TRT] [I] [GpuLayer] SOFTMAX: /model.10/m/m.0/attn/Softmax
[TRT] [I] [GpuLayer] MATRIX_MULTIPLY: /model.10/m/m.0/attn/MatMul_1
[TRT] [I] [GpuLayer] SHUFFLE: /model.10/m/m.0/attn/Reshape_2
[TRT] [I] [GpuLayer] SHUFFLE: /model.23/Reshape
[TRT] [I] [GpuLayer] COPY: /model.23/Reshape_copy_output
[TRT] [I] [GpuLayer] SHUFFLE: /model.23/Reshape_1
[TRT] [I] [GpuLayer] COPY: /model.23/Reshape_1_copy_output
[TRT] [I] [GpuLayer] SHUFFLE: /model.23/Reshape_2
[TRT] [I] [GpuLayer] COPY: /model.23/Reshape_2_copy_output
[TRT] [I] [GpuLayer] SHUFFLE: /model.23/dfl/Reshape + /model.23/dfl/Transpose
[TRT] [I] [GpuLayer] SOFTMAX: /model.23/dfl/Softmax
[TRT] [I] [GpuLayer] CONVOLUTION: /model.23/dfl/conv/Conv
[TRT] [I] [GpuLayer] CONSTANT: scale constant of /model.23/Sub_1
[TRT] [I] [GpuLayer] CONSTANT: scale constant of /model.23/Sub
[TRT] [I] [GpuLayer] SHUFFLE: /model.23/dfl/Reshape_1
[TRT] [I] [GpuLayer] CONSTANT: /model.23/Constant_9_output_0
[TRT] [I] [GpuLayer] CONSTANT: /model.23/Constant_10_output_0
[TRT] [I] [GpuLayer] ELEMENTWISE: PWN(scale eltwise of /model.23/Sub, /model.23/Sub)
[TRT] [I] [GpuLayer] ELEMENTWISE: /model.23/Add_1
[TRT] [I] [GpuLayer] ELEMENTWISE: PWN(scale eltwise of /model.23/Sub_1, /model.23/Sub_1)
[TRT] [I] [GpuLayer] ELEMENTWISE: PWN(/model.23/Constant_11_output_0 + (Unnamed Layer* 386) [Shuffle], PWN(/model.23/Add_2, /model.23/Div_1))
[TRT] [I] [GpuLayer] COPY: /model.23/Div_1_output_0_copy
[TRT] [I] [GpuLayer] CONSTANT: /model.23/Constant_12_output_0 + (Unnamed Layer* 390) [Shuffle]
[TRT] [I] [GpuLayer] ELEMENTWISE: PWN(/model.23/Sigmoid)
[TRT] [I] [GpuLayer] ELEMENTWISE: /model.23/Mul_2
[TRT] [I] [GpuLayer] COPY: /model.23/Mul_2_output_0_copy
```

Figura 4.9: Exportación del modelo YOLO11n con TensorRT a FP16 para su ejecución en la DLA.

En la precisión FP16, como se ilustra en la Figura 4.9, la DLA se encarga de 6 de las 34 operaciones/capas del modelo, mientras que la GPU procesa las 28 restantes. En esencia, la DLA no logra descargar completamente a la GPU, lo que limita su capacidad para operar de forma independiente y optimizar el rendimiento general del sistema. Si bien en modelos más simples la DLA podría asumir una mayor proporción de las operaciones, la complejidad del modelo YOLO11n (y de forma similar para YOLOv5 y YOLOv8) impide que la DLA asuma una carga de trabajo más significativa.

Considerando la ejecución en GPU, la optimización con TensorRT, incluso manteniendo la precisión FP32, produce una notable aceleración en comparación con la ejecución del modelo sin optimizar en GPU. El modelo optimizado para GPU con TensorRT en FP32 alcanza un speedup de $\frac{183,29}{84,44} = 2,17 \times$ frente a la GPU. Esto demuestra que TensorRT mejora drásticamente el rendimiento mediante optimizaciones como la fusión de capas y la eliminación de operaciones redundantes, resultando en una ejecución más eficiente.

En cuanto a las diferentes precisiones numéricas en la GPU (todas optimizadas con TensorRT), los resultados muestran un rendimiento en FPS similar entre FP32, FP16 e INT8.

Finalmente, los resultados obtenidos indican que la precisión de validación se mantiene prácticamente constante entre todas las variantes del modelo. Tal como se observa en la Tabla 4.6, el valor de mAP50-95 es de 0.77 en todos los casos, lo que sugiere que la reducción de la precisión numérica no afecta de manera significativa al rendimiento del modelo. Esta estabilidad en los resultados puede deberse a la relativa simplicidad de la tarea de detección de objetos planteada en este experimento.

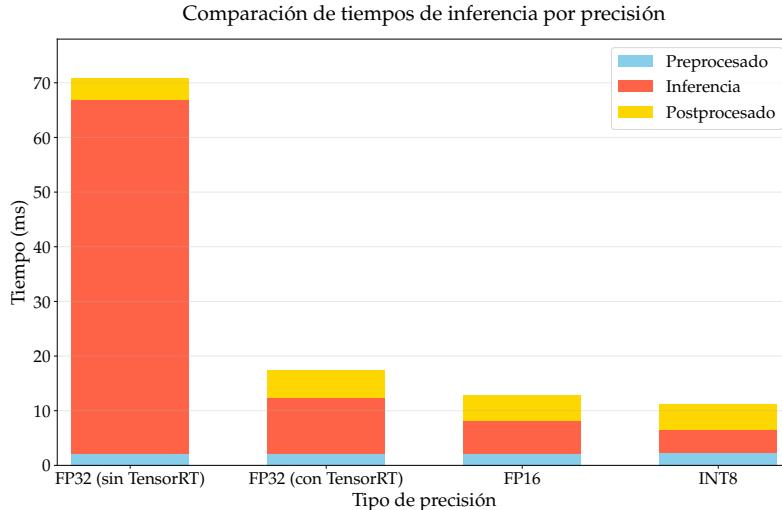


Figura 4.10: Tiempos de ejecución de la etapa de inferencia para las diferentes precisiones en GPU con TensorRT.

Un análisis más detallado de los tiempos de la etapa de inferencia, como se observa en la Figura 4.10, revela una reducción en el tiempo de inferencia al utilizar precisiones más bajas. Aunque el impacto en los FPS totales del sistema es limitado por otros factores del *pipeline* en este escenario de alta carga de objetos, es importante destacar que las precisiones FP16 e INT8 logran un menor consumo energético. Esto se debe a que estas precisiones reducidas requieren menos recursos computacionales, lo que se traduce en una mayor eficiencia energética. En particular, la precisión INT8 se presenta como la opción más eficiente en términos de consumo energético, alcanzando un rendimiento de 1.96 Frames/W.

Acelerador	Precisión	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	GPU Avg. (%) ↓	CPU Avg. (%) ↓	mAP ₅₀₋₉₅ ↑
CPU	FP32	65	81,97	0,79	28,49	12,988	1,064,25	0,06	0	49,37	0,7757
GPU	FP32(sin exportar a TensorRT)	1,006	80,59	12,48	17,3	15,829	1,275,39	0,79	22,56	30,97	0,7757
GPU	FP32(exportado TensorRT)	2,292	80,62	28,43	1,34	19,273	1,553,41	1,48	28,98	37,25	0,7836
GPU	FP16	2,302	80,58	28,57	1,22	15,769	1,270,4	1,81	16,2	36,98	0,7831
GPU	INT8	2,279	80,06	28,47	1,51	14,32	1,153,87	1,98	12,09	36,72	0,7792
DLA0	FP16	2,229	80,56	27,67	2,12	16,616	1,338,17	1,67	6,09	36,03	0,7827
DLA0	INT8	2,137	80,62	26,51	3,26	14,445	1,164,98	1,83	10,94	37,75	0,7727

Tabla 4.7: Resultados del experimento con distintas precisiones y aceleradores a 30 fps con un vídeo de carga alta y constante.

En la Tabla 4.7, se observan las ejecuciones en tiempo real del modelo con diferentes precisiones y aceleradores. Observando los resultados, como se ha mencionado anteriormente, la ejecución del modelo en CPU es completamente inviable para tiempo real, con una tasa de 0.79 FPS y 28.49 LFPS.

En cuanto a la GPU, todas las precisiones (FP32, FP16 e INT8 con TensorRT) logran operar en tiempo real, con tasas de FPS de 28.42, 28.73 y 28.44 respectivamente, lo que indica que el sistema puede procesar los fotogramas a la velocidad requerida sin perder ninguno.

La combinación de la GPU con la precisión INT8 es la más eficiente en términos de consumo energético, alcanzando 1.96 Frames/W.

4.2.5. Modo de energía y cores de la CPU

En esta sección, se analiza el impacto de la configuración energética del dispositivo NVIDIA Jetson AGX Xavier en el rendimiento del sistema. Se exploran distintos perfiles de energía predefinidos (10W, 15W, 30W y MAXN), los cuales ajustan tanto el número de núcleos de CPU activos como su frecuencia máxima de operación. El objetivo es evaluar cómo estas configuraciones influyen en la velocidad de procesamiento (FPS), el consumo energético y, por ende, la eficiencia energética global del sistema.

Para la realización de estas pruebas, se ha utilizado el modelo YOLO11n con la precisión FP16 optimizado con TensorRT para su ejecución en la GPU, un video de 84 objetos (carga alta y constante) y la segmentación por procesos con memoria compartida.

Modo de Energía	Núcleos de CPU	Frecuencia (GHz)	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	GPU Avg. (%) ↓	CPU Avg. (%) ↓
10W	2	1,2	2,400	343,03	7	0	5,518	1,903,73	1,26	22,29	23,73
15W	4	1,2	2,400	174,44	13,76	0	6,897	1,199,61	2	12,64	39,22
30W	2	2,11	2,400	179,66	13,36	0	8,621	1,548,38	1,55	10,34	23,18
30W	4	1,8	2,400	119,21	20,13	0	9,535	1,136,27	2,11	17,94	36,89
30W	6	1,41	2,400	135,58	17,7	0	8,398	1,139,58	2,11	38,51	38,81
30W	8	1,2	2,400	163,04	14,72	0	7,95	1,295,48	1,85	16,2	38,35
MAXN	8	2,23	2,400	89,92	26,69	0	15,676	1,409,25	1,7	15,74	37,63

Tabla 4.8: Resultados del experimento con distintos modelos y tallas a máxima capacidad con un vídeo de carga alta y constante.

Observando los resultados de la Tabla 4.8, se aprecia una clara influencia del perfil de energía y del número de núcleos de CPU activos en el rendimiento del sistema. En general, al aumentar el perfil de energía y la cantidad de núcleos habilitados, se observa una mejora en la tasa de fotogramas por segundo (FPS). Sin embargo, el perfil de energía de 30W presenta un comportamiento particular: la configuración óptima se alcanza con 4 núcleos de CPU. Esto se debe a que esta combinación ofrece la frecuencia máxima del procesador más alta (1.8 GHz) en comparación con las configuraciones de 6 y 8 núcleos bajo el mismo perfil (1.41 GHz y 1.2 GHz, respectivamente). Aunque la configuración de 30W con 4 núcleos exhibe el mayor consumo de potencia medio (9.535 W), resulta en el menor consumo de energía total (1136.27 J) debido a la reducción en el tiempo de ejecución en comparación con las configuraciones de 6 y 8 núcleos bajo el mismo perfil. Esto subraya la importancia de considerar tanto el consumo de potencia instantáneo como la duración total de la ejecución al evaluar la eficiencia energética del sistema.

Para evaluar el rendimiento en condiciones de tiempo real, se utilizó el mismo vídeo de 84 objetos (carga alta y constante) y la segmentación por procesos con memoria compartida. Se estableció un límite de 33.3 ms por fotograma (30 FPS) para evaluar la la capacidad del sistema para operar en tiempo real y medir la tasa de fotogramas perdidos (LFPS). Es importante señalar que, en los perfiles de energía más bajos, no se logró simular el tiempo real, lo que resultó en tiempos de ejecución superiores a los 80 segundos del vídeo a 30 FPS.

Modo de Energía	Núcleos de CPU	Frecuencia (GHz)	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	GPU Avg. (%) ↓	CPU Avg. (%) ↓
10W	2	1,2	537	118,74	4,52	15,69	5,219	619,43	0,87	11,35	24,44
15W	4	1,2	959	83,1	11,54	17,34	6,806	565,46	1,7	11,5	41,48
30W	2	2,11	1,017	88,48	11,49	15,63	8,515	752,92	1,35	10,98	23,65
30W	4	1,8	1,546	81,76	18,91	10,45	9,488	775,38	1,99	15,37	37,87
30W	6	1,41	1,309	81,51	16,06	13,38	8,21	668,79	1,96	36,97	29,62
30W	8	1,2	1,102	82,27	13,39	15,78	7,855	646,02	1,71	10,4	40,47
MAXN	8	2,23	2,141	80,73	26,52	3,21	15,653	1,263,31	1,69	14,92	37,79

Tabla 4.9: Resultados del experimento con distintos modelos y tallas a 30 fps con un vídeo de carga alta y constante.

La Tabla 4.9 presenta los resultados de la simulación en tiempo real. Como se mencionó anteriormente, los perfiles de energía más bajos no lograron simular el tiempo real, evidenciado por tiempos de ejecución superiores a la duración del vídeo de prueba. De todos los perfiles de energía evaluados, únicamente la configuración de máximas prestaciones (MAXN) con 8 núcleos de CPU activos logró operar a una tasa de 26.69 FPS, lo que indica que el sistema puede procesar los fotogramas a una velocidad cercana a la requerida, aunque con cierta pérdida de fotogramas.

4.2.6. Dispositivos Jetson

En esta sección se analizará el rendimiento de la solución propuesta en diferentes dispositivos Jetson, específicamente en el Jetson Orin Nano, Jetson AGX Xavier y Jetson AGX Orin. Se utilizará el modelo YOLO11n con la precisión FP16 optimizado con TensorRT para su ejecución en la GPU, un video de 84 objetos (carga alta y constante) y la segmentación por procesos con memoria compartida. Repasando lo visto en la sección 2.2.3, los dispositivos Jetson evaluados pertenecen a dos generaciones distintas: la Jetson AGX Xavier, que forma parte de una generación anterior, y los modelos Jetson Orin Nano y Jetson AGX Orin, que pertenecen a la generación más reciente de dispositivos Jetson.

La Jetson AGX Xavier es un dispositivo de la generación anterior que cuenta con 8 procesadores ARM Carmel v8.2 a 2.2GHz, una GPU NVIDIA con 512 núcleos Volta y 64 Tensor Cores, junto con 32GB de memoria LPDDR4x. Además, incluye 2 NVIDIA DLAs v1.

Por otro lado, el Jetson Orin Nano, perteneciente a la generación más reciente, está equipado con 6 procesadores ARM Cortex-A78AE a 1.7GHz, una GPU NVIDIA con 1024 núcleos Ampere y 32 Tensor Cores, y 8GB de memoria LPDDR5. Aunque es el modelo más compacto de la serie Orin, ofrece un rendimiento notable para aplicaciones de IA en el *edge*.

Finalmente, la Jetson AGX Orin, también de la generación más reciente, es el modelo más avanzado de los tres. Cuenta con 12 procesadores ARM Cortex-A78AE a 2.2GHz, una GPU NVIDIA con 2048 núcleos Ampere y 64 Tensor Cores, y 64GB de memoria LPDDR5. Además, incluye 2 NVIDIA DLAs v2, que representan una mejora respecto a los DLA v1 de la Jetson AGX Xavier, proporcionando mayor eficiencia y capacidad para tareas de inferencia en redes neuronales profundas.

Dispositivo	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	GPU Avg. (%) ↓	CPU Avg. (%) ↓	T. Prepro. (ms) ↓	T. Infer. (ms) ↓	T. Postpro. (ms) ↓
Jetson AGX Xavier	2,400	78,06	30,75	0	16,47	1,285,27	1,87	17,72	40,08	2,14	5,77	3,8
Jetson Orin Nano	2,400	54,53	44,01	0	7,568	412,49	5,82	19,97	53,38	1,76	4,44	3,96
Jetson AGX Orin	2,400	42,53	56,43	0	13,029	553,97	4,33	21,32	26,93	1,32	2,51	2,89

Tabla 4.10: Resultados del experimento con distintos dispositivos Jetson a máxima capacidad con un vídeo de carga alta y constante.

Analizando los resultados de la Tabla 4.10, se observa que todos los dispositivos Jetson logran procesar el vídeo de 84 objetos (carga alta y constante) a una tasa de FPS superior a 30, lo que indica que posiblemente pueden operar en tiempo real. Sin embargo, el rendimiento varía significativamente entre los dispositivos. El que mayor rendimiento ofrece es el Jetson AGX Orin, con una tasa de FPS de 56.43, seguido por el Jetson Orin Nano con 44.01 FPS, y finalmente el Jetson AGX Xavier con 30.75 FPS. Esta diferencia en el rendimiento se debe a las especificaciones de hardware de cada dispositivo, como la cantidad de núcleos de CPU, la potencia de la GPU.

De manera sorprendente, el Jetson Orin Nano, a pesar de ser un dispositivo más compacto y de menor potencia que el Jetson AGX Orin, logra un rendimiento notablemente

alto, superando al Jetson AGX Xavier y siendo el dispositivo más eficiente en términos de consumo energético, alcanzando 5.82 Frames/W frente a los 4.33 Frames/W del Jetson AGX Orin.

Dispositivo	Frames Procesados ↑	Tiempo Ejec. (s) ↓	FPS ↑	LFPS ↓	Potencia media (W) ↓	Energía (J) ↓	Frames/W ↑	GPU Avg. (%) ↓	CPU Avg. (%) ↓	T. Prepro. (ms) ↓	T. Infer. (ms) ↓	T. Postpro. (ms) ↓
Jetson AGX Xavier	2,383	80,52	29,6	0	15,965	1,285,37	1,85	16,73	37,61	2,09	5,73	3,78
Jetson Orin Nano	2,400	80,33	29,88	0	6,324	507,68	4,73	13,04	34,68	1,66	4,41	3,77
Jetson AGX Orin	2,399	80,28	29,88	0	9,581	768,75	3,12	12,69	14	1,34	2,48	2,85

Tabla 4.11: Resultados del experimento con distintos dispositivos Jetson a 30 fps con un vídeo de carga alta y constante.

Los resultados de la Tabla 4.11 reflejan una tendencia similar a la observada en la Tabla 4.10. Todos los dispositivos evaluados tienen la capacidad de operar en tiempo real. Sin embargo, para determinar cuál es el más adecuado para la solución propuesta, es necesario considerar no solo el rendimiento en FPS, sino también el consumo energético y la eficiencia.

En este contexto, el Jetson Orin Nano destaca como el dispositivo más eficiente, alcanzando una tasa de 4.73 Frames/W en condiciones de tiempo real, superando a los otros dos dispositivos. Además, su precio es significativamente más bajo, siendo de 300€, en comparación con los 2400€ del Jetson AGX Orin. Por lo tanto, para este caso específico, donde se utiliza un modelo de detección de objetos relativamente sencillo y se requiere un rendimiento en tiempo real, el Jetson Orin Nano se presenta como la opción más adecuada. Su principal limitación radica en la memoria, que es de 8GB LPDDR5, en comparación con los 64GB LPDDR5 del Jetson AGX Orin y los 32GB LPDDR4x del Jetson AGX Xavier.

Sin embargo, si la aplicación requiere un mayor rendimiento para modelos más complejos o tareas más exigentes, el Jetson AGX Orin sería la elección ideal debido a su superior capacidad de procesamiento.

4.3 Evaluación del seguimiento de objetos

Tras implementar la solución propuesta, se evaluó su capacidad para realizar un seguimiento preciso de los objetos detectados. Este análisis se centró en determinar cómo el sistema mantiene la coherencia en la identificación de los objetos a lo largo del tiempo, un aspecto esencial para aplicaciones como la inspección de calidad y el conteo automatizado, donde la trayectoria y la identidad de cada elemento son cruciales.

Para medir el rendimiento de manera exhaustiva, se recurrió a un conjunto de métricas estándar en el campo del seguimiento de objetos múltiples (MOT), tal como se describió en la subsección 2.3.3. Estas métricas, incluyendo MOTA para la precisión global del seguimiento, MOTP para la exactitud en la localización, IDF1 para la consistencia en la asignación de identidades, y HOTA para una visión equilibrada de detección y asociación, ofrecen una perspectiva multidimensional del comportamiento del sistema.

El experimento se llevó a cabo utilizando un vídeo de prueba compuesto por 600 fotogramas. Este vídeo fue cuidadosamente etiquetado de forma manual para generar una referencia precisa (*ground truth*), contra la cual se compararon las salidas del sistema. En esta evaluación, el modelo YOLO11n, optimizado con NVIDIA TensorRT y configurado para operar con precisión FP16, se encargó de la detección inicial de objetos en cada fotograma. Posteriormente, el algoritmo BYTETrack[43] gestionó la tarea de seguimiento, asociando las detecciones a lo largo de la secuencia para construir y mantener las trayectorias de los objetos.

MOTA (%)	MOTP (%)	IDF1 (%)	HOTA (%)
81.8	80.4	90.2	71.7

Tabla 4.12: Resultados de la evaluación de métricas de seguimiento de objetos

Los resultados obtenidos fueron notablemente positivos y validan la robustez del enfoque implementado. El sistema alcanzó un MOTA del 81.8 %, lo que indica una alta precisión general en el seguimiento, minimizando errores como falsos positivos, detecciones omitidas o cambios incorrectos de identidad.

En términos de localización, el MOTP fue del 80.4 %, reflejando que las cajas delimitadoras predichas por el sistema se alinearon con gran exactitud con las posiciones reales de los objetos, un factor importante para cualquier análisis espacial posterior.

De particular relevancia para la aplicación objetivo, el sistema demostró una excelente consistencia en la asignación de identidades, logrando un IDF1 del 90.2 %. Este alto valor de IDF1 es un fuerte indicador de que el sistema puede mantener la identidad correcta de cada objeto a lo largo del tiempo, incluso en presencia de occlusiones o interacciones complejas, lo cual es fundamental para un seguimiento fiable a nivel individual.

Finalmente, la métrica HOTA, que ofrece una evaluación más holística, alcanzó un valor del 71.7 %, confirmando un equilibrio sólido y competente entre la calidad de la detección inicial y la efectividad de la asociación de trayectorias. Colectivamente, estos resultados cuantitativos no solo demuestran la eficacia del sistema para el seguimiento de objetos múltiples, sino que también justifican las elecciones de diseño y la combinación de tecnologías empleadas, subrayando su preparación para aplicaciones que requieren un seguimiento robusto y preciso en entornos dinámicos.

CAPÍTULO 5

Prueba de concepto

Tras la fase de desarrollo y análisis de la solución propuesta, que se llevó a cabo utilizando vídeos pregrabados en un entorno de laboratorio, se procedió a la implementación de una prueba de concepto. Esta prueba simula un entorno de producción real mediante la construcción de una cinta transportadora sencilla. El principal objetivo de esta etapa es validar la viabilidad y el rendimiento de la solución en un escenario práctico, evaluando su capacidad para detectar y clasificar objetos (canicas) en movimiento sobre dicha cinta.

5.1 Diseño y construcción del sistema físico

La implementación práctica de la solución propuesta requirió el diseño y construcción de una plataforma física que intenta simular un entorno de producción industrial real. Para esta tarea, se tomaron como referencia los planos y la lista de materiales detallados en un proyecto de acceso público [20], que proporcionó una base sólida para el desarrollo de una cinta transportadora funcional y económicamente viable.

La selección de este diseño de referencia se basó en varios criterios fundamentales: su simplicidad constructiva, la disponibilidad de los materiales necesarios, la posibilidad de adaptación para diferentes tipos de objetos, y su capacidad para integrar sistemas de visión artificial. Además, el proyecto de referencia había sido previamente validado en aplicaciones similares, lo que reducía los riesgos técnicos asociados con el desarrollo desde cero.

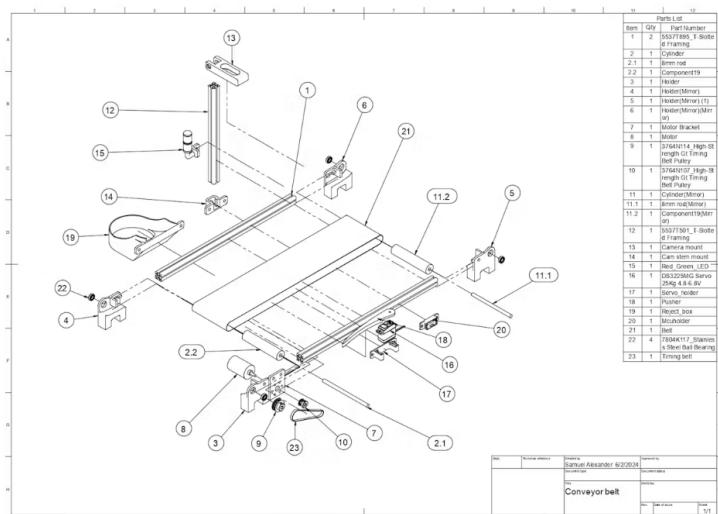


Figura 5.1: Planos de la cinta transportadora. Extraído de [20].

Los materiales empleados para la construcción de la cinta transportadora son los siguientes:

- **Motor de corriente continua (DC) de 12V y 319 RPM:** Proporciona movimiento constante a la banda transportadora.
- **Filamento PLA:** Material para imprimir en 3D las piezas estructurales de la cinta (soportes, poleas, guías).
- **Cuatro rodamientos 688ZZ:** Permiten el movimiento suave de los ejes rotativos con baja fricción.
- **Servomotor DS3225:** Servomotor para mover la palanca de desvío de canicas defectuosas.
- **Banda transportadora:** Cinta de 7.5 cm de ancho y 1.2 m de longitud con superficie texturizada.
- **Correa de sincronización GT2:** Correa dentada para transmisión de potencia entre motor y poleas.
- **Raspberry Pi Pico WH:** Microcontrolador con WiFi para control del servomotor.
- **Cámara USB:** Dispositivo de captura de imágenes a 30 fps.
- **Componentes auxiliares:** Cables, tornillería y elementos de fijación.
- **Fuente de alimentación:** Fuente de 12V/3A para alimentar el motor y componentes electrónicos.

El montaje de la cinta transportadora se realizó siguiendo una metodología sistemática basada en los planos ilustrados en la Figura 5.1. El proceso se dividió en varias fases para garantizar la precisión del ensamblaje y facilitar las pruebas intermedias de funcionamiento.

La primera fase consistió en la impresión 3D de todas las piezas estructurales utilizando una impresora Prusa MK2[35] con filamento PLA.

La segunda fase abarcó el ensamblaje de la estructura mecánica principal. Los rodamientos se instalaron en sus alojamientos correspondientes utilizando un ajuste a presión controlado, verificando la ausencia de holguras excesivas que pudieran generar vibraciones durante el funcionamiento.

Durante la tercera fase se procedió a la instalación del sistema de transmisión. El motor DC se montó en su soporte específico, asegurando su correcta alineación con la polea motriz. La correa de sincronización GT2 se instaló con la tensión adecuada para evitar tanto el deslizamiento como el desgaste prematuro por sobretensión.

La cuarta fase incluyó la instalación de la banda transportadora y el ajuste de su tensión. Este proceso requirió varios ciclos de ajuste iterativo para lograr un movimiento uniforme sin ondulaciones ni desviaciones laterales. Se verificó que la superficie de transporte mantuviera una planicidad adecuada en toda su longitud.

Finalmente, la cámara se ubicó estratégicamente en la parte superior de la cinta para obtener una vista clara y perpendicular de las canicas durante su tránsito, minimizando las distorsiones ópticas y las sombras que pudieran interferir con el proceso de detección.

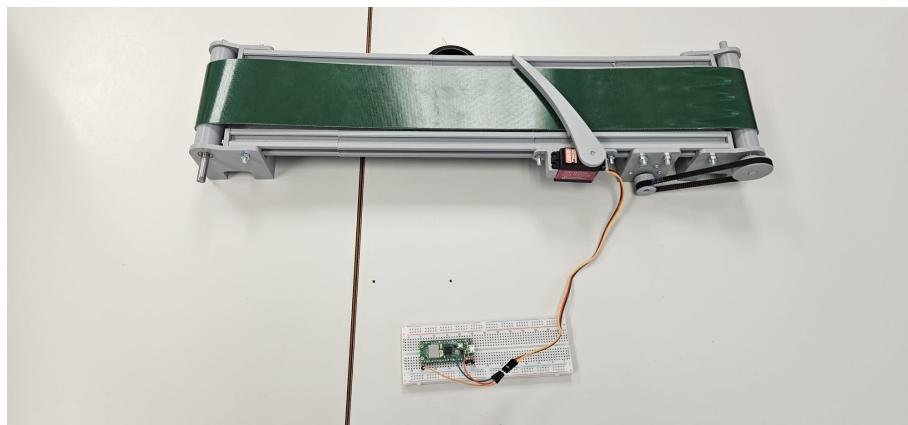


Figura 5.2: Cinta transportadora construida para la prueba de concepto.

En la Figura 5.2 se muestra la cinta transportadora construida para la prueba de concepto.

5.2 Integración del sistema de visión artificial

La integración del sistema de visión artificial constituye el núcleo tecnológico de la prueba de concepto, ya que combina los algoritmos de detección desarrollados durante las fases anteriores del proyecto con la infraestructura física de la cinta transportadora. Esta integración requirió el diseño de una arquitectura distribuida que optimiza el procesamiento en tiempo real y garantiza una respuesta adecuada del sistema de actuación.

El sistema completo se configuró utilizando una arquitectura distribuida que separa las responsabilidades de procesamiento y control entre dos dispositivos especializados. El procesamiento principal, incluyendo la ejecución del modelo de detección de objetos y los algoritmos de seguimiento, se realiza en un dispositivo NVIDIA Jetson AGX Xavier.

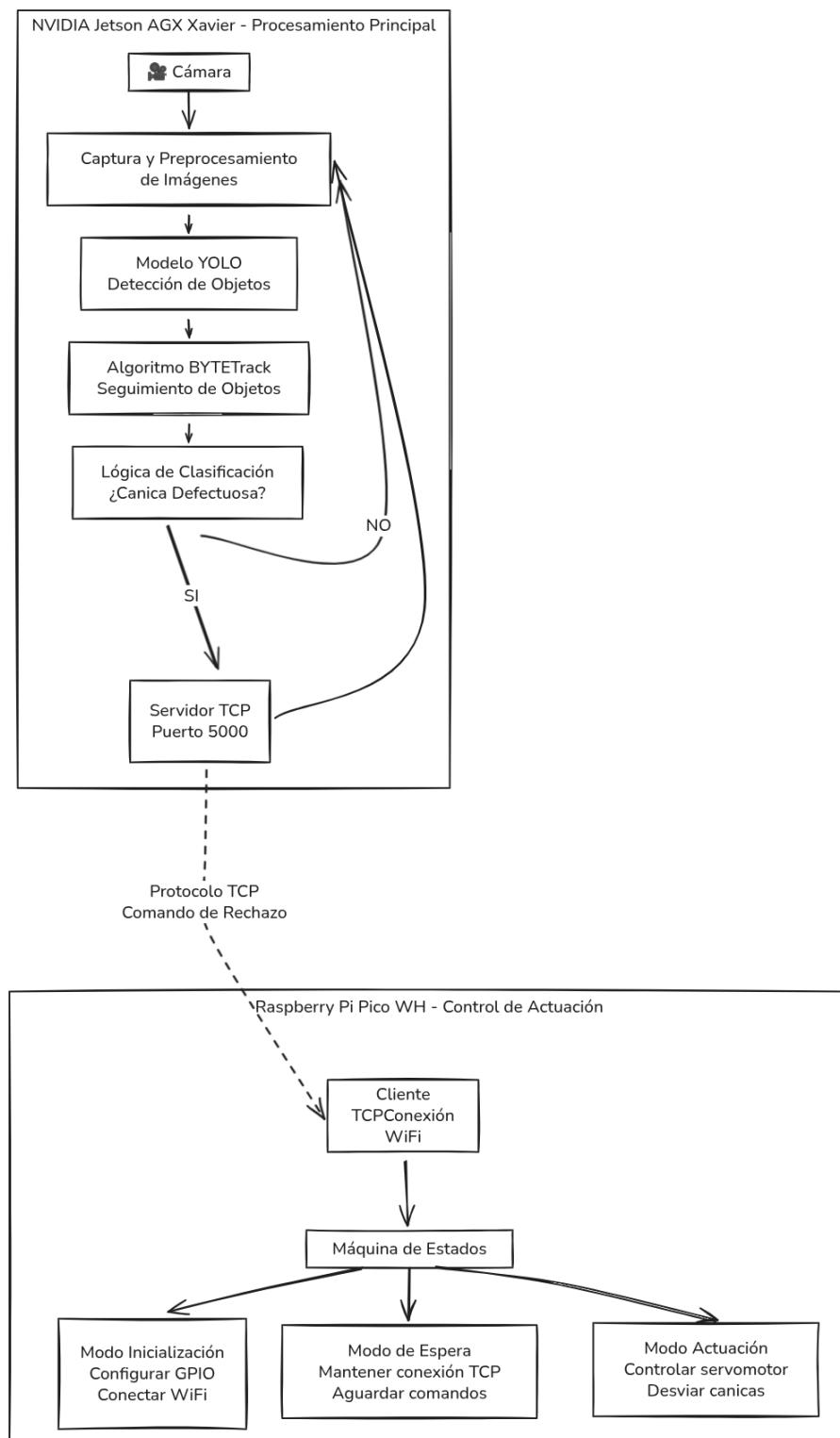


Figura 5.3: Diagrama de la arquitectura del sistema de visión artificial.

El Jetson AGX Xavier actúa como el cerebro del sistema, ejecutando continuamente los siguientes procesos en paralelo: captura y preprocesamiento de imágenes de la cámara, inferencia del modelo de detección YOLO entrenado, algoritmos de seguimiento BYTETrack para mantener la identidad de los objetos, lógica de clasificación para determinar si una canica es defectuosa, y comunicación en tiempo real con el sistema de actuación.

Como complemento, la Raspberry Pi Pico WH se programó específicamente para controlar el servomotor que desvía las canicas identificadas como defectuosas, tal como se ilustra en la Figura 5.3. Esta separación de responsabilidades permite que el sistema de actuación responda con latencia mínima, mientras que el Jetson puede dedicar todos sus recursos al procesamiento de visión artificial sin interrupciones por tareas de control de hardware.

Para la programación de la Raspberry Pi Pico WH se utilizó MicroPython[26], un lenguaje que ofrece la simplicidad de Python con las capacidades de tiempo real necesarias para aplicaciones de control. La elección de MicroPython se basó en su facilidad de desarrollo, su amplia documentación, y su capacidad para manejar comunicaciones de red y control de hardware de forma simultánea.

El firmware desarrollado para el microcontrolador implementa una máquina de estados que gestiona tres modos principales de operación: modo de inicialización, donde se configuran los pines GPIO y se establece la conexión WiFi; modo de espera, donde el sistema aguarda comandos del Jetson manteniendo una conexión TCP activa; y modo de actuación, donde se ejecuta la secuencia de movimiento del servomotor para desviar canicas defectuosas.

Para la comunicación entre el Jetson AGX Xavier y la Raspberry Pi Pico WH, se utilizó el protocolo TCP que garantiza la confiabilidad y el orden de los mensajes intercambiados. La elección de TCP sobre UDP se justifica por la necesidad de garantizar la entrega de todos los comandos de actuación, ya que la pérdida de un mensaje podría resultar en que una canica defectuosa no sea rechazada correctamente. La arquitectura de comunicación se muestra en la Figura 5.3, donde se aprecia la interconexión entre los componentes del sistema.

Al iniciar el sistema, el Jetson AGX Xavier actúa como servidor TCP, abriendo un socket en el puerto 5000 y configurando un buffer de recepción optimizado para mensajes de tamaño reducido. El servidor implementa un mecanismo de reconexión automática que permite recuperarse de desconexiones temporales sin perder el estado del sistema. Se configuró un timeout de conexión de 30 segundos y un sistema de heartbeat cada 5 segundos para detectar desconexiones de manera proactiva.

Durante el funcionamiento del sistema, cuando el modelo de detección ejecutándose en el Jetson identifica una canica defectuosa basándose en los criterios de clasificación establecidos, se activa una secuencia temporizada cuidadosamente diseñada. El sistema de seguimiento BYTETrack proporciona información de posición y velocidad de la canica, permitiendo calcular el tiempo exacto en que llegará a la zona de actuación del servomotor.

Al recibir la señal de activación, la Raspberry Pi Pico WH ejecuta la rutina de rechazo siguiendo una secuencia optimizada. Primero se posiciona el servomotor en la posición de espera, luego se ejecuta un movimiento rápido hasta la posición de desviación, se mantiene la posición durante el tiempo necesario para que la canica pase completamente, y finalmente se retorna a la posición de espera con un movimiento suave.

Este accionamiento preciso mueve un mecanismo mecánico compuesto por una paleta de desviación que intercepta la trayectoria de la canica defectuosa, desviándola fuera de la trayectoria principal de la cinta transportadora, tal como se observa en el diagrama de la Figura 5.3. El diseño del mecanismo garantiza que las canicas válidas no sean afectadas por el movimiento del actuador, separando efectivamente los productos defectuosos del flujo de productos válidos.

5.3 Resultados experimentales y evaluación

La evaluación experimental de la prueba de concepto se diseñó para validar tanto la funcionalidad técnica del sistema como su viabilidad práctica en un entorno que simula condiciones industriales reales. Esta evaluación abarcó múltiples aspectos del sistema, desde la precisión de detección hasta la efectividad del mecanismo de rechazo, proporcionando una base sólida para evaluar el potencial de escalabilidad de la solución propuesta.

En la Figura 5.4 se muestra un frame representativo de la prueba de concepto, donde se observan tres canicas de color blanco transitando por la cinta transportadora. Entre ellas, se identifica claramente una canica defectuosa que presenta una mancha roja, correspondiente al criterio de clasificación establecido para distinguir los productos defectuosos de los válidos.



Figura 5.4: Imagen de ejemplo de la prueba de concepto.

La prueba de concepto validó la capacidad del sistema para detectar y clasificar canicas en movimiento sobre una cinta transportadora en tiempo real. El sistema demostró efectividad en la identificación de objetos defectuosos, activando exitosamente el mecanismo de rechazo cuando se detectaban canicas con manchas.

No obstante, el mecanismo de rechazo implementado presenta limitaciones significativas debido a su simplicidad. La precisión del sistema para desviar únicamente las canicas defectuosas no fue óptima, ya que el actuador mecánico requiere un tiempo de respuesta que no siempre se sincroniza perfectamente con la velocidad de la cinta transportadora.

Además, la infraestructura de la cinta transportadora construida carece de un sistema de recirculación continua, lo que provoca que las canicas se acumulen al final del recorrido o caigan fuera de la estructura. Esta limitación impide evaluar el comportamiento del sistema durante períodos prolongados de operación continua.

Aunque la implementación actual es demasiado rudimentaria para considerarse una solución industrial completa, los resultados obtenidos confirman la viabilidad técnica del enfoque propuesto. La prueba establece una base sólida que demuestra el potencial del sistema para ser escalado y refinado hacia aplicaciones industriales más complejas y robustas.

CAPÍTULO 6

Conclusiones

En este capítulo se presentan las conclusiones obtenidas a partir del desarrollo y la implementación de la solución propuesta para la detección y clasificación de objetos en movimiento sobre una cinta transportadora. Se reflexiona sobre los logros alcanzados, las lecciones aprendidas durante el proceso y las posibles direcciones futuras para mejorar y ampliar el sistema.

6.1 Objetivos alcanzados y dificultades

A lo largo de este Trabajo de Fin de Grado, se han abordado diversos desafíos y se han alcanzado los objetivos propuestos, aunque no sin encontrar ciertas dificultades inherentes al desarrollo de un sistema complejo de visión artificial. A continuación, se detallan los principales logros y los obstáculos superados.

Los objetivos fundamentales del proyecto se han alcanzado con éxito. Se llevó a cabo un estudio exhaustivo del estado del arte, abarcando tanto las CNNs, con experimentación en diversos modelos de detección, como los aceleradores hardware, con especial atención a las GPUs y la plataforma NVIDIA Jetson. Durante esta fase, se identificaron y superaron desafíos significativos, principalmente relacionados con la gestión de la compatibilidad entre las múltiples versiones de *frameworks* de aprendizaje profundo y el software de NVIDIA. La arquitectura ARM64 de los dispositivos Jetson también presentó particularidades que complejizaron la instalación y configuración de dependencias críticas, como ciertas bibliotecas de Python y componentes específicos del ecosistema de NVIDIA. A pesar de estos obstáculos técnicos, se logró una configuración estable y funcional del entorno de desarrollo y experimentación.

Se generó un conjunto de datos y vídeos de entrenamiento, fundamental para optimizar la precisión del modelo de detección. Estos recursos, proporcionaron un entorno de prueba que intenta simular condiciones reales de operación. El principal desafío fue seleccionar un objeto de prueba fácilmente identificable y con variabilidad controlada en forma y color. Aunque se hubiera preferido un objeto industrial más complejo, no se dispuso de medios para su adquisición.

Se entrenaron múltiples modelos de detección de las familias YOLOv5, YOLOv8 y YOLO11, en diversas precisiones y tamaños. Esto permitió evaluar su evolución y rendimiento en precisión y velocidad. La principal dificultad fue el ajuste fino de hiperparámetros para optimizar el rendimiento, proceso que requirió numerosas iteraciones y pruebas.

Se implementó un sistema de seguimiento que integra los modelos de detección entrenados con el algoritmo BYTETrack. Esta combinación permitió mantener la identidad

de los objetos a lo largo del tiempo, mejorando significativamente la precisión y robustez del sistema. El principal obstáculo fue la integración del algoritmo de seguimiento con la salida de detección, debido a incompatibilidades entre versiones de *frameworks* y dependencias. No obstante, se logró una integración exitosa y eficaz en la práctica.

Para analizar los cuellos de botella, se exploraron diversas estrategias de segmentación para identificar componentes críticos y optimizar su rendimiento. La segmentación por procesos con memoria compartida fue la más efectiva, facilitando la comunicación eficiente entre detección y seguimiento, y mejorando notablemente el rendimiento global. La mayor complejidad residió en la programación de la comunicación interproceso (transmisión de datos y sincronización), que exigió un conocimiento profundo de programación concurrente y gestión de memoria compartida.

Para cuantificar las métricas de rendimiento bajo diferentes configuraciones de hardware, se realizaron múltiples experimentos. Estos incluyeron la evaluación de distintos modelos y tamaños, operando a diversas tasas de fotogramas por segundo (FPS) y utilizando diferentes dispositivos Jetson. Estos experimentos permitieron evaluar el rendimiento en condiciones realistas y ajustar la configuración para un equilibrio óptimo entre precisión y velocidad. La principal dificultad fue la ejecución de numerosas pruebas y ajustes para identificar la configuración óptima, lo que demandó una considerable inversión de tiempo y recursos.

6.2 Aprendizaje

Durante el desarrollo de este Trabajo de Fin de Grado, se adquirieron conocimientos y habilidades valiosas en diversas áreas. En primer lugar, se profundizó en el ciclo de vida completo de las redes neuronales, desde la creación y anotación de conjuntos de datos, pasando por el entrenamiento y ajuste de hiperparámetros, hasta la optimización de los modelos para una inferencia eficiente. Este proceso implicó un aprendizaje significativo sobre la arquitectura interna de las redes, las métricas de evaluación y las técnicas para mejorar su rendimiento.

Paralelamente, se exploró en detalle el uso de aceleradores hardware, con un enfoque particular en las GPUs y la plataforma NVIDIA Jetson. Esta exploración permitió comprender cómo estas tecnologías pueden mejorar drásticamente el rendimiento de los sistemas de visión artificial en tiempo real. La experiencia práctica obtenida en la configuración y optimización del entorno de desarrollo, así como en la resolución de problemas complejos relacionados con la compatibilidad de software y las particularidades de la arquitectura ARM64, constituyó un aspecto crucial del aprendizaje.

Asimismo, se desarrollaron habilidades prácticas en la implementación de sistemas de seguimiento de objetos. La integración de los modelos de detección entrenados con algoritmos avanzados como BYTETrack supuso un aprendizaje profundo sobre la gestión de identidades y la reconstrucción de trayectorias de objetos en movimiento, aspectos esenciales para aplicaciones industriales y comerciales.

Además, se adquirieron conocimientos sólidos sobre técnicas de programación concurrente, específicamente la segmentación por procesos y la comunicación interproceso mediante memoria compartida. Esta área permitió optimizar significativamente el rendimiento del sistema al identificar y abordar cuellos de botella críticos en el flujo de datos.

Finalmente, la planificación y ejecución de una batería exhaustiva de experimentos para cuantificar métricas de rendimiento bajo diversas configuraciones de hardware y software proporcionó una comprensión práctica y metodológica sobre cómo evaluar, comparar y mejorar el rendimiento en sistemas complejos de visión artificial.

6.3 Relación con los estudios cursados

Este Trabajo de Fin de Grado se relaciona estrechamente con los estudios cursados en el Grado en Ingeniería Informática, especialmente en las siguientes asignaturas han sido las que han proporcionado la base teórica y práctica necesaria para llevar a cabo este proyecto:

- **SDL (Sistemas basados en Deep Learning para la Industria):** La asignatura se encuentra en el segundo cuatrimestre del tercer curso del grado en Ingeniería Informática en la mención de Ingeniería de Computadores. La asignatura tiene como objetivos «Conocer, configurar y utilizar los sistemas actuales específicos para inteligencia artificial (en concreto aprendizaje profundo) que existen para la industria. Especial énfasis en sistemas reales con capacidad de procesamiento de imágenes (clasificación/detección de objetos)» y «Conocer y profundizar en sistemas basados en GPU para el procesamiento de modelos de redes neuronales». Esta asignatura es la base teórica y práctica de este Trabajo de Fin de Grado, ya que se ha utilizado el conocimiento adquirido en ella para implementar la solución propuesta, incluyendo la selección y entrenamiento de modelos de detección de objetos, así como la optimización para su ejecución en dispositivos Jetson.
- **CPA (Computación Paralela):** La asignatura se encuentra en el primer cuatrimestre del tercer curso del grado en Ingeniería Informática. La asignatura tiene como objetivo «Conocer la computación paralela y los modelos de programación paralela a través de los modelos más extendidos: memoria compartida y memoria distribuida». Los conocimientos adquiridos en esta asignatura han sido de especial relevancia para la implementación de técnicas de paralelización para distribuir la carga computacional del procesado de vídeos entre hilos y/o procesos para reducir el tiempo de procesado.

6.4 Trabajo futuro

Como trabajo futuro para mejorar la solución propuesta, se plantean varias líneas de investigación y desarrollo que podrían ampliar las capacidades del sistema y optimizar su rendimiento.

Una primera línea de mejora consistiría en **evaluar el sistema con objetos de mayor complejidad**. Aunque la prueba de concepto se realizó con canicas, un objeto relativamente simple, sería beneficioso probar el sistema con una gama más amplia de objetos que presenten mayor variabilidad en forma, tamaño, textura y color. Esto permitiría validar la robustez y la capacidad de generalización del sistema en escenarios más desafiantes, acercándose a las condiciones reales de un entorno industrial. La utilización de objetos industriales auténticos, con sus irregularidades y posibles defectos específicos, proporcionaría una evaluación más fidedigna del rendimiento del sistema en aplicaciones prácticas.

Otra dirección prometedora es la **incorporación de múltiples cámaras** para obtener una visión más completa de la escena. Un sistema multicámara permitiría capturar imágenes desde diferentes ángulos y perspectivas, lo que podría mejorar significativamente la detección y el seguimiento de objetos, especialmente en situaciones de oclusión parcial o cuando los objetos se mueven rápidamente y cambian de orientación. La fusión de información de múltiples vistas podría conducir a una reconstrucción 3D más precisa de los objetos y sus trayectorias, enriqueciendo el análisis y permitiendo una inspección más detallada.

Además de la visión, se podría **integrar información de otros tipos de sensores** para complementar los datos visuales. Sensores de proximidad, sensores de peso, cámaras térmicas o incluso sensores hiperespectrales podrían proporcionar información adicional valiosa. Por ejemplo, los sensores de peso podrían ayudar a verificar la cantidad de producto, mientras que las cámaras térmicas podrían detectar anomalías de temperatura no visibles. Esta fusión multisensorial podría mejorar la capacidad del sistema para identificar defectos sutiles o características que no son fácilmente discernibles solo con cámaras RGB.

Para completar el ciclo de inspección, se podría **desarrollar un mecanismo de rechazo de objetos defectuosos más sofisticado y automatizado**. Esto podría implicar el diseño e integración de actuadores neumáticos, brazos robóticos pequeños o desviadores mecánicos controlados por el sistema. Al detectar un objeto no conforme, el sistema enviaría una señal para activar el mecanismo de rechazo, retirando el producto defectuoso de la línea de producción de manera eficiente y sin intervención manual, lo que mejoraría la productividad y reduciría errores.

Finalmente, para escalar la capacidad de procesamiento y manejar flujos de vídeo de mayor resolución o múltiples líneas de producción, se podría explorar la **implementación de un sistema distribuido utilizando varios dispositivos Jetson**. Cada dispositivo podría encargarse de una porción del flujo de vídeo o de una tarea específica (p. ej., un dispositivo para detección y otro para seguimiento avanzado). Esta arquitectura distribuida permitiría un procesamiento paralelo más efectivo, mejorando la velocidad general del sistema y su capacidad para manejar cargas de trabajo más intensivas, haciéndolo adecuado para entornos industriales de mayor escala.

Bibliografía

- [1] Anders S. G. Andrae and Tomas Edler. On global electricity usage of communication technology: Trends to 2030. *Challenges*, 6(1):117–157, 2015.
- [2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, September 2016.
- [3] Google Cloud. Tensor processing units (tpus). <https://cloud.google.com/tpu>, 2025. Accedido el 21 de mayo de 2025.
- [4] Computer Science Wiki. File:maxpoolsample2.png. <https://computersciencewiki.org/index.php/File:MaxpoolSample2.png>, 2018. Accedido el 20 de mayo de 2025.
- [5] T. Conte. IEEE rebooting computing initiative & international roadmap of devices and systems. In *Proc. IEEE Rebooting Computer Architecture 2030 Workshop*.
- [6] CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT), November 2023.
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [8] Robert H. Dennard, Fritz H. Gaenslen, Hwa-Nien Yu, V. Leo Rideout, Ernest Bassous, and Andre R. Leblanc. Design of ion-implanted mosfets with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, SC-9(5):256–268, October 1974.
- [9] ENCCS. The gpu hardware and software ecosystem. <https://enccs.github.io/gpu-programming/2-gpu-ecosystem/>, 2025. Parte del curso "GPU programming: why, when and how?".
- [10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014.
- [11] Wen-mei W. Hwu, David B. Kirk, and Izzat El Hajj. *Programming Massively Parallel Processors: A Hands-on Approach*. Elsevier, 4th edition, 2022.
- [12] Glenn Jocher. Ultralytics yolov5, 2020.
- [13] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.
- [14] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024.
- [15] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023.
- [16] Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

- [17] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336, 2009.
- [18] Swapna Kategaru. Convolution neural network in deep learning. <https://developersbreach.com/convolution-neural-network-deep-learning/>, 2025. Accedido el 26 de abril de 2025.
- [19] Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun. *A Guide to Convolutional Neural Networks for Computer Vision*. Synthesis Lectures on Computer Vision. Springer Cham, 1 edition, 2018.
- [20] Kinetica. Counting for inspection and quality control with tensorrt. <https://www.hackster.io/kinetika/counting-for-inspection-and-quality-control-with-tensorrt-550b91>, 2024. Accedido el 26 de mayo de 2025.
- [21] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.
- [22] Tushar Kumar. R-cnn explained. <https://youtu.be/5DvljLV4S1E?si=oeQUo9Cmv4NInTns>, 2024. Video. Accedido: 14 de abril de 2025.
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.
- [24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. *SSD: Single Shot MultiBox Detector*, page 21–37. Springer International Publishing, 2016.
- [25] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129:1–31, 02 2021.
- [26] MicroPython. Micropython - python for microcontrollers. <https://micropython.org/>, 2025. Accedido el 2 de junio de 2025.
- [27] Karthik Mittal. Object detection with histogram of oriented gradients (hog). <https://iq.opengenus.org/object-detection-with-histogram-of-oriented-gradients-hog/>, 2020. Accedido el 20 de mayo de 2025.
- [28] Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114–117, April 1965.
- [29] NVIDIA Corporation. Nvidia deep learning accelerator (nvdla). <https://developer.nvidia.com/deep-learning-accelerator>, 2024. Accedido el 21 de mayo de 2025.
- [30] NVIDIA Corporation. Cuda toolkit. <https://developer.nvidia.com/cuda-toolkit>, 2025. Accedido el 19 de junio de 2025.
- [31] NVIDIA Corporation. Jetson modules, support, ecosystem, and lineup. <https://developer.nvidia.com/embedded/jetson-modules>, 2025. Accedido el 24 de abril de 2025.

- [32] NVIDIA Corporation. Nvidia. <https://www.nvidia.com/>, 2025. Accedido el 19 de junio de 2025.
- [33] NVIDIA Corporation. Tegrastats utility. https://docs.nvidia.com/drive/drive_os_5.1.6.1L/nvvib_docs/index.html#page/DRIVE_OS_Linux_SDK_Development_Guide/Utilities/util_tegrastats.html, 2025. Accedido el 24 de junio de 2025.
- [34] ONNX. Open neural network exchange (onnx), 2025.
- [35] Prusa Research. Mk2/mk2s. <https://www.prusa3d.com/es/categoría/mk2-mk2s/>, 2025. Accedido el 17 de junio de 2025.
- [36] Python documentation. multiprocessing — process-based parallelism. <https://docs.python.org/3.8/library/multiprocessing.html>. Accedido el 24 de junio de 2025.
- [37] Python documentation. multiprocessing.shared_memory — shared memory for direct access across processes. https://docs.python.org/3.8/library/multiprocessing.shared_memory.html. Accedido el 24 de junio de 2025.
- [38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.
- [39] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking, 2016.
- [40] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. Technical report, University of Trento and University of Amsterdam, 2012. Technical Report, submitted to IJCV.
- [41] Ultralytics. Docker quickstart. <https://docs.ultralytics.com/es/guides/docker-quickstart/#verify-nvidia-runtime-with-docker>, 2024. Accedido el 9 de junio de 2025.
- [42] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017.
- [43] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2022.

APÉNDICE A

Objetivos de desarrollo sostenible

Objetivos de Desarrollo Sostenible	Alto	Medio	Bajo	No procede
ODS 1. Fin de la pobreza.				X
ODS 2. Hambre cero.		X		
ODS 3. Salud y bienestar.				X
ODS 4. Educación de calidad.				X
ODS 5. Igualdad de género.				X
ODS 6. Agua limpia y saneamiento.				X
ODS 7. Energía asequible y no contaminante.				X
ODS 8. Trabajo decente y crecimiento económico.	X			
ODS 9. Industria, innovación e infraestructuras.	X			
ODS 10. Reducción de las desigualdades.				X
ODS 11. Ciudades y comunidades sostenibles.				X
ODS 12. Producción y consumo responsables.	X			
ODS 13. Acción por el clima.			X	
ODS 14. Vida submarina.				X
ODS 15. Vida de ecosistemas terrestres.				X
ODS 16. Paz, justicia e instituciones sólidas.				X
ODS 17. Alianzas para lograr objetivos.				X

Justificación de los Objetivos de Desarrollo Sostenible

ODS-8. Trabajo decente y crecimiento económico: Este proyecto contribuye directamente al crecimiento económico sostenible mediante la automatización inteligente de procesos de control de calidad. La implementación de sistemas de detección de defectos basados en IA permite reducir costes operativos, minimizar desperdicios y optimizar la cadena de producción, lo que se traduce en mayor productividad y competitividad empresarial. Se alinea específicamente con la meta 8.2 de la ONU: «Lograr niveles más elevados de productividad económica mediante la diversificación, la modernización tecnológica y la innovación», al incorporar tecnologías avanzadas de procesamiento de imágenes y ML en entornos industriales tradicionales.

ODS-9. Industria, innovación e infraestructura: El desarrollo de sistemas inteligentes para la detección de defectos representa una clara apuesta por la innovación industrial. Este proyecto no solo implementa tecnologías emergentes como la IA en procesos productivos, sino que además optimiza su rendimiento mediante el uso eficiente de hardware especializado como GPUs, algoritmos de seguimiento multi-objeto y técnicas de pa-

ralelización. Esto responde directamente a la meta 9.4 de la ONU: «Modernizar la infraestructura y reconvertir las industrias para que sean sostenibles, utilizando los recursos con mayor eficacia y promoviendo la adopción de tecnologías y procesos industriales limpios y ambientalmente racionales», al permitir mejoras significativas en eficiencia energética y uso de recursos mediante sistemas de inspección automatizados.

ODS-12. Producción y consumo responsables: La implementación de sistemas de detección temprana de defectos contribuye sustancialmente a la producción responsable mediante: 1) la reducción del descarte de productos y materias primas al identificar problemas en etapas iniciales del proceso productivo, 2) la optimización del consumo energético al evitar el procesamiento completo de productos defectuosos, y 3) la mejora de la calidad final que aumenta la vida útil de los productos. Estas aportaciones se vinculan directamente con la meta 12.5 de la ONU: «De aquí a 2030, reducir considerablemente la generación de desechos mediante actividades de prevención, reducción, reciclado y reutilización», ya que el sistema desarrollado actúa preventivamente evitando la generación de residuos industriales y facilitando la reutilización de materiales recuperados.

APÉNDICE B

Código fuente

En este capítulo se presenta el código fuente más relevante de la solución propuesta eliminando las partes que no aportan valor al lector. El código completo se encuentra disponible en el repositorio de GitHub del proyecto¹

El primero de los archivos que se muestra a continuación es el objeto DetectionTrackingPipeline, que implementa las funciones de las cuatro fases del pipeline de detección y seguimiento de objetos. Este objeto se encarga de inicializar el modelo de detección, el algoritmo de seguimiento, la cámara y el socket para la comunicación con la Raspberry Pi Pico WH. Además, implementa las funciones para procesar los fotogramas de vídeo, detectar los objetos, realizar el seguimiento y pintarlos en el fotograma.

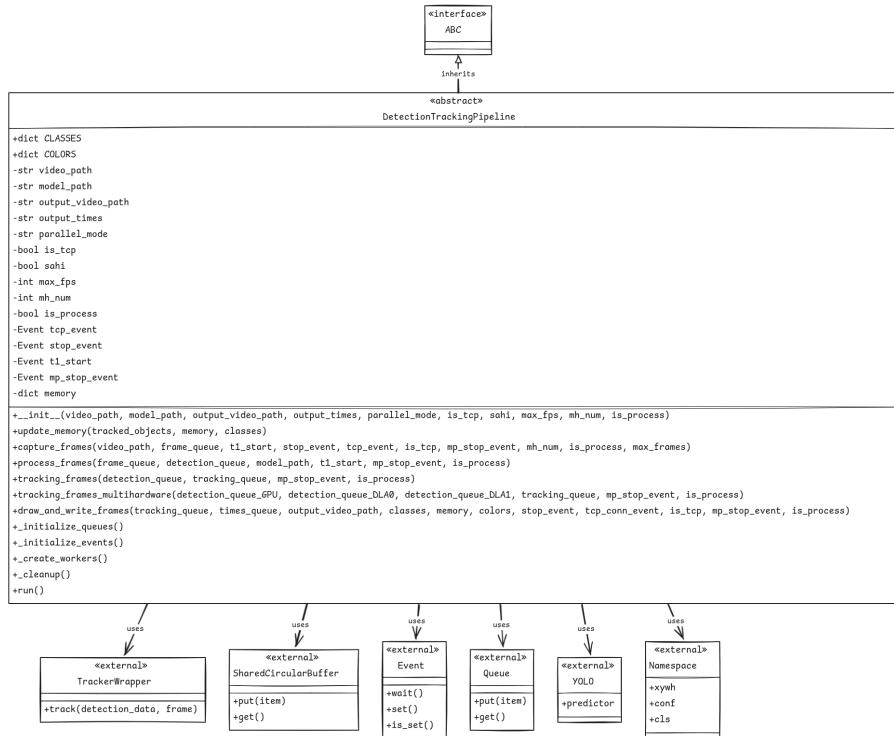


Figura B.1: Diagrama de clases del objeto DetectionTrackingPipeline.

La figura B.1 muestra el diagrama de clases del objeto DetectionTrackingPipeline. Durante la ejecución del programa, las funciones *capture_frames*, *process_frames*, *tracking_frames* y *draw_and_write_frames* se ejecutarán en paralelo, cada una en un hilo o proceso diferente.

¹<https://github.com/AbelHaro/TFG>

rente, según la configuración elegida. El objeto `DetectionTrackingPipeline` se encarga de gestionar la comunicación entre estos hilos o procesos, asegurando que los datos se transmitan correctamente entre ellos y que el sistema funcione de manera eficiente.

Listing B.1: detection_tracking_pipeline.py

```
1  from abc import ABC, abstractmethod
2  import cv2
3  import os
4  from argparse import Namespace
5  from classes.tracker_wrapper import TrackerWrapper
6  from lib.tcp import handle_send, tcp_server
7  import logging
8  from typing import Union, Optional
9  import torch.multiprocessing as mp
10 from classes.shared_circular_buffer import SharedCircularBuffer
11 import time
12
13
14 class DetectionTrackingPipeline(ABC):
15     """Abstract base class for object detection and tracking pipelines.
16
17     This class defines the structure and common functionality for various
18     pipelines
19     that integrate object detection and their subsequent tracking through
20     video
21     sequences. It allows the implementation of different parallelization
22     strategies
23     and hardware management.
24     """
25
26     CLASSES = {
27         0: "negra",
28         1: "blanca",
29         2: "verde",
30         3: "azul",
31         4: "negra-d",
32         5: "blanca-d",
33         6: "verde-d",
34         7: "azul-d",
35     }
36
37     COLORS = {
38         "negra": (0, 0, 255),
39         "blanca": (0, 255, 0),
40         "verde": (255, 0, 0),
41         "azul": (255, 255, 0),
42         "negra-d": (0, 165, 255),
43         "blanca-d": (255, 165, 0),
44         "verde-d": (255, 105, 180),
45         "azul-d": (255, 0, 255),
46     }
47
48     def update_memory(self, tracked_objects, memory, classes) -> None:
49         """Updates tracking memory with detected and tracked objects.
50
51         Maintains a record of objects, their state (defective or not),
52         and their visibility throughout frames. An object is considered
53         permanently defective if detected as defective during a
54         consecutive number of frames defined by 'PERMANENT_DEFECT_THRESHOLD'
55         .
56
57         Objects that have not been seen for 'FRAME_AGE' frames are removed
58         from memory.
59
60         Args:
61             tracked_objects (List[dict]): A list of dictionaries containing
62                 information about tracked objects. Each dictionary has keys
63                 'id', 'bbox' (tuple), 'state' (bool), and 'last_seen' (int).
64             memory (SharedCircularBuffer): The shared circular buffer used
65                 to store tracked objects.
66             classes (Namespace): A namespace object containing class definitions.
67
68         Returns:
69             None
70
71         Raises:
72             None
73
74         Notes:
75             This method updates the tracking memory with the latest
76             information from the tracked objects. It also maintains a
77             record of objects that have been seen for a long time
78             ('PERMANENT_DEFECT_THRESHOLD') and removes them from memory.
```

```

56     tracked_objects: List of tracked objects in the current frame.
57         Each object is a tuple or list with
58             information like
59                 tracking ID, detected class, etc.
60             memory: Dictionary that stores the state of tracked objects
61                 between frames. Keys are tracking IDs.
62             classes: Dictionary that maps class IDs to their names.
63             """
64
65             FRAME_AGE = 60 # Number of frames to keep an object in memory if
66             not visible
67             PERMANENT_DEFECT_THRESHOLD = 3 # Consecutive frames to mark as "
68             permanent defect"
69
70
71             for obj in tracked_objects:
72                 track_id = int(obj[4])
73                 detected_class = classes[int(obj[6])]
74                 is_defective = detected_class.endswith("-d")
75
76                 if track_id in memory:
77                     entry = memory[track_id]
78
79                     # If already marked as permanent defect, only update its
80                     visibility
81                     if entry.get("permanent_defect", False):
82                         entry["visible_frames"] = FRAME_AGE
83                         continue
84
85                     # Updates the consecutive defect counter
86                     if is_defective:
87                         entry["defect_counter"] = entry.get("defect_counter",
88                             0) + 1
89                     else:
90                         entry["defect_counter"] = 0 # Reset if not defective
91                         in this frame
92
93                     # Mark as permanent defect if it reaches the threshold
94                     if entry["defect_counter"] >= PERMANENT_DEFECT_THRESHOLD:
95                         entry["permanent_defect"] = True
96                         # The class already includes '-d', no need to reassign
97                         'detected_class' here
98                         # entry["defective"] will be updated below
99
100
101                     # Update defective status and visibility
102                     entry["defective"] = entry.get("permanent_defect", False)
103                     or is_defective
104                     entry["visible_frames"] = FRAME_AGE
105                     entry["class"] = detected_class
106
107                     else:
108                         # New detected object
109                         memory[track_id] = {
110                             "defective": is_defective,
111                             "visible_frames": FRAME_AGE,
112                             "class": detected_class,
113                             "defect_counter": 1 if is_defective else 0,
114                             "permanent_defect": False, # Initialize as not
115                                         permanent
116                         }
117
118
119                     # Decrement visibility and remove old objects
120                     for track_id in list(memory): # Iterate over a copy of the keys
121                         memory[track_id]["visible_frames"] -= 1
122                         if memory[track_id]["visible_frames"] <= 0:
123                             del memory[track_id]

```

```

111     def capture_frames(
112         self,
113         video_path: str,
114         frame_queue: Union[mp.Queue, SharedCircularBuffer],
115         t1_start: mp.Event,
116         stop_event: mp.Event,
117         tcp_event: mp.Event,
118         is_tcp: bool,
119         mp_stop_event: Optional[mp.Event] = None,
120         mh_num: int = 1,
121         is_process: bool = False,
122         max_frames: Optional[int] = None,
123     ):
124         """Captures frames from a video file and queues them for processing
125
126         Reads frames from a video specified by 'video_path'. If 'is_tcp' is
127         True,
128         waits for 'tcp_event' to be set before starting capture.
129         Frames are put into 'frame_queue'. If 'max_frames' is defined,
130         attempts to maintain that FPS rate by limiting capture speed.
131         When finished or if 'stop_event' is set, sends 'None' to the queue
132         (as many times as 'mh_num') to signal the end of capture.
133
134         Args:
135             video_path: Path to the video file.
136             frame_queue: Queue (multiprocessing or shared circular buffer)
137                 to send frames.
138             t1_start: Multiprocessing event to synchronize start.
139             stop_event: Event to stop capture.
140             tcp_event: Event for synchronization in TCP mode.
141             is_tcp: Boolean indicating if operating in TCP mode.
142             mp_stop_event: Optional event to wait before terminating
143                 process/thread.
144             mh_num: Number of queue consumers (to send multiple 'None' at
145                 the end).
146             is_process: Boolean, True if running as a separate process.
147             max_frames: Maximum desired FPS for capture.
148
149         Raises:
150             FileNotFoundError: If 'video_path' does not exist.
151             IOError: If the video cannot be opened.
152
153
154     if not os.path.exists(video_path):
155         logging.error(f"Video file does not exist: {video_path}")
156         for _ in range(mh_num): # Notify all consumers
157             frame_queue.put(None)
158         raise FileNotFoundError(f"Video file does not exist: {video_path}")
159
160     cap = cv2.VideoCapture(video_path)
161
162     if not cap.isOpened():
163         logging.error(f"Error opening video file: {video_path}")
164         for _ in range(mh_num): # Notify all consumers
165             frame_queue.put(None)
166         raise IOError(f"Error opening video file: {video_path}")
167
168     # Wait for TCP signal if enabled
169     if is_tcp:
170         tcp_event.wait()
171
172     frame_count = 0

```

```

169     first_time = True # To record the time of the first processed
170         frame
171
172     # Wait for t1_start signal
173     t1_start.wait()
174
175     # Calculate time per frame if max_fps is specified
176     frame_time_target = 1 / max_frames if max_frames else None
177
178     logging.info("Starting frame capture...")
179     while cap.isOpened() and not stop_event.is_set():
180         loop_start_time = time.time()
181
182         if first_time: # This t1 seems to be for a benchmark, not for
183             FPS logic
184             t1 = cv2.getTickCount()
185             first_time = False
186
187         ret, frame = cap.read()
188
189         if not ret:
190             logging.info("End of video or read error.")
191             break
192
193         try:
194             # Try to put the frame in the queue, without waiting if
195             # there's an FPS limit
196             # (to discard frames if the queue is full and maintain the
197             # pace)
198             if max_frames:
199                 frame_queue.put_nowait((frame, frame_count))
200             else:
201                 frame_queue.put((frame, frame_count))
202
203             except Exception as e: # It would be better to catch a more
204                 specific exception if known
205                 logging.warning(f"Could not queue frame {frame_count}: {e}")
206
207             # Decide whether to continue or not, here the frame is
208             # simply skipped
209             pass
210
211
212             # If there's a frame_time_target, sleep to not exceed
213             # max_frames
214             if frame_time_target:
215                 elapsed_time = time.time() - loop_start_time
216                 if elapsed_time < frame_time_target:
217                     time.sleep(frame_time_target - elapsed_time)
218
219             frame_count += 1
220
221             cap.release()
222             logging.info(f"Capture finished. Total frames read: {frame_count}")
223
224             # Signal the end to consumers
225             for _ in range(mh_num):
226                 frame_queue.put(None)
227
228             # Wait for the main process/thread stop signal if necessary
229             if mp_stop_event:
230                 mp_stop_event.wait()
231
232             # Terminate the process if running as such
233             if is_process:

```

```

225     logging.info("Terminating capture process.")
226     os._exit(0)
227
228     def process_frames(
229         self,
230         frame_queue: Union[mp.Queue, SharedCircularBuffer],
231         detection_queue: Union[mp.Queue, SharedCircularBuffer],
232         model_path: str,
233         t1_start: mp.Event,
234         mp_stop_event: Optional[mp.Event] = None,
235         is_process: bool = False,
236     ):
237         """Processes frames from a queue, performs object detection, and
238         queues the results.
239
240         Consumes frames from 'frame_queue', uses a YOLO model (loaded from
241         'model_path') to detect objects, and then queues the original frame
242         along with detection results in 'detection_queue'.
243         Signals 't1_start' after initializing the model.
244
245         Args:
246             frame_queue: Input queue with frames to process.
247             detection_queue: Output queue for frames with detections.
248             model_path: Path to the YOLO model file.
249             t1_start: Event to signal that model initialization has
250                     finished.
251             mp_stop_event: Optional event to wait before terminating the
252                         process/thread.
253             is_process: Boolean, True if running as a separate process.
254
255         """
256
257         from ultralytics import YOLO
258
259         logging.info(f"Loading model from: {model_path}")
260         model = YOLO(model_path, task="detect")
261
262         # Model warm-up
263         # This can improve the speed of the first real inferences.
264         logging.info("Performing model warm-up...")
265         # It's assumed that the model has these default parameters or they
266         # are configurable.
267         # It's good practice to do warm-up with data similar to the input.
268         # Here a generic configuration is used.
269         try:
270             model(conf=0.5, half=True, imgsz=(640, 640), augment=True,
271                   verbose=False)
272         except Exception as e:
273             logging.warning(f"Error during model warm-up: {e}")
274
275         logging.info("Model loaded and ready. Signaling t1_start.")
276         t1_start.set() # Signal that the model is ready
277
278         while True:
279             item = frame_queue.get()
280             if item is None: # End signal
281                 detection_queue.put(None) # Propagate the signal
282                 logging.info("End signal received in process_frames.")
283                 break
284
285             frame, frame_count = item
286
287             # Perform preprocessing (assuming model.predictor exists and
288             # has these methods)
289             # It's important to verify the API of the ultralytics version
290             # being used.

```



```

339     # Perform tracking
340     # 'result_detections' must be compatible with what 'tracker_wrapper.track' expects
341     tracked_outputs = tracker_wrapper.track(result_detections,
342                                              frame)
343
344     tracking_queue.put((frame, tracked_outputs))
345
346     if mp_stop_event:
347         mp_stop_event.wait()
348
349     if is_process:
350         logging.info("Terminating tracking process.")
351         os._exit(0)
352
353 def tracking_frames_multihardware(
354     self,
355     detection_queue_GPU: Union[mp.Queue, SharedCircularBuffer],
356     detection_queue_DLAO: Union[mp.Queue, SharedCircularBuffer],
357     detection_queue_DLAI: Union[mp.Queue, SharedCircularBuffer],
358     tracking_queue: Union[mp.Queue, SharedCircularBuffer],
359     mp_stop_event: Optional[mp.Event] = None,
360     is_process: bool = False,
361 ):
362     """Performs object tracking from multiple detection queues (multi-hardware).
363
364     Consumes frames and detections from three different queues ('detection_queue_GPU',
365     'detection_queue_DLAI', 'detection_queue_DLAI'), which are assumed to come from
366     different hardware accelerators. Orders frames by their frame number
367     before processing them with 'TrackerWrapper' to maintain temporal coherence.
368     Tracking results are queued in 'tracking_queue'.
369
370     Args:
371         detection_queue_GPU: Detection queue for GPU.
372         detection_queue_DLAI: Detection queue for DLAI.
373         detection_queue_DLAI: Detection queue for DLAI.
374         tracking_queue: Output queue for frames with tracked objects.
375         mp_stop_event: Optional event to wait before terminating the process/thread.
376         is_process: Boolean, True if running as a separate process.
377     """
378     tracker_wrapper = TrackerWrapper(frame_rate=30) # Adjust frame_rate if necessary
379     logging.info("Multi-hardware tracker initialized.")
380
381     # Flags to control if each input queue has finished
382     stop_gpu, stop_dla0, stop_dla1 = False, False, False
383     # Buffers to store the last item read from each queue
384     item_gpu, item_dla0, item_dla1 = None, None, None
385
386     while True:
387         # Try to get a new item from each queue if it's not stopped and the buffer is empty
388         if not stop_gpu and item_gpu is None:
389             item_gpu = detection_queue_GPU.get()
390             if item_gpu is None:
391                 stop_gpu = True
392                 logging.info("GPU queue finished.")

```

```

393     if not stop_dla0 and item_dla0 is None:
394         item_dla0 = detection_queue_DLA0.get()
395         if item_dla0 is None:
396             stop_dla0 = True
397             logging.info("DLA0 queue finished.")
398
399     if not stop_dla1 and item_dla1 is None:
400         item_dla1 = detection_queue_DLA1.get()
401         if item_dla1 is None:
402             stop_dla1 = True
403             logging.info("DLA1 queue finished.")
404
405     # If all input queues have finished, terminate this process/
406     # thread
407     if stop_gpu and stop_dla0 and stop_dla1:
408         tracking_queue.put(None) # Signal the end to the next in
409         the chain
410         logging.info("All detection queues finished. Terminating multi-hardware tracking.")
411         if mp_stop_event:
412             mp_stop_event.wait()
413         if is_process:
414             os._exit(0)
415         break # Exit the while loop
416
417     # Extract frame numbers from current items (if they exist)
418     # The expected item format is (frame, result, times,
419     # frame_number)
420     # A very high value (float('inf')) is used if the item is None
421     # or doesn't have frame_number,
422     # so that valid items have priority.
423     frame_number_gpu = item_gpu[3] if item_gpu else float("inf")
424     frame_number_dla0 = item_dla0[3] if item_dla0 else float("inf")
425     frame_number_dla1 = item_dla1[3] if item_dla1 else float("inf")
426
427     # Select the item with the lowest frame number to process
428     # This ensures frames are processed in chronological order
429     selected_item = None
430     if (
431         frame_number_gpu <= frame_number_dla0
432         and frame_number_gpu <= frame_number_dla1
433         and item_gpu is not None
434     ):
435         selected_item = item_gpu
436         item_gpu = None # Empty the buffer so the next item from
437                     # this queue is read
438     elif (
439         frame_number_dla0 <= frame_number_gpu
440         and frame_number_dla0 <= frame_number_dla1
441         and item_dla0 is not None
442     ):
443         selected_item = item_dla0
444         item_dla0 = None
445     elif (
446         frame_number_dla1 <= frame_number_gpu
447         and frame_number_dla1 <= frame_number_dla0
448         and item_dla1 is not None
449     ):
449         selected_item = item_dla1
450         item_dla1 = None
451     else:
452         # If there are no valid items or all buffers are empty (and
453         # some queue hasn't finished)

```

```

450         # wait a bit to not consume CPU unnecessarily.
451         # This can happen if one queue is much faster than the
452         # others and the others are waiting for data.
453         if (
454             item_gpu is None
455             and item_dla0 is None
456             and item_dla1 is None
457             and not (stop_gpu and stop_dla0 and stop_dla1)
458         ):
459             time.sleep(0.001) # Small pause
460             continue # Return to the beginning of the loop to re-
461             evaluate or read new inputs
462
463             frame, result_detections, _, _ = (
464                 selected_item # times and frame_number are not used
465                 directly here
466             )
467
468             # Perform tracking
469             tracked_outputs = tracker_wrapper.track(result_detections,
470                 frame)
471             tracking_queue.put((frame, tracked_outputs))
472
473     def draw_and_write_frames(
474         self,
475         tracking_queue: Union[mp.Queue, SharedCircularBuffer],
476         times_queue: Union[
477             mp.Queue, SharedCircularBuffer
478         ], # Assume this queue is for benchmark times
479         output_video_path: str,
480         classes: dict, # Mapping of class ID to name
481         memory: dict, # Shared/updated tracking memory
482         colors: dict, # Mapping of class name to color for drawing
483         stop_event: mp.Event, # Event to stop this process/thread
484         tcp_conn_event: mp.Event, # Event to signal TCP connection
485             establishment (renamed from tcp_conn)
486         is_tcp: bool,
487         mp_stop_event: Optional[mp.Event] = None,
488         is_process: bool = False,
489     ):
490         """Draws tracked objects on frames, writes output video and handles
491             TCP communication.
492
493             Consumes frames with tracked objects from 'tracking_queue'. Draws
494             rectangles
495             and labels for each object using information from 'memory' and 'colors'.
496             Writes processed frames to a video file at 'output_video_path'.
497             If 'is_tcp' is True, establishes a TCP server and sends "
498                 DETECTED_DEFECT"
499             messages when a defective object is detected for the first time (
500                 according to 'sended_id').
501             Signals 'tcp_conn_event' after starting the TCP server.
502             Upon completion, puts 'None' in 'times_queue' and activates 'stop_event'.
503
504             Args:
505                 tracking_queue: Input queue with frames and tracked objects.
506                 times_queue: Queue to send a completion signal (or times).
507                 output_video_path: Path to save the output video.
508                 classes: Dictionary mapping class ID to name. (Used indirectly
509                     via 'update_memory')
510                 memory: Tracking memory dictionary.
511                 colors: Dictionary mapping class name to color.

```

```

502     stop_event: Global event to stop all processes/threads in the
503         pipeline.
504     tcp_conn_event: Event to signal that TCP connection is ready.
505     is_tcp: Boolean indicating if operating in TCP mode.
506     mp_stop_event: Optional event to wait before terminating the
507         process/thread.
508     is_process: Boolean, True if running as a separate process.
509     """
510
511     from concurrent.futures import ThreadPoolExecutor
512
513     # ThreadPoolExecutor to handle background tasks (e.g. TCP sending)
514     # max_workers could be adjusted according to expected load.
515     thread_pool = ThreadPoolExecutor(max_workers=8)
516     video_writer = None # Initialize VideoWriter to None
517     frame_number_counter = 0 # Counter for written frames
518
519     # Dictionary to track defect IDs already sent via TCP
520     # to avoid sending multiple messages for the same defect.
521     sended_defect_ids = {}
522
523     client_socket, server_socket = None, None # Initialize sockets
524
525     if is_tcp:
526         try:
527             logging.info("Starting TCP server on 0.0.0.0:8765...")
528             client_socket, server_socket = tcp_server("0.0.0.0", 8765)
529             # Send "READY" in a separate thread to not block.
530             thread_pool.submit(handle_send, client_socket, "READY")
531             tcp_conn_event.set() # Signal that TCP server is ready
532         except Exception as e:
533             raise RuntimeError(f"Error starting TCP server: {e}")
534
535     while True:
536         item = tracking_queue.get()
537         if item is None: # End signal
538             logging.info("End signal received in draw_and_write_frames.")
539             break
540
541         frame, tracked_objects = item
542
543         # Initialize VideoWriter with the first frame to get dimensions
544         if video_writer is None:
545             try:
546                 frame_height, frame_width = frame.shape[:2]
547                 fourcc = cv2.VideoWriter_fourcc(*"mp4v") # Codec for .mp4
548                 video_writer = cv2.VideoWriter(
549                     output_video_path, fourcc, 30, (frame_width,
550                         frame_height))
551                 logging.info(f"VideoWriter initialized for: {output_video_path}")
552             except Exception as e:
553                 logging.error(f"Error initializing VideoWriter: {e}")
554                 break
555
556         # Update memory with current tracked objects
557         self.update_memory(tracked_objects, memory, classes)
558
559         tcp_message_sent_this_frame = False
560
561         # Internal function to draw a single object on the frame

```

```
560     def draw_single_object(obj_data):
561         nonlocal frame, memory, colors, tcp_message_sent_this_frame
562         , is_tcp, client_socket, sended_defect_ids
563
564         # Expected format: (xmin, ymin, xmax, ymax, obj_id, conf,
565         # ...
566         xmin, ymin, xmax, ymax, obj_id = map(int, obj_data[:5])
567         confidence = float(obj_data[5])
568
569         # Confidence threshold to draw the object
570         if confidence < 0.4:
571             return
572
573         # Get updated class and memory state
574         obj_id_in_memory = memory.get(obj_id)
575         if not obj_id_in_memory:
576             return
577
578         current_class_name = obj_id_in_memory["class"]
579         is_currently_defective = current_class_name.endswith("-d")
580
581         # TCP sending logic for defects
582         if (
583             is_tcp
584             and is_currently_defective
585             and not tcp_message_sent_this_frame
586             and not sended_defect_ids.get(obj_id)
587         ):
588             sended_defect_ids[obj_id] = True
589
590             # Send TCP message in a pool thread to not block
591             # drawing
592             thread_pool.submit(handle_send, client_socket, "
593                 DETECTED_DEFECT")
594             tcp_message_sent_this_frame =
595                 True # Mark that a message was already sent in
596                 this frame
597             )
598             logging.debug(f"[TCP] Sending 'DETECTED_DEFECT' for ID {obj_id}")
599
600             # Draw rectangle and text
601             object_color = colors.get(
602                 current_class_name, (255, 255, 255)
603             ) # Default color: white
604             cv2.rectangle(frame, (xmin, ymin), (xmax, ymax),
605             object_color, 2)
606             text_label = f"ID:{obj_id} {current_class_name} {confidence
607             :.2f}"
608             cv2.putText(
609                 frame,
610                 text_label,
611                 (xmin, ymin - 10), # Text position above the rectangle
612                 cv2.FONT_HERSHEY_SIMPLEX,
613                 0.5, # Font size
614                 (255, 255, 255), # Text color (white)
615                 2, # Line thickness
616             )
617
618             # Draw all tracked objects using the thread pool
619             draw_tasks = [thread_pool.submit(draw_single_object, obj) for
620             obj in tracked_objects]
621             for task in draw_tasks: # Wait for all drawing tasks to
622             complete
```

```

task.result()

# Draw frame number on the video
cv2.putText(
    frame,
    f"Frame:{frame_number_counter}",
    (10, 30), # Position
    cv2.FONT_HERSHEY_SIMPLEX,
    1, # Size
    (0, 255, 0), # Color (green)
    2, # Thickness
)

if video_writer:
    video_writer.write(frame)
frame_number_counter += 1

# Finalization and cleanup
if video_writer:
    video_writer.release()

thread_pool.shutdown(wait=True) # Close thread pool waiting for
                                pending tasks to finish

stop_event.set() # Activate global stop event for other processes/
                  threads

if mp_stop_event:
    mp_stop_event.wait()

if is_tcp and client_socket:
    try:
        client_socket.close()
    except Exception as e:
        print(f"Error closing TCP client socket:{e}")

if is_tcp and server_socket:
    try:
        server_socket.close()
    except Exception as e:
        print(f"Error closing TCP server socket:{e}")

if is_process:
    os._exit(0)

def __init__(
    self,
    video_path: str,
    model_path: str,
    output_video_path: str,
    output_times: str,
    parallel_mode: str,
    is_tcp: bool = False,
    sahi: bool = False,
    max_fps: int = None,
    mh_num: int = 1,
    is_process: bool = True,
):
    """Initializes the pipeline with common configuration and control
    events.

Args:
    video_path: Path to the input video file.
    model_path: Path to the detection model file.
    output_video_path: Path to save the processed video.

```

```
675     output_times: Path to save timing/benchmark information.
676     parallel_mode: Parallelization mode (e.g. 'sequential', ,
677                     processes', 'threads').
678     is_tcp: Enables TCP communication for defect notifications.
679     sahi: Enables the use of SAHI (Slice-Aided Hyper Inference) for
680           detection. (Not implemented in this fragment)
681     max_fps: Limits the FPS of video capture.
682     mh_num: Number of handlers/consumers for certain queues (multi-
683               hardware/processing).
684     is_process: Indicates if pipeline components run as separate
685                  processes.
686
687     """
688     self.video_path = video_path
689     self.model_path = model_path
690     self.output_video_path = output_video_path
691     self.output_times = output_times
692     self.parallel_mode = parallel_mode
693     self.is_tcp = is_tcp
694     self.sahi = sahi
695     self.max_fps = max_fps
696     self.mh_num = mh_num
697     self.is_process = is_process
698
699     # Common control events
700     self.tcp_event = mp.Event()
701     self.stop_event = mp.Event()
702     self.t1_start = mp.Event()
703     self.mp_stop_event = mp.Event() if is_process else None
704
705     # Shared memory
706     self.memory = {}
707
708     @abstractmethod
709     def _initialize_queues(self):
710         """Abstract method to initialize communication queues between
711             stages.
712
713             Must be implemented by derived classes to configure queues
714             (e.g. 'mp.Queue', 'SharedCircularBuffer') according to the
715                 parallelization
716             strategy and pipeline type.
717             """
718
719         pass
720
721     @abstractmethod
722     def _initialize_events(self):
723         """Abstract method to initialize pipeline control events.
724
725             Must be implemented by derived classes to configure events
726             necessary for synchronization and pipeline flow control.
727             """
728
729         pass
730
731     @abstractmethod
732     def _create_workers(self):
733         """Abstract method to create pipeline workers.
734             Must be implemented by derived classes to start processes,
735             threads, or any other parallel execution mechanism needed
736             for pipeline stages.
737             """
738
739         pass
740
741     @abstractmethod
742     def _cleanup(self):
```

```
733     """Abstract method for resource cleanup when finishing the pipeline
734     .
735     Must be implemented by derived classes to free resources such as
736     processes, threads, queues, events, or any other handlers opened
737     during pipeline execution.
738     """
739     pass
740
741     @abstractmethod
742     def run(self):
743         """Executes the complete pipeline.
744
745         This is the main method that orchestrates the startup, execution
746         and
747         orderly finalization of all pipeline stages.
748         Must be implemented by derived classes.
749         """
750         pass
```

El segundo archivo es el programa que configura las llamadas a las funciones del objeto `DetectionTrackingPipeline`. Este programa se encarga de inicializar el objeto según el tipo de segmentación elegida, ya sea por hilos, procesos, multiproceso o memoria compartida.

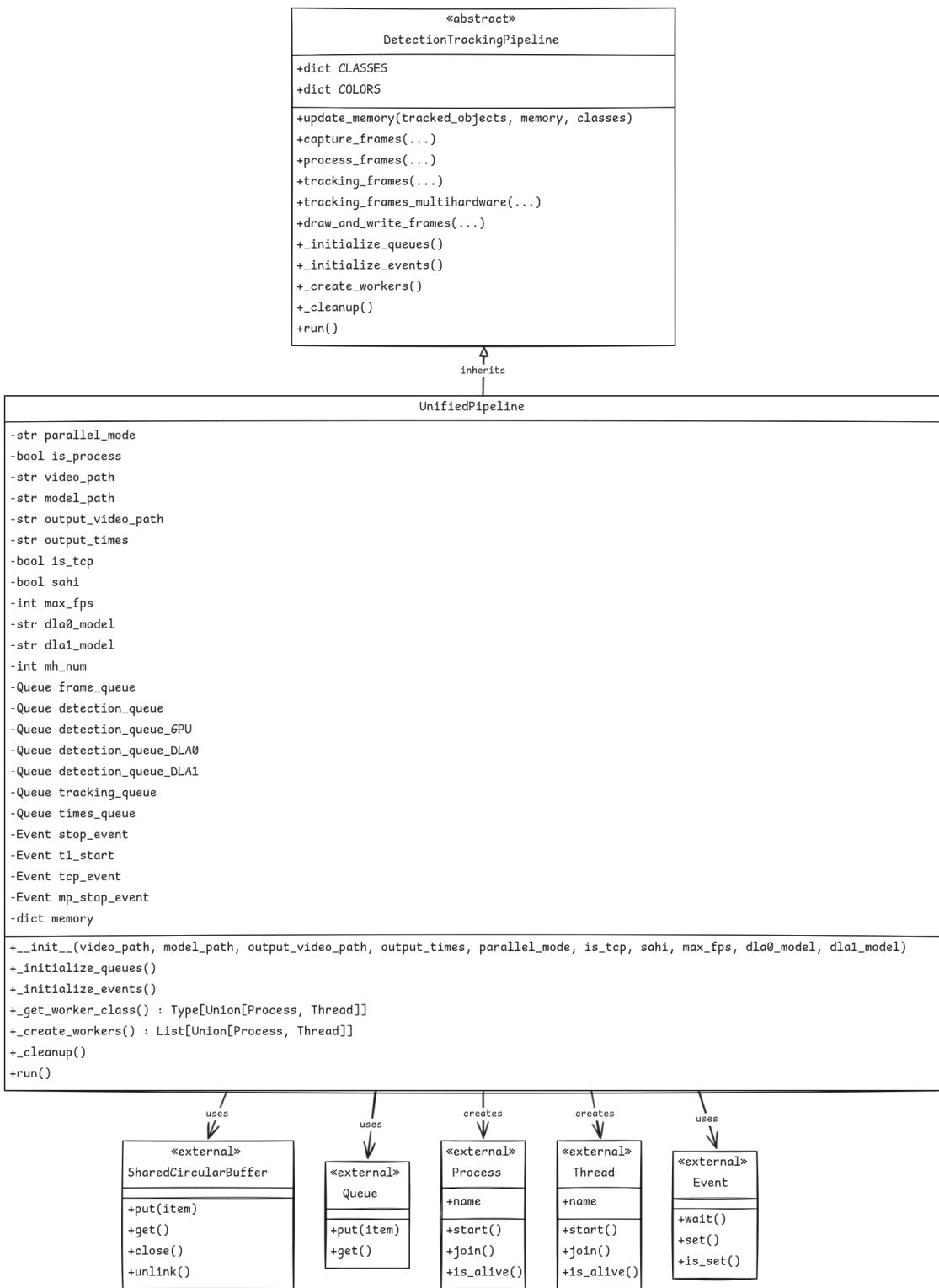


Figura B.2: Diagrama de clases del programa principal.

La figura B.2 muestra el diagrama de clases de objeto UnifiedPipeline. Esta hereda del objeto DetectionTrackingPipeline y añade la funcionalidad de inicializar el objeto según el tipo de segmentación elegida. La función *run* se encarga de iniciar el proceso de captura, procesamiento, seguimiento y escritura de fotogramas. Dependiendo del tipo de segmentación elegida, se ejecutará en un hilo o proceso diferente.

Listing B.2: unified_pipeline.py

```
1 import cv2
2 import torch.multiprocessing as mp
3 from queue import Queue
4 from classes.shared_circular_buffer import SharedCircularBuffer
5 from detection_tracking_pipeline import DetectionTrackingPipeline
6 import threading
7 import logging
8 from typing import Union, List, Type
9
10
11 class UnifiedPipeline(DetectionTrackingPipeline):
12     """Unified pipeline that supports different parallelization
13     strategies.
14
15     This class inherits from 'DetectionTrackingPipeline' and
16     provides a concrete
17     implementation that allows configuring the pipeline to run in
18     different modes:
19     - 'mp.hardware': Uses multiple processes and shared memory
20     queues,
21             optimized for scenarios with multiple hardware
22             accelerators
23             (e.g. GPU, DLA0, DLA1).
24     - 'threads': Uses threads for concurrency within the same
25     process.
26     - 'mp.shared_memory': Uses multiple processes with shared
27     memory queues
28             (SharedCircularBuffer).
29     - Default (any other string): Uses multiple processes with
30     standard
31             'torch.multiprocessing' queues.
32
33     """
34
35     def __init__(
36         self,
37         video_path: str,
38         model_path: str,
39         output_video_path: str,
40         output_times: str,
41         parallel_mode: str,
42         is_tcp: bool = False,
43         sahi: bool = False,
44         max_fps: int = None,
45         dla0_model: str = None,
46         dla1_model: str = None,
47     ):
48         """Initializes the unified pipeline.
49
50         Args:
51             video_path: Path to the input video file.
52             model_path: Path to the main detection model (e.g. for
53                         GPU).
54             output_video_path: Path to save the processed video
55                         with detections.
56             output_times: Path to the CSV file to save processing
57                         times.
58             parallel_mode: Parallelization strategy to use.
59             is_tcp: Boolean to enable TCP communication (e.g. to
60                         notify defects).
```

```

48         sahi: Boolean to enable SAHI (Slice-Aided Hyper
49             Inference).
50         max_fps: Optional. Limits the frames per second of
51             processing.
52         dla0_model: Optional. Path to the model for DLA0
53             accelerator (if 'parallel_mode' is 'mp_hardware').
54         dla1_model: Optional. Path to the model for DLA1
55             accelerator (if 'parallel_mode' is 'mp_hardware').
56         """
57     self.parallel_mode = parallel_mode
58     # Determines if workers will be processes or threads based
59     # on the parallelization mode.
60     self.is_process = parallel_mode != "threads"
61     self.video_path = video_path
62     self.model_path = model_path
63     self.output_video_path = output_video_path
64     self.output_times = output_times
65     self.is_tcp = is_tcp
66     self.sahi = sahi
67     self.max_fps = max_fps
68     self.dla0_model = dla0_model
69     self.dla1_model = dla1_model
70     # mh_num is used to indicate to capture_frames how many ,
71     # None' signals to send
72     # when finishing, so that all frame_queue consumers
73     # terminate.
74     self.mh_num = 1 # Default: one frame consumer (
75         process_frames), if using mp_hardware, it would be 3 (
76         GPU + DLA0 + DLA1)
77
78     self._initialize_queues()
79     self._initialize_events()
80     # Initializes shared memory for object tracking.
81     # This memory is used by 'draw_and_write_frames' and
82     # updated by 'update_memory'.
83     self.memory = {}
84
85     def _initialize_queues(self):
86         """Initializes communication queues between pipeline stages
87         .
88
89         The choice of queue type (standard Queue, mp.Queue,
90             SharedCircularBuffer)
91         and its size is based on 'parallel_mode' and whether 'max_fps'
92         is defined.
93         'SharedCircularBuffer' is used for modes with explicit
94         shared memory.
95         """
96
97         # Queue size: 1 if max_fps is limited (to avoid
98             accumulation), otherwise 10.
99         queue_size = 1 if self.max_fps else 10
100
101         if self.parallel_mode == "mp_hardware":
102             # Multiple detection queues for different hardware (GPU
103                 , DLA0, DLA1)
104             # and a common frame queue. All use
105                 SharedCircularBuffer.
106             logging.info("mp_hardware\u2014mode:\u2014Using\u2014
107                 SharedCircularBuffer\u2014for\u2014all\u2014queues.\")
```

```

89         self.frame_queue = SharedCircularBuffer(
90             queue_size=queue_size, max_item_size=16
91         ) # Arbitrary item size
92         self.detection_queue_GPU = SharedCircularBuffer(
93             queue_size=queue_size, max_item_size=16)
94         self.detection_queue_DLao = SharedCircularBuffer(
95             queue_size=queue_size, max_item_size=16
96         )
97         self.detection_queue_DLao1 = SharedCircularBuffer(
98             queue_size=queue_size, max_item_size=16
99         )
100        self.tracking_queue = SharedCircularBuffer(queue_size=
101            queue_size, max_item_size=16)
102        self.times_queue = SharedCircularBuffer(queue_size=
103            queue_size, max_item_size=16)
104        self.mh_num = 3 # One frame capturer feeds 3 frame
105            processors (GPU, DLao, DLao1)

106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
self.parallel_mode == "threads":
    # Standard 'queue.Queue' queues for communication
    # between threads.
    logging.info("threads mode: Using queue.Queue .")
    self.frame_queue = Queue(maxsize=queue_size)
    self.detection_queue = Queue(maxsize=queue_size)
    self.tracking_queue = Queue(maxsize=queue_size)
    self.times_queue = Queue(maxsize=queue_size)

elif self.parallel_mode == "mp_shared_memory":
    # 'SharedCircularBuffer' queues for multiprocessing
    # with shared memory.
    logging.info("mp_shared_memory mode: Using
    SharedCircularBuffer .")
    self.frame_queue = SharedCircularBuffer(queue_size=
        queue_size, max_item_size=16)
    self.detection_queue = SharedCircularBuffer(queue_size=
        queue_size, max_item_size=16)
    self.tracking_queue = SharedCircularBuffer(queue_size=
        queue_size, max_item_size=16)
    self.times_queue = SharedCircularBuffer(queue_size=
        queue_size, max_item_size=16)

else: # Default mode: standard multiprocessing
    # 'mp.Queue' queues from 'torch.multiprocessing'.
    # If max_fps is defined, the frame queue has size 1 to
    # process frame by frame.
    logging.info("Standard multiprocessing mode: Using mp.
    Queue .")
    self.frame_queue = mp.Queue(maxsize=1) if self.max_fps
        else mp.Queue(maxsize=queue_size)
    self.detection_queue = mp.Queue(maxsize=queue_size)
    self.tracking_queue = mp.Queue(maxsize=queue_size)
    self.times_queue = mp.Queue(maxsize=queue_size)

def _initialize_events(self):
    """Initializes synchronization events.

    All events are from 'torch.multiprocessing.Event'
    regardless of mode,
    since they can be shared between processes if necessary ,
```

```

133     and also work correctly with threads.
134     - 'stop_event': Signal to stop all pipeline workers.
135     - 't1_start': Signal to synchronize the start of timing and
136                     certain operations after model initialization
137
138     - 'tcp_event': Signal to synchronize TCP operations.
139     - 'mp_stop_event': Signal for workers to wait before
140                       exiting,
141                           allowing an orderly shutdown.
142
143     # Multiprocessing events are used for all modes for
144     # consistency
145     # and because they work for both processes and threads.
146     self.stop_event = mp.Event()
147     self.t1_start = mp.Event()
148     self.tcp_event = mp.Event()
149     self.mp_stop_event = (
150         mp.Event()
151     ) # Used for workers to wait before os._exit if they are
152       processes
153
154     def _get_worker_class(self) -> Type[Union[mp.Process, threading.
155     Thread]]:
156         """Determines the base class for workers (process or thread
157         )."""
158
159         Returns:
160             The 'threading.Thread' class if 'parallel_mode' is 'threads',
161             otherwise, 'torch.multiprocessing.Process'.
162
163         if self.parallel_mode == "threads":
164             return threading.Thread
165             return mp.Process
166
167         def _create_workers(self) -> List[Union[mp.Process, threading.
168     Thread]]:
169             """Creates and initializes the list of workers (processes
170             or threads) for the pipeline.
171
172             Each worker is an instance of the class returned by '_get_worker_class()'.
173             The configuration of workers (target functions and
174             arguments) depends
175             on the 'parallel_mode' and whether SAHI or multiple
176             hardware is used.
177
178             Returns:
179                 A list of Worker objects (Process or Thread) ready to
180                 be started.
181
182             Worker = self._get_worker_class()
183             workers: List[Union[mp.Process, threading.Thread]] = []
184
185             # 1. Frame Capture Worker (common to all modes)
186             # This worker reads frames from the video and puts them in
187             # 'frame_queue'.
188             workers.append(
189                 Worker(
190

```

```

178         name="CaptureWorker",
179         target=self.capture_frames,
180         args=(
181             self.video_path,
182             self.frame_queue,
183             self.t1_start, # Event to start capture after
184             others are ready
185             self.stop_event, # Event to stop capture
186             self.tcp_event, # Event for TCP
187             synchronization (if is_tcp)
188             self.is_tcp,
189             self.mp_stop_event, # Event to wait before
190             exiting (if process)
191             self.mh_num, # Number of frame_queue consumers
192             self.is_process, # True if the worker is a
193             process
194             self.max_fps, # FPS limit for capture
195         ),
196     )
197 )
198
199 # 2. Frame Processing Workers (Detection)
200 if self.parallel_mode == "mp_hardware":
201     # Multiple detection workers, one for each specified
202     # model/hardware.
203     # Each consumes from 'frame_queue' and produces to its
204     # 'detection_queue_*'.
205     hardware_setups = [
206         (self.model_path, self.detection_queue_GPU, "GPU"),
207         (self.dla0_model, self.detection_queue_DLA0, "DLA0"
208             ),
209         (self.dla1_model, self.detection_queue_DLA1, "DLA1"
210             ),
211     ]
212     for model_p, detection_q, hw_name in hardware_setups:
213         if model_p: # Only create the worker if a model
214             path was provided
215             workers.append(
216                 Worker(
217                     name=f"ProcessWorker_{hw_name}",
218                     target=self.process_frames_sahi if self
219                     .sahi else self.process_frames,
220                     args=(
221                         self.frame_queue,
222                         detection_q,
223                         model_p,
224                         self.t1_start, # Signals when the
225                         model is ready
226                         self.mp_stop_event,
227                         self.is_process,
228                     ),
229                 )
230             )
231
232             # Multi-Hardware specific Tracking Worker
233             # Consumes from all 'detection_queue_*' and produces to
234             # 'tracking_queue'.
235             workers.append(
236                 Worker(

```

```
225         name="TrackingWorker_MH",
226         target=self.tracking_frames_multihardware,
227         args=(
228             self.detection_queue_GPU,
229             self.detection_queue_DL40,
230             self.detection_queue_DL41,
231             self.tracking_queue,
232             self.mp_stop_event,
233             self.is_process,
234         ),
235     )
236 )
237 else:
238     # Single Frame Processing Worker (Detection)
239     # Consumes from 'frame_queue' and produces to 'detection_queue'.
240     workers.append(
241         Worker(
242             name="ProcessWorker",
243             target=self.process_frames_sahi if self.sahi
244             else self.process_frames,
245             args=(
246                 self.frame_queue,
247                 self.detection_queue,
248                 self.model_path,
249                 self.t1_start,
250                 self.mp_stop_event,
251                 self.is_process,
252             ),
253         )
254     )
255
256     # Standard Tracking Worker
257     # Consumes from 'detection_queue' and produces to 'tracking_queue'.
258     workers.append(
259         Worker(
260             name="TrackingWorker",
261             target=self.tracking_frames,
262             args=(
263                 self.detection_queue,
264                 self.tracking_queue,
265                 self.mp_stop_event,
266                 self.is_process,
267             ),
268         )
269     )
270
271     # 3. Workers common to all modes (Drawing, CSV Writing,
272     # Hardware Usage)
273
274     # Worker to draw detections/tracking on frames and write
275     # the output video.
276     # Consumes from 'tracking_queue' and can interact with TCP.
277     workers.append(
278         Worker(
279             name="DrawWriteWorker",
280             target=self.draw_and_write_frames,
281             args=
```



```

326             queue_buffer.unlink()
327             logging.debug(f"Buffer {i} cleaned.")
328         else:
329             logging.warning(
330                 f"Expected SharedCircularBuffer in "
331                 "cleanup, got {type(queue_buffer)}"
332             )
333     except Exception as e:
334         logging.error(f"Error cleaning buffer {i}: {e}")
335
336     else:
337         logging.info(
338             f"{self.parallel_mode} mode: No manual cleanup "
339             "required for standard queues."
340         )
341     logging.info("Resource cleanup finished.")

342 def run(self):
343     """Runs the unified pipeline.
344
345     This method orchestrates the creation, startup and
346     termination of workers.
347     Measures the total processing time from when 't1_start' is
348     activated
349     (usually after models are ready) until 'stop_event'
350     is activated (usually by 'draw_and_write_frames' when
351     finishing processing).
352     Finally, performs resource cleanup and waits for thread
353     termination
354     if that is the parallelization mode.
355     """
356     logging.info(f"Starting pipeline in mode: {self.
357     parallel_mode}")

358     # Create and start all workers (processes or threads)
359     workers = self._create_workers()
360     for worker in workers:
361         logging.info(f"Starting worker: {worker.name}")
362         worker.start()

363     # Wait for the initialization stage (e.g. model loading in
364     # process_frames)
365     # to activate the t1_start event. This marks the actual
366     # start of measurable processing.
367     logging.info("Waiting for t1_start signal (models ready/
368     measurement start)...")
369     self.t1_start.wait()
370     logging.info("t1_start signal received.")

371     t_start_processing = cv2.getTickCount()

372     # Wait for the pipeline to complete its main task.
373     # 'stop_event' is usually activated by the last worker in
374     # the chain
375     # (draw_and_write_frames) when there are no more frames to
376     # process.
377     logging.info("Pipeline running. Waiting for stop event/
378     signal (end of processing)...")
379     self.stop_event.wait()

```

```

371     logging.info("stop_event signal received.")
372
373     t_end_processing = cv2.getTickCount()
374
375     # Clean up resources (e.g. shared memory queues)
376     # It's important to do this before processes terminate
377     # completely.
378     self._cleanup()
379
380     # Calculate and display performance statistics
381     tiempo_total_segundos = (t_end_processing -
382                               t_start_processing) / cv2.getTickFrequency()
383
384     # Get the total number of frames from the video to
385     # calculate average FPS.
386     # I assume that get_total_frames is a method (possibly
387     # static or instance)
388     # that reads video metadata.
389
390     print(f"Total processing time: {tiempo_total_segundos:.2f} seconds")
391
392     # Wait for all workers to finish, especially important for
393     # threads.
394     # For processes, mp_stop_event and os._exit() handle their
395     # termination.
396     if self.parallel_mode == "threads":
397         logging.info("Waiting for thread termination...")
398         for worker in workers:
399             if worker.is_alive(): # Only join if the thread is
400                 still alive
401                 logging.info(f"Waiting for {worker.name}...")
402                 worker.join(timeout=10) # Add a timeout to
403                 # avoid indefinite blocking
404                 if worker.is_alive():
405                     logging.warning(f"Thread {worker.name} did not finish after timeout.")
406                 else:
407                     logging.info(f"Thread {worker.name} finished.")
408
409     # Signal processes that they can terminate (if they are
410     # waiting for mp_stop_event)
411     if self.is_process:
412         self.mp_stop_event.set()
413
414     print("Pipeline finished.")

```

El tercer archivo define ‘TrackerWrapper’, una clase que encapsula la funcionalidad del algoritmo de seguimiento BYTETrack. Su propósito es simplificar la integración del algoritmo de seguimiento en el pipeline principal.

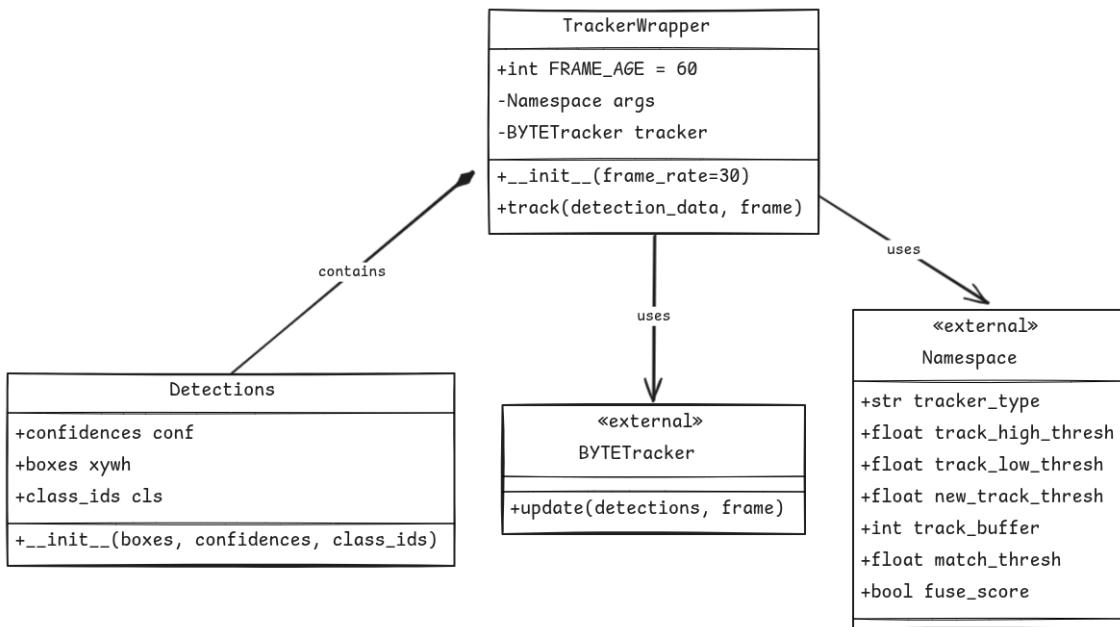


Figura B.3: Diagrama de clases del objeto TrackerWrapper.

La Figura B.3 presenta el diagrama de clases de ‘TrackerWrapper’. A partir de ‘BaseTracker’, esta clase se encarga de inicializar el algoritmo BYTETrack y provee la función ‘track’ para procesar cada fotograma y retornar las trayectorias de los objetos detectados. En esencia, ‘TrackerWrapper’ abstrae la complejidad de la interacción directa con BYTETrack, ofreciendo una interfaz limpia y consistente al resto del sistema.

Listing B.3: `tracker_wrapper.py`

```

1  from argparse import Namespace
2  from ultralytics.trackers.byte_tracker import BYTETracker # type:
3      ignore
4
5
6
7  class TrackerWrapper:
8      FRAME_AGE = 60
9
10     def __init__(self, frame_rate=30):
11         self.args = Namespace(
12             tracker_type="bytetrack",
13             track_high_thresh=0.25,
14             track_low_thresh=0.1,
15             new_track_thresh=0.25,
16             track_buffer=self.FRAME_AGE,
17             match_thresh=0.8,
18             fuse_score=True,
19         )
20         self.tracker = BYTETracker(self.args, frame_rate=frame_rate
21             )
22
23     class Detections:
24         def __init__(self, boxes, confidences, class_ids):
25             self.conf = confidences
26             self.xywh = boxes
27             self.cls = class_ids
  
```

```
27
28     def track(self, detection_data, frame):
29         detections = self.Detections(
30             detection_data.xywh.numpy(),
31             detection_data.conf.numpy(),
32             detection_data.cls.numpy().astype(int),
33         )
34         return self.tracker.update(detections, frame)
```

El cuarto archivo es el objeto ‘SharedCircularBuffer’, que implementa un buffer circular con memoria compartida. Este buffer permite crear la abstracción de una cola de mensajes que puede ser utilizada por diferentes procesos que comparten memoria.

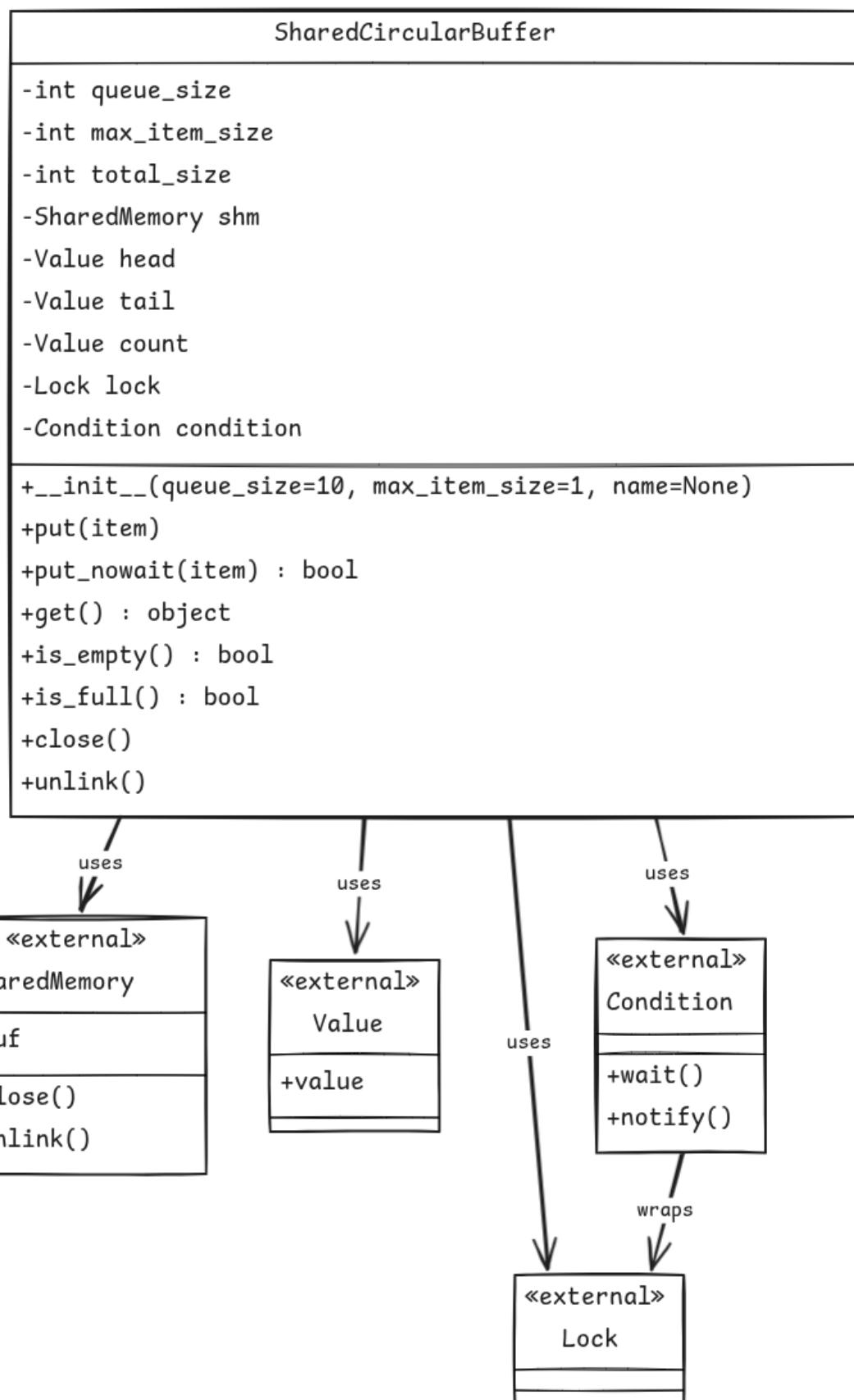


Figura B.4: Diagrama de clases del objeto SharedCircularBuffer.

La Figura B.4 muestra el diagrama de clases del objeto ‘SharedCircularBuffer’. Este objeto implementa las funciones para inicializar el buffer, escribir y leer mensajes, y gestionar la memoria compartida. El buffer circular permite que los procesos que comparten memoria puedan comunicarse de manera eficiente, evitando bloqueos y garantizando un acceso seguro a los datos.

Listing B.4: shared_circular_buffer.py

```

1 import pickle
2 import numpy as np
3 from multiprocessing import shared_memory, Value, Lock, Condition
4
5
6 class SharedCircularBuffer:
7     def __init__(self, queue_size=10, max_item_size=1, name=None):
8         """
9             Initializes a circular buffer in shared memory.
10
11             :param queue_size: Maximum number of elements in the queue.
12             :param max_item_size: Maximum size in MB of each element.
13             :param name: Name of the shared memory (None to create a
14                         new one).
15         """
16         self.queue_size = queue_size
17
18         if max_item_size not in (1, 2, 4, 8, 16, 32, 64, 128, 256,
19                                  512):
20             raise ValueError(
21                 "The maximum item size must be 1, 2, 4, 8, 16, 32, 64, 128, 256 or 512 MB."
22             )
23
24         self.max_item_size = max_item_size * 1024 * 1024
25         self.total_size = queue_size * self.max_item_size
26
27         if name:
28             self.shm = shared_memory.SharedMemory(name=name)
29         else:
30             self.shm = shared_memory.SharedMemory(create=True, size
31                                                 =self.total_size)
32
33         self.head = Value("i", 0) # Read index
34         self.tail = Value("i", 0) # Write index
35         self.count = Value("i", 0) # Number of elements in the
36                                   queue
37         self.lock = Lock()
38         self.condition = Condition(self.lock) # For
39                                   synchronization
40
41     def put(self, item):
42         """
43             Adds an item to the queue in shared memory without
44             overwriting unread elements.
45         """
46         if isinstance(item, np.ndarray):
47             reshaped_item = item.reshape(-1)
48             item_data = {"data": reshaped_item, "shape": item.shape
49                         }
50         else:
51             item_data = {"data": item, "shape": None}
52
53         with self.lock:
54             while self.count.value == self.queue_size:
55                 self.condition.wait()
56
57             self.tail.value += 1
58             self.shm.write(item_data)
59
59             self.condition.notify()
60
61             self.count.value += 1
62
63     def get(self):
64         """
65             Returns the first item in the queue.
66         """
67         with self.lock:
68             while self.count.value == 0:
69                 self.condition.wait()
70
71             self.head.value += 1
72             item_data = self.shm.read()
73
74             self.condition.notify()
75
76             self.count.value -= 1
77
78             return item_data["data"]
79
80     def clear(self):
81         """
82             Clears the queue by setting the head and tail pointers
83             to the same value.
84         """
85         self.head.value = self.tail.value
86
87         self.condition.notify()
88
89     def __del__(self):
90         self.shm.close()
91
92

```

```
44     data_bytes = pickle.dumps(item_data)
45
46     if len(data_bytes) > self.max_item_size:
47         raise ValueError("The item is too large for the queue.")
48
49     with self.condition:
50         while self.count.value == self.queue_size:
51             # The queue is full, wait until there is space
52             self.condition.wait()
53
54         pos = (self.tail.value % self.queue_size) * self.
55             max_item_size
56         self.shm.buf[pos : pos + len(data_bytes)] = memoryview(
57             data_bytes)
58
59         self.tail.value = (self.tail.value + 1) % self.
60             queue_size
61         self.count.value += 1
62
63         self.condition.notify() # Notify 'get' that a new
64             element is available
65
66     def put_nowait(self, item):
67         """Adds an item to the queue in shared memory. If the queue
68             is full, returns False without adding the item."""
69
70         if isinstance(item, np.ndarray):
71             reshaped_item = item.reshape(-1)
72             item_data = {"data": reshaped_item, "shape": item.shape}
73         else:
74             item_data = {"data": item, "shape": None}
75
76         data_bytes = pickle.dumps(item_data)
77
78         if len(data_bytes) > self.max_item_size:
79             raise ValueError("The item is too large for the queue.")
80
81         with self.condition:
82             if self.count.value >= self.queue_size:
83                 return False # Queue is full, cannot add the item
84
85             pos = (self.tail.value % self.queue_size) * self.
86                 max_item_size
87             self.shm.buf[pos : pos + len(data_bytes)] = memoryview(
88                 data_bytes)
89
90             self.tail.value = (self.tail.value + 1) % self.
91                 queue_size
92             self.count.value += 1
93
94             self.condition.notify() # Notify 'get' that a new
95                 element is available
96
97             return True
```

```

91     def get(self):
92         """Extracts an item from the queue in shared memory,
93             waiting if it is empty."""
94         with self.condition:
95             while self.count.value == 0: # Wait if the queue is
96                 empty
97                 self.condition.wait()
98
99             pos = (self.head.value % self.queue_size) * self.
100                max_item_size
101            data_bytes = bytes(self.shm.buf[pos : pos + self.
102                max_item_size])
103
104            self.head.value = (self.head.value + 1) % self.
105                queue_size
106            self.count.value -= 1
107
108            self.condition.notify() # Notify 'put' that space is
109                available
110
111            item_data = pickle.loads(data_bytes)
112
113        if item_data["shape"] is not None:
114            return np.array(item_data["data"]).reshape(item_data["
115                shape"])
116
117        return item_data["data"]
118
119    def is_empty(self):
120        """Returns True if the queue is empty."""
121        with self.lock:
122            return self.count.value == 0
123
124    def is_full(self):
125        """Returns True if the queue is full."""
126        with self.lock:
127            return self.count.value == self.queue_size
128
129    def close(self):
130        """Closes the shared memory."""
131        self.shm.close()
132
133    def unlink(self):
134        """Releases the shared memory (should only be called once).
135            """
136        self.shm.unlink()

```

Por último, el quinto archivo es el programa que se encarga de proporcionar los argumentos necesarios para ejecutar el programa principal. Este programa permite al usuario elegir todas las opciones de configuración del pipeline, como el tipo de segmentación, el modelo de detección, la precisión, el tamaño del modelo, el dispositivo a utilizar y la tasa de fotogramas por segundo.

Listing B.5: inference.py

```

1  """
2  Object Detection and Tracking Inference Pipeline
3

```

```
4 This module provides a comprehensive inference pipeline for object
5     detection and tracking
6 using YOLO models with support for different hardware configurations,
7     parallelization
8 modes.
9
10 Features:
11 - Multiple YOLO model variants (YOLOv8, YOLOv11, etc.)
12 - Hardware acceleration (GPU, DLA0, DLA1, CPU)
13 - Various precision modes (FP32, FP16, INT8)
14 - Parallelization strategies (threads, multiprocessing, shared memory)
15 - Configurable FPS limits
16 """
17
18 import argparse
19 import os
20 import torch.multiprocessing as mp
21 from unified_pipeline import UnifiedPipeline
22
23 def parse_arguments():
24     """
25         Parse command line arguments for the inference pipeline.
26
27     Returns:
28         argparse.Namespace: Parsed command line arguments containing:
29             - num_objects: Number of objects to count in video
30             - model: YOLO model variant to use
31             - precision: Model precision (FP32, FP16, INT8)
32             - hardware: Target hardware (GPU, DLA0, DLA1, CPU)
33             - mode: Power mode configuration
34             - tcp: Whether to use TCP communication
35             - version: Dataset version
36             - parallel: Parallelization strategy
37             - max_fps: Maximum FPS limit
38
39     parser = argparse.ArgumentParser(description="Object-detection-and-
40                                     tracking-inference-pipeline")
41
42     # Object counting configuration
43     parser.add_argument(
44         "--num_objects",
45         default="free",
46         type=str,
47         choices=["free", "variable", "0", "18", "40", "48", "60", "70",
48             "88", "176"],
49         help="Number of objects to count in the video, default=free",
50     )
51
52     # Model selection
53     parser.add_argument(
54         "--model",
55         default="yolo11n",
56         type=str,
57         choices=[
58             "yolo11n",
59             "yolo11s",
60             "yolo11m",
61             "yolo11l",
62             "yolo11x",
63             "yolov5nu",
64             "yolov5mu",
65             "yolov8n",
66             "yolov8s",
67         ],
68     )
69
70     # Video input configuration
71     parser.add_argument(
72         "--video",
73         default=None,
74         type=str,
75         help="Path to the video file to process, default=None"
76     )
77
78     # Output configuration
79     parser.add_argument(
80         "--output",
81         default=None,
82         type=str,
83         help="Path to save the output results, default=None"
84     )
85
86     # Logging and metrics
87     parser.add_argument(
88         "--log_level",
89         default="INFO",
90         type=str,
91         choices=["INFO", "WARNING", "ERROR"]
92     )
93
94     # Parallelization and performance
95     parser.add_argument(
96         "--parallel",
97         default="multiprocessing",
98         type=str,
99         choices=["threads", "multiprocessing", "shared_memory"]
100    )
101
102    # Performance and constraints
103    parser.add_argument(
104        "--max_fps",
105        default=None,
106        type=int,
107        help="Maximum FPS limit, default=None"
108    )
109
110    # Precision and hardware
111    parser.add_argument(
112        "--precision",
113        default="FP32",
114        type=str,
115        choices=["FP32", "FP16", "INT8"]
116    )
117
118    # Power mode
119    parser.add_argument(
120        "--mode",
121        default="standard",
122        type=str,
123        choices=["standard", "low_power"]
124    )
125
126    # TCP communication
127    parser.add_argument(
128        "--tcp",
129        default=False,
130        type=bool,
131        help="Whether to use TCP communication, default=False"
132    )
133
134    # Dataset version
135    parser.add_argument(
136        "--version",
137        default="latest",
138        type=str,
139        choices=["latest", "v1", "v2", "v3"]
140    )
141
142    # Parallelization strategy
143    parser.add_argument(
144        "--parallel",
145        default="multiprocessing",
146        type=str,
147        choices=["threads", "multiprocessing", "shared_memory"]
148    )
149
150    # Shared memory
151    parser.add_argument(
152        "--shared_memory",
153        default=False,
154        type=bool,
155        help="Use shared memory for parallel processing, default=False"
156    )
157
158    # GPU usage
159    parser.add_argument(
160        "--gpu",
161        default=None,
162        type=int,
163        help="Index of the GPU to use, default=None"
164    )
165
166    # DLA usage
167    parser.add_argument(
168        "--dla",
169        default=None,
170        type=int,
171        help="Index of the DLA to use, default=None"
172    )
173
174    # CPU usage
175    parser.add_argument(
176        "--cpu",
177        default=False,
178        type=bool,
179        help="Use CPU for inference, default=False"
180    )
181
182    # DLA0 usage
183    parser.add_argument(
184        "--dla0",
185        default=False,
186        type=bool,
187        help="Use DLA0 for inference, default=False"
188    )
189
190    # DLA1 usage
191    parser.add_argument(
192        "--dla1",
193        default=False,
194        type=bool,
195        help="Use DLA1 for inference, default=False"
196    )
197
198    # Model variants
199    parser.add_argument(
200        "--model",
201        default="yolo11n",
202        type=str,
203        choices=[
204            "yolo11n",
205            "yolo11s",
206            "yolo11m",
207            "yolo11l",
208            "yolo11x",
209            "yolov5nu",
210            "yolov5mu",
211            "yolov8n",
212            "yolov8s",
213        ],
214    )
215
216    # Model weights
217    parser.add_argument(
218        "--weights",
219        default=None,
220        type=str,
221        help="Path to the model weights, default=None"
222    )
223
224    # Model configuration
225    parser.add_argument(
226        "--config",
227        default=None,
228        type=str,
229        help="Path to the model configuration file, default=None"
230    )
231
232    # Model parameters
233    parser.add_argument(
234        "--params",
235        default=None,
236        type=str,
237        help="Path to the model parameters file, default=None"
238    )
239
240    # Model quantization
241    parser.add_argument(
242        "--quantize",
243        default=False,
244        type=bool,
245        help="Quantize the model, default=False"
246    )
247
248    # Model optimization
249    parser.add_argument(
250        "--optimize",
251        default=False,
252        type=bool,
253        help="Optimize the model, default=False"
254    )
255
256    # Model pruning
257    parser.add_argument(
258        "--prune",
259        default=False,
260        type=bool,
261        help="Prune the model, default=False"
262    )
263
264    # Model fusion
265    parser.add_argument(
266        "--fuse",
267        default=False,
268        type=bool,
269        help="Fuse the model, default=False"
270    )
271
272    # Model quantization
273    parser.add_argument(
274        "--quantize",
275        default=False,
276        type=bool,
277        help="Quantize the model, default=False"
278    )
279
280    # Model optimization
281    parser.add_argument(
282        "--optimize",
283        default=False,
284        type=bool,
285        help="Optimize the model, default=False"
286    )
287
288    # Model pruning
289    parser.add_argument(
290        "--prune",
291        default=False,
292        type=bool,
293        help="Prune the model, default=False"
294    )
295
296    # Model fusion
297    parser.add_argument(
298        "--fuse",
299        default=False,
300        type=bool,
301        help="Fuse the model, default=False"
302    )
303
304    # Model quantization
305    parser.add_argument(
306        "--quantize",
307        default=False,
308        type=bool,
309        help="Quantize the model, default=False"
310    )
311
312    # Model optimization
313    parser.add_argument(
314        "--optimize",
315        default=False,
316        type=bool,
317        help="Optimize the model, default=False"
318    )
319
320    # Model pruning
321    parser.add_argument(
322        "--prune",
323        default=False,
324        type=bool,
325        help="Prune the model, default=False"
326    )
327
328    # Model fusion
329    parser.add_argument(
330        "--fuse",
331        default=False,
332        type=bool,
333        help="Fuse the model, default=False"
334    )
335
336    # Model quantization
337    parser.add_argument(
338        "--quantize",
339        default=False,
340        type=bool,
341        help="Quantize the model, default=False"
342    )
343
344    # Model optimization
345    parser.add_argument(
346        "--optimize",
347        default=False,
348        type=bool,
349        help="Optimize the model, default=False"
350    )
351
352    # Model pruning
353    parser.add_argument(
354        "--prune",
355        default=False,
356        type=bool,
357        help="Prune the model, default=False"
358    )
359
360    # Model fusion
361    parser.add_argument(
362        "--fuse",
363        default=False,
364        type=bool,
365        help="Fuse the model, default=False"
366    )
367
368    # Model quantization
369    parser.add_argument(
370        "--quantize",
371        default=False,
372        type=bool,
373        help="Quantize the model, default=False"
374    )
375
376    # Model optimization
377    parser.add_argument(
378        "--optimize",
379        default=False,
380        type=bool,
381        help="Optimize the model, default=False"
382    )
383
384    # Model pruning
385    parser.add_argument(
386        "--prune",
387        default=False,
388        type=bool,
389        help="Prune the model, default=False"
390    )
391
392    # Model fusion
393    parser.add_argument(
394        "--fuse",
395        default=False,
396        type=bool,
397        help="Fuse the model, default=False"
398    )
399
400    # Model quantization
401    parser.add_argument(
402        "--quantize",
403        default=False,
404        type=bool,
405        help="Quantize the model, default=False"
406    )
407
408    # Model optimization
409    parser.add_argument(
410        "--optimize",
411        default=False,
412        type=bool,
413        help="Optimize the model, default=False"
414    )
415
416    # Model pruning
417    parser.add_argument(
418        "--prune",
419        default=False,
420        type=bool,
421        help="Prune the model, default=False"
422    )
423
424    # Model fusion
425    parser.add_argument(
426        "--fuse",
427        default=False,
428        type=bool,
429        help="Fuse the model, default=False"
430    )
431
432    # Model quantization
433    parser.add_argument(
434        "--quantize",
435        default=False,
436        type=bool,
437        help="Quantize the model, default=False"
438    )
439
440    # Model optimization
441    parser.add_argument(
442        "--optimize",
443        default=False,
444        type=bool,
445        help="Optimize the model, default=False"
446    )
447
448    # Model pruning
449    parser.add_argument(
450        "--prune",
451        default=False,
452        type=bool,
453        help="Prune the model, default=False"
454    )
455
456    # Model fusion
457    parser.add_argument(
458        "--fuse",
459        default=False,
460        type=bool,
461        help="Fuse the model, default=False"
462    )
463
464    # Model quantization
465    parser.add_argument(
466        "--quantize",
467        default=False,
468        type=bool,
469        help="Quantize the model, default=False"
470    )
471
472    # Model optimization
473    parser.add_argument(
474        "--optimize",
475        default=False,
476        type=bool,
477        help="Optimize the model, default=False"
478    )
479
480    # Model pruning
481    parser.add_argument(
482        "--prune",
483        default=False,
484        type=bool,
485        help="Prune the model, default=False"
486    )
487
488    # Model fusion
489    parser.add_argument(
490        "--fuse",
491        default=False,
492        type=bool,
493        help="Fuse the model, default=False"
494    )
495
496    # Model quantization
497    parser.add_argument(
498        "--quantize",
499        default=False,
500        type=bool,
501        help="Quantize the model, default=False"
502    )
503
504    # Model optimization
505    parser.add_argument(
506        "--optimize",
507        default=False,
508        type=bool,
509        help="Optimize the model, default=False"
510    )
511
512    # Model pruning
513    parser.add_argument(
514        "--prune",
515        default=False,
516        type=bool,
517        help="Prune the model, default=False"
518    )
519
520    # Model fusion
521    parser.add_argument(
522        "--fuse",
523        default=False,
524        type=bool,
525        help="Fuse the model, default=False"
526    )
527
528    # Model quantization
529    parser.add_argument(
530        "--quantize",
531        default=False,
532        type=bool,
533        help="Quantize the model, default=False"
534    )
535
536    # Model optimization
537    parser.add_argument(
538        "--optimize",
539        default=False,
540        type=bool,
541        help="Optimize the model, default=False"
542    )
543
544    # Model pruning
545    parser.add_argument(
546        "--prune",
547        default=False,
548        type=bool,
549        help="Prune the model, default=False"
550    )
551
552    # Model fusion
553    parser.add_argument(
554        "--fuse",
555        default=False,
556        type=bool,
557        help="Fuse the model, default=False"
558    )
559
560    # Model quantization
561    parser.add_argument(
562        "--quantize",
563        default=False,
564        type=bool,
565        help="Quantize the model, default=False"
566    )
567
568    # Model optimization
569    parser.add_argument(
570        "--optimize",
571        default=False,
572        type=bool,
573        help="Optimize the model, default=False"
574    )
575
576    # Model pruning
577    parser.add_argument(
578        "--prune",
579        default=False,
580        type=bool,
581        help="Prune the model, default=False"
582    )
583
584    # Model fusion
585    parser.add_argument(
586        "--fuse",
587        default=False,
588        type=bool,
589        help="Fuse the model, default=False"
590    )
591
592    # Model quantization
593    parser.add_argument(
594        "--quantize",
595        default=False,
596        type=bool,
597        help="Quantize the model, default=False"
598    )
599
600    # Model optimization
601    parser.add_argument(
602        "--optimize",
603        default=False,
604        type=bool,
605        help="Optimize the model, default=False"
606    )
607
608    # Model pruning
609    parser.add_argument(
610        "--prune",
611        default=False,
612        type=bool,
613        help="Prune the model, default=False"
614    )
615
616    # Model fusion
617    parser.add_argument(
618        "--fuse",
619        default=False,
620        type=bool,
621        help="Fuse the model, default=False"
622    )
623
624    # Model quantization
625    parser.add_argument(
626        "--quantize",
627        default=False,
628        type=bool,
629        help="Quantize the model, default=False"
630    )
631
632    # Model optimization
633    parser.add_argument(
634        "--optimize",
635        default=False,
636        type=bool,
637        help="Optimize the model, default=False"
638    )
639
640    # Model pruning
641    parser.add_argument(
642        "--prune",
643        default=False,
644        type=bool,
645        help="Prune the model, default=False"
646    )
647
648    # Model fusion
649    parser.add_argument(
650        "--fuse",
651        default=False,
652        type=bool,
653        help="Fuse the model, default=False"
654    )
655
656    # Model quantization
657    parser.add_argument(
658        "--quantize",
659        default=False,
660        type=bool,
661        help="Quantize the model, default=False"
662    )
663
664    # Model optimization
665    parser.add_argument(
666        "--optimize",
667        default=False,
668        type=bool,
669        help="Optimize the model, default=False"
670    )
671
672    # Model pruning
673    parser.add_argument(
674        "--prune",
675        default=False,
676        type=bool,
677        help="Prune the model, default=False"
678    )
679
680    # Model fusion
681    parser.add_argument(
682        "--fuse",
683        default=False,
684        type=bool,
685        help="Fuse the model, default=False"
686    )
687
688    # Model quantization
689    parser.add_argument(
690        "--quantize",
691        default=False,
692        type=bool,
693        help="Quantize the model, default=False"
694    )
695
696    # Model optimization
697    parser.add_argument(
698        "--optimize",
699        default=False,
700        type=bool,
701        help="Optimize the model, default=False"
702    )
703
704    # Model pruning
705    parser.add_argument(
706        "--prune",
707        default=False,
708        type=bool,
709        help="Prune the model, default=False"
710    )
711
712    # Model fusion
713    parser.add_argument(
714        "--fuse",
715        default=False,
716        type=bool,
717        help="Fuse the model, default=False"
718    )
719
720    # Model quantization
721    parser.add_argument(
722        "--quantize",
723        default=False,
724        type=bool,
725        help="Quantize the model, default=False"
726    )
727
728    # Model optimization
729    parser.add_argument(
730        "--optimize",
731        default=False,
732        type=bool,
733        help="Optimize the model, default=False"
734    )
735
736    # Model pruning
737    parser.add_argument(
738        "--prune",
739        default=False,
740        type=bool,
741        help="Prune the model, default=False"
742    )
743
744    # Model fusion
745    parser.add_argument(
746        "--fuse",
747        default=False,
748        type=bool,
749        help="Fuse the model, default=False"
750    )
751
752    # Model quantization
753    parser.add_argument(
754        "--quantize",
755        default=False,
756        type=bool,
757        help="Quantize the model, default=False"
758    )
759
760    # Model optimization
761    parser.add_argument(
762        "--optimize",
763        default=False,
764        type=bool,
765        help="Optimize the model, default=False"
766    )
767
768    # Model pruning
769    parser.add_argument(
770        "--prune",
771        default=False,
772        type=bool,
773        help="Prune the model, default=False"
774    )
775
776    # Model fusion
777    parser.add_argument(
778        "--fuse",
779        default=False,
780        type=bool,
781        help="Fuse the model, default=False"
782    )
783
784    # Model quantization
785    parser.add_argument(
786        "--quantize",
787        default=False,
788        type=bool,
789        help="Quantize the model, default=False"
790    )
791
792    # Model optimization
793    parser.add_argument(
794        "--optimize",
795        default=False,
796        type=bool,
797        help="Optimize the model, default=False"
798    )
799
800    # Model pruning
801    parser.add_argument(
802        "--prune",
803        default=False,
804        type=bool,
805        help="Prune the model, default=False"
806    )
807
808    # Model fusion
809    parser.add_argument(
810        "--fuse",
811        default=False,
812        type=bool,
813        help="Fuse the model, default=False"
814    )
815
816    # Model quantization
817    parser.add_argument(
818        "--quantize",
819        default=False,
820        type=bool,
821        help="Quantize the model, default=False"
822    )
823
824    # Model optimization
825    parser.add_argument(
826        "--optimize",
827        default=False,
828        type=bool,
829        help="Optimize the model, default=False"
830    )
831
832    # Model pruning
833    parser.add_argument(
834        "--prune",
835        default=False,
836        type=bool,
837        help="Prune the model, default=False"
838    )
839
840    # Model fusion
841    parser.add_argument(
842        "--fuse",
843        default=False,
844        type=bool,
845        help="Fuse the model, default=False"
846    )
847
848    # Model quantization
849    parser.add_argument(
850        "--quantize",
851        default=False,
852        type=bool,
853        help="Quantize the model, default=False"
854    )
855
856    # Model optimization
857    parser.add_argument(
858        "--optimize",
859        default=False,
860        type=bool,
861        help="Optimize the model, default=False"
862    )
863
864    # Model pruning
865    parser.add_argument(
866        "--prune",
867        default=False,
868        type=bool,
869        help="Prune the model, default=False"
870    )
871
872    # Model fusion
873    parser.add_argument(
874        "--fuse",
875        default=False,
876        type=bool,
877        help="Fuse the model, default=False"
878    )
879
880    # Model quantization
881    parser.add_argument(
882        "--quantize",
883        default=False,
884        type=bool,
885        help="Quantize the model, default=False"
886    )
887
888    # Model optimization
889    parser.add_argument(
890        "--optimize",
891        default=False,
892        type=bool,
893        help="Optimize the model, default=False"
894    )
895
896    # Model pruning
897    parser.add_argument(
898        "--prune",
899        default=False,
900        type=bool,
901        help="Prune the model, default=False"
902    )
903
904    # Model fusion
905    parser.add_argument(
906        "--fuse",
907        default=False,
908        type=bool,
909        help="Fuse the model, default=False"
910    )
911
912    # Model quantization
913    parser.add_argument(
914        "--quantize",
915        default=False,
916        type=bool,
917        help="Quantize the model, default=False"
918    )
919
920    # Model optimization
921    parser.add_argument(
922        "--optimize",
923        default=False,
924        type=bool,
925        help="Optimize the model, default=False"
926    )
927
928    # Model pruning
929    parser.add_argument(
930        "--prune",
931        default=False,
932        type=bool,
933        help="Prune the model, default=False"
934    )
935
936    # Model fusion
937    parser.add_argument(
938        "--fuse",
939        default=False,
940        type=bool,
941        help="Fuse the model, default=False"
942    )
943
944    # Model quantization
945    parser.add_argument(
946        "--quantize",
947        default=False,
948        type=bool,
949        help="Quantize the model, default=False"
950    )
951
952    # Model optimization
953    parser.add_argument(
954        "--optimize",
955        default=False,
956        type=bool,
957        help="Optimize the model, default=False"
958    )
959
960    # Model pruning
961    parser.add_argument(
962        "--prune",
963        default=False,
964        type=bool,
965        help="Prune the model, default=False"
966    )
967
968    # Model fusion
969    parser.add_argument(
970        "--fuse",
971        default=False,
972        type=bool,
973        help="Fuse the model, default=False"
974    )
975
976    # Model quantization
977    parser.add_argument(
978        "--quantize",
979        default=False,
980        type=bool,
981        help="Quantize the model, default=False"
982    )
983
984    # Model optimization
985    parser.add_argument(
986        "--optimize",
987        default=False,
988        type=bool,
989        help="Optimize the model, default=False"
990    )
991
992    # Model pruning
993    parser.add_argument(
994        "--prune",
995        default=False,
996        type=bool,
997        help="Prune the model, default=False"
998    )
999
1000   # Model fusion
1001  parser.add_argument(
1002      "--fuse",
1003      default=False,
1004      type=bool,
1005      help="Fuse the model, default=False"
1006  )
1007
1008  # Model quantization
1009  parser.add_argument(
1010     "--quantize",
1011     default=False,
1012     type=bool,
1013     help="Quantize the model, default=False"
1014  )
1015
1016  # Model optimization
1017  parser.add_argument(
1018     "--optimize",
1019     default=False,
1020     type=bool,
1021     help="Optimize the model, default=False"
1022  )
1023
1024  # Model pruning
1025  parser.add_argument(
1026     "--prune",
1027     default=False,
1028     type=bool,
1029     help="Prune the model, default=False"
1030  )
1031
1032  # Model fusion
1033  parser.add_argument(
1034     "--fuse",
1035     default=False,
1036     type=bool,
1037     help="Fuse the model, default=False"
1038  )
1039
1040  # Model quantization
1041  parser.add_argument(
1042     "--quantize",
1043     default=False,
1044     type=bool,
1045     help="Quantize the model, default=False"
1046  )
1047
1048  # Model optimization
1049  parser.add_argument(
1050     "--optimize",
1051     default=False,
1052     type=bool,
1053     help="Optimize the model, default=False"
1054  )
1055
1056  # Model pruning
1057  parser.add_argument(
1058     "--prune",
1059     default=False,
1060     type=bool,
1061     help="Prune the model, default=False"
1062  )
1063
1064  # Model fusion
1065  parser.add_argument(
1066     "--fuse",
1067     default=False,
1068     type=bool,
1069     help="Fuse the model, default=False"
1070  )
1071
1072  # Model quantization
1073  parser.add_argument(
1074     "--quantize",
1075     default=False,
1076     type=bool,
1077     help="Quantize the model, default=False"
1078  )
1079
1080  # Model optimization
1081  parser.add_argument(
1082     "--optimize",
1083     default=False,
1084     type=bool,
1085     help="Optimize the model, default=False"
1086  )
1087
1088  # Model pruning
1089  parser.add_argument(
1090     "--prune",
1091     default=False,
1092     type=bool,
1093     help="Prune the model, default=False"
1094  )
1095
1096  # Model fusion
1097  parser.add_argument(
1098     "--fuse",
1099     default=False,
1100     type=bool,
1101     help="Fuse the model, default=False"
1102  )
1103
1104  # Model quantization
1105  parser.add_argument(
1106     "--quantize",
1107     default=False,
1108     type=bool,
1109     help="Quantize the model, default=False"
1110  )
1111
1112  # Model optimization
1113  parser.add_argument(
1114     "--optimize",
1115     default=False,
1116     type=bool,
1117     help="Optimize the model, default=False"
1118  )
1119
1120  # Model pruning
1121  parser.add_argument(
1122     "--prune",
1123     default=False,
1124     type=bool,
1125     help="Prune the model, default=False"
1126  )
1127
1128  # Model fusion
1129  parser.add_argument(
1130     "--fuse",
1131     default=False,
1132     type=bool,
1133     help="Fuse the model, default=False"
1134  )
1135
1136  # Model quantization
1137  parser.add_argument(
1138     "--quantize",
1139     default=False,
1140     type=bool,
1141     help="Quantize the model, default=False"
1142  )
1143
1144  # Model optimization
1145  parser.add_argument(
1146     "--optimize",
1147     default=False,
1148     type=bool,
1149     help="Optimize the model, default=False"
1150  )
1151
1152  # Model pruning
1153  parser.add_argument(
1154     "--prune",
1155     default=False,
1156     type=bool,
1157     help="Prune the model, default=False"
1158  )
1159
1160  # Model fusion
1161  parser.add_argument(
1162     "--fuse",
1163     default=False,
1164     type=bool,
1165     help="Fuse the model, default=False"
1166  )
1167
1168  # Model quantization
1169  parser.add_argument(
1170     "--quantize",
1171     default=False,
1172     type=bool,
1173     help="Quantize the model, default=False"
1174  )
1175
1176  # Model optimization
1177  parser.add_argument(
1178     "--optimize",
1179     default=False,
1180     type=bool,
1181     help="Optimize the model, default=False"
1182  )
1183
1184  # Model pruning
1185  parser.add_argument(
1186     "--prune",
1187     default=False,
1188     type=bool,
1189     help="Prune the model, default=False"
1190  )
1191
1192  # Model fusion
1193  parser.add_argument(
1194     "--fuse",
1195     default=False,
1196     type=bool,
1197     help="Fuse the model, default=False"
1198  )
1199
1200  # Model quantization
1201  parser.add_argument(
1202     "--quantize",
1203     default=False,
1204     type=bool,
1205     help="Quantize the model, default=False"
1206  )
1207
1208  # Model optimization
1209  parser.add_argument(
1210     "--optimize",
1211     default=False,
1212     type=bool,
1213     help="Optimize the model, default=False"
1214  )
1215
1216  # Model pruning
1217  parser.add_argument(
1218     "--prune",
1219     default=False,
1220     type=bool,
1221     help="Prune the model, default=False"
1222  )
1223
1224  # Model fusion
1225  parser.add_argument(
1226     "--fuse",
1227     default=False,
1228     type=bool,
1229     help="Fuse the model, default=False"
1230  )
1231
1232  # Model quantization
1233  parser.add_argument(
1234     "--quantize",
1235     default=False,
1236     type=bool,
1237     help="Quantize the model, default=False"
1238  )
1239
1240  # Model optimization
1241  parser.add_argument(
1242     "--optimize",
1243     default=False,
1244     type=bool,
1245     help="Optimize the model, default=False"
1246  )
1247
1248  # Model pruning
1249  parser.add_argument(
1250     "--prune",
1251     default=False,
1252     type=bool,
1253     help="Prune the model, default=False"
1254  )
1255
1256  # Model fusion
1257  parser.add_argument(
1258     "--fuse",
1259     default=False,
1260     type=bool,
1261     help="Fuse the model, default=False"
1262  )
1263
1264  # Model quantization
1265  parser.add_argument(
1266     "--quantize",
1267     default=False,
1268     type=bool,
1269     help="Quantize the model, default=False"
1270  )
1271
1272  # Model optimization
1273  parser.add_argument(
1274     "--optimize",
1275     default=False,
1276     type=bool,
1277     help="Optimize the model, default=False"
1278  )
1279
1280  # Model pruning
1281  parser.add_argument(
1282     "--prune",
1283     default=False,
1284     type=bool,
1285     help="Prune the model, default=False"
1286  )
1287
1288  # Model fusion
1289  parser.add_argument(
1290     "--fuse",
1291     default=False,
1292     type=bool,
1293     help="Fuse the model, default=False"
1294  )
1295
1296  # Model quantization
1297  parser.add_argument(
1298     "--quantize",
1299     default=False,
1300     type=bool,
1301     help="Quantize the model, default=False"
1302  )
1303
1304  # Model optimization
1305  parser.add_argument(
1306     "--optimize",
1307     default=False,
1308     type=bool,
1309     help="Optimize the model, default=False"
1310  )
1311
1312  # Model pruning
1313  parser.add_argument(
1314     "--prune",
1315     default=False,
1316     type=bool,
1317     help="Prune the model, default=False"
1318  )
1319
1320  # Model fusion
1321  parser.add_argument(
1322     "--fuse",
1323     default=False,
1324     type=bool,
1325     help="Fuse the model, default=False"
1326  )
1327
1328  # Model quantization
1329  parser.add_argument(
1330     "--quantize",
1331     default=False,
1332     type=bool,
1333     help="Quantize the model, default=False"
1334  )
1335
1336  # Model optimization
1337  parser.add_argument(
1338     "--optimize",
1339     default=False,
1340     type=bool,
1341     help="Optimize the model, default=False"
1342  )
1343
1344  # Model pruning
1345  parser.add_argument(
1346     "--prune",
1347     default=False,
1348     type=bool,
1349     help="Prune the model, default=False"
1350  )
1351
1352  # Model fusion
1353  parser.add_argument(
1354     "--fuse",
1355     default=False,
1356     type=bool,
1357     help="Fuse the model, default=False"
1358  )
1359
1360  # Model quantization
1361  parser.add_argument(
1362     "--quantize",
1363     default=False,
1364     type=bool,
1365     help="Quantize the model, default=False"
1366  )
1367
1368  # Model optimization
1369  parser.add_argument(
1370     "--optimize",
1371     default=False,
1372     type=bool,
1373     help="Optimize the model, default=False"
1374  )
1375
1376  # Model pruning
1377  parser.add_argument(
1378     "--prune",
1379     default=False,
1380     type=bool,
1381     help="Prune the model, default=False"
1382  )
1383
1384  # Model fusion
1385  parser.add_argument(
1386     "--fuse",
1387     default=False,
1388     type=bool,
1389     help="Fuse the model, default=False"
1390  )
1391
1392  # Model quantization
1393  parser.add_argument(
1394     "--quantize",
1395     default=False,
1396     type=bool,
1
```

```

64     ],
65     help="YOLO model variant to use, default=yolo1in",
66 )
67
68 # Model precision configuration
69 parser.add_argument(
70     "--precision",
71     default="FP16",
72     type=str,
73     choices=["FP32", "FP16", "INT8"],
74     help="Model precision format, default=FP16",
75 )
76
77 # Hardware acceleration target
78 parser.add_argument(
79     "--hardware",
80     default="GPU",
81     type=str,
82     choices=["GPU", "DLA0", "DLA1", "ALL", "CPU"],
83     help="Target hardware for inference, default=GPU",
84 )
85
86 # Power mode configuration
87 parser.add_argument(
88     "--mode",
89     required=True,
90     type=str,
91     choices=["MAXN", "30W", "15W", "10W"],
92     help="Power mode configuration (required)",
93 )
94
95 # Network communication
96 parser.add_argument(
97     "--tcp", default=False, type=bool, help="Enable TCP communication",
98     default=False
99 )
100
101 # Dataset version
102 parser.add_argument(
103     "--version",
104     default="2025_02_24",
105     type=str,
106     choices=["2025_02_24", "2024_11_28"],
107     help="Dataset version to use, default=2025_02_24",
108 )
109
110 # Parallelization strategy
111 parser.add_argument(
112     "--parallel",
113     default="mp_shared_memory",
114     type=str,
115     choices=["threads", "mp", "mp_shared_memory", "mp_hardware"],
116     help="Parallelization strategy, default=mp_shared_memory",
117 )
118
119 # FPS limiting
120 parser.add_argument(
121     "--max_fps",
122     default=None,
123     type=int,
124     help="Maximum FPS limit for processing, default=None (unlimited)",
125 )
126
127     return parser.parse_args()

```

```

127
128
129 def initialize_pipeline(args):
130     """Initialize the detection and tracking pipeline according to
131         parallelization mode."""
132     mode = f"{args.mode}_{mp.cpu_count()}CORE"
133     model_name = args.model
134
135     batch_size = 1
136     batch_suffix = f"_batch{batch_size}" if batch_size > 1 else ""
137
138     # Define model paths for different hardware configurations
139     base_path = (
140         f"../../models/canicas/{args.version}/\u2022\
141         {args.version}_canicas_{model_name}_{args.precision}"
142     )
143
144     GPU_model_path = f"{base_path}_GPU{batch_suffix}.engine"
145     DLA0_model_path = f"{base_path}_DLA0{batch_suffix}.engine"
146     DLA1_model_path = f"{base_path}_DLA1{batch_suffix}.engine"
147     CPU_model_path = f"../../models/canicas/\u2022\
148         {args.version}/{args.version}_canicas_{model_name}.pt"
149
150     model_path = (
151         GPU_model_path
152         if args.hardware == "GPU"
153         else (
154             DLA0_model_path
155             if args.hardware == "DLA0"
156             else DLA1_model_path if args.hardware == "DLA1" else
157                 CPU_model_path
158         )
159     )
160
161     # Select video path based on number of objects
162     if args.num_objects == "free":
163         video_path = "../../datasets_labeled/videos/\u2022\
164         contar_objetos_variable_2min.mp4"
165     else:
166         video_path = f"../../datasets_labeled/videos/\u2022\
167         contar_objetos_{args.num_objects}_2min.mp4"
168
169     output_dir = "../../inference_predictions/custom_tracker"
170
171     os.makedirs(output_dir, exist_ok=True)
172
173     output_video_path = os.path.join(
174         output_dir,
175         f"{args.parallel}_{model_name}_{args.precision}_{args.hardware}_\
176         f'{args.num_objects}_objects_{mode}.mp4",
177     )
178     fps_prefix = f"_-{args.max_fps}fps" if args.max_fps else "maxfps"
179     output_times = (
180         f"{model_name}_{args.precision}_{args.hardware}_\
181         f'{args.num_objects}_objects_{mode}{fps_prefix}'"
182     )
183
184     print("\n\n[PROGRAM] \u2022Selected\u2022options:", args, "\n\n")
185
186     if not args.parallel in ["threads", "mp", "mp_shared_memory",
187         "mp.hardware"]:
188         raise ValueError(
189             "Invalid\u2022parallelization\u2022mode.\u2022Must\u2022be\u2022'threads', '\u2022'mp', '\u2022'\
190             '\u2022'mp_shared_memory'\u2022or\u2022'mp.hardware'."
191

```

```
188     )
189
190     # Create a unified pipeline instance
191     if args.parallel == "mp.hardware":
192         pipeline = UnifiedPipeline(
193             video_path,
194             GPU_model_path,
195             output_video_path,
196             output_times,
197             args.parallel,
198             is_tcp=args.tcp,
199             max_fps=args.max_fps,
200             dla0_model=DLA0_model_path,
201             dla1_model=DLA1_model_path,
202         )
203     else:
204         pipeline = UnifiedPipeline(
205             video_path,
206             model_path,
207             output_video_path,
208             output_times,
209             args.parallel,
210             is_tcp=args.tcp,
211             max_fps=args.max_fps,
212         )
213
214     return pipeline
215
216
217 def main():
218     args = parse_arguments()
219     detection_tracking_pipeline = initialize_pipeline(args)
220     detection_tracking_pipeline.run()
221
222
223 if __name__ == "__main__":
224     mp.set_start_method("spawn")
225     print(f"[PROGRAM] Number of CPUs: {mp.cpu_count()}")
226     main()
```