UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Detección y seguimiento de individuos en entornos complejos.

Sebastián Josué Endara Revelo

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Ciencias de la Computación

Quito, 13 de mayo de 2025

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Detección y seguimiento de individuos en entornos complejos.

Sebastián Josué Endara Revelo

Nombre del profesor, Título académico **Noel Pérez Pérez, Ph.D.**

Quito, 13 de mayo de 2025

© DERECHOS DE AUTOR

Nombres y apellidos: Sebastián Josué Endara Revelo

Código: 00323096

Cédula de identidad: 1719319905

Lugar y fecha: Quito, 13 de mayo de 2025

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Esta investigación aborda el desafío de detectar, seguir y analizar la saturación de individuos en entornos complejos utilizando técnicas de visión por computadora. El objetivo principal es desarrollar un sistema de detección automatizado que maximice la precisión en la identificación de individuos mientras mantiene la eficiencia computacional. Para lograr esto, se implementaron y evaluaron de manera comparativa tres modelos de detección de objetos de última generación: YOLOv10, YOLOv11 y YOLOv12. Los modelos fueron entrenados y validados utilizando un conjunto de datos derivado de grabaciones de cámaras de vigilancia universitarias, aplicando una estrategia de validación cruzada de 10 pliegues para asegurar una evaluación robusta del rendimiento. Entre los modelos evaluados, YOLOv10 con un umbral de confianza de 0.5 resultó ser la configuración óptima, ofreciendo el mejor equilibrio entre precisión y eficiencia. Alcanzó un mAP@0.5 de 0.9132, un mAP@0.5–0.95 de 0.7385, una Precisión de 0.9131 y un Recall de 0.8644. Estas métricas destacan la efectividad del modelo para detectar individuos con precisión en escenarios de vigilancia. El sistema propuesto establece las bases para un análisis automatizado de saturación, permitiendo la toma de decisiones basada en datos en aplicaciones como el monitoreo de espacios públicos y la planificación de seguridad. Un proceso de evaluación riguroso, que incluyó pruebas de significancia estadística y análisis de costos computacionales, guió la selección del modelo. Los resultados demuestran que el modelo YOLOv10 seleccionado ofrece una solución confiable y eficiente para la detección de individuos, contribuyendo al avance de los sistemas de vigilancia inteligente.

Palabras clave: Deep Learning, visión computacional, detección de objetos, análisis de densidad de multitudes, YOLO, videovigilancia.

ABSTRACT

This research tackles the challenge of detecting, tracking, and analyzing the saturation of individuals in complex environments using computer vision techniques. The primary objective is to develop an automated detection system that maximizes individual identification accuracy while maintaining computational efficiency. To achieve this, three state-of-the-art object detection models—YOLOv10, YOLOv11, and YOLOv12—were implemented and comparatively evaluated. The models were trained and validated using a dataset derived from university surveillance camera footage, applying a 10-fold cross-validation strategy to ensure robust performance assessment. Among the evaluated models, YOLOv10 with a confidence threshold of 0.5 emerged as the optimal configuration, offering the best trade-off between accuracy and efficiency. It achieved a mAP@0.5 of 0.9132, a mAP@0.5–0.95 of 0.7385, a Precision of 0.9131, and a Recall of 0.8644. These metrics highlight the model's effectiveness in accurately detecting individuals in real-world surveillance scenarios. The proposed system lays the groundwork for automated saturation analysis, enabling data-driven decision-making in applications such as public space monitoring and safety planning. A rigorous evaluation process—including statistical significance testing and computational cost analysis—guided the model selection. The results demonstrate that the chosen YOLOv10 model offers a reliable and efficient solution for individual detection, contributing to the advancement of intelligent surveillance systems.

Key words: Deep Learning, computer vision, object detection, YOLO, crowd density analysis, surveillance.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

El análisis de individuos en entornos poblados es esencial para la gestión de espacios y la garantía de la seguridad en grandes instituciones como las universidades. El creciente número de personas en estas áreas plantea desafíos para la planificación eficaz de eventos, el control de acceso y la respuesta a emergencias. La necesidad de sistemas robustos y automatizados se ve aún más subrayada por la creciente sofisticación de las técnicas de vigilancia, incluido el Resumen y Seguimiento de Objetos en Video (VOST), que requieren una detección de objetos inicial precisa y confiable como un prerrequisito fundamental. Por lo tanto, la transición a soluciones de visión por computadora, particularmente aquellas que utilizan aprendizaje profundo, ofrece una vía para superar estas limitaciones y proporcionar las herramientas necesarias para protocolos eficaces de gestión de multitudes y seguridad. Esta transición exige una mirada más atenta a los avances recientes en la detección de objetos, que forman la base de muchas aplicaciones de análisis de video.

La detección de objetos ha experimentado avances significativos gracias a la visión por computadora. Redmon et al. (2016) introdujeron el modelo *You Only Look Once* (YOLO), que proporciona un equilibrio entre velocidad y precisión en la detección de objetos en tiempo real y ha sido ampliamente adoptado. Kirillov et al. (2023) presentaron el Modelo de Segmentación de Cualquier Cosa (SAM, por sus siglas en inglés), que permite la segmentación automatizada de objetos en imágenes y videos para simplificar la identificación de regiones relevantes. Además, Ravi et al. (2024) desarrollaron SAM 2, una extensión de SAM, para mejorar el rendimiento de la segmentación en secuencias de video. Hua et al. (2024) abordan la detección de comportamiento anómalo en peatones mediante la introducción de YOLO-ABD, un método ligero que mejora YOLOv8n con detección de objetos pequeños, reorganización de canales (channel shuffling) y mecanismos de atención para mejorar la precisión en escenarios

complejos. Su evaluación en el conjunto de datos IITB-Corridor logró una puntuación mAP50 del 89.3 %. Jiao y Abdullah (2024) propusieron DP-YOLO para mejorar la detección de peatones en escenas concurridas, centrándose en desafíos como objetivos pequeños y oclusión, al reemplazar las convoluciones estándar con convoluciones deformables y utilizar Varifocal Loss, logrando un mAP50 del 89.7 %. Estas puntuaciones mAP50 de YOLO-ABD y DP-YOLO representan puntos de referencia sólidos para evaluar el rendimiento de los modelos de detección de personas, particularmente en escenarios desafiantes.

Además, se han desarrollado enfoques de seguimiento de múltiples objetos como una extensión de los detectores de objetos para mejorar la precisión del conteo de personas. Por ejemplo, Bewley et al. (2016) propusieron el método SORT, que utiliza un filtro de Kalman para un seguimiento eficiente en tiempo real. Wojke et al. (2017) lo mejoraron con DeepSORT, incorporando una métrica de aprendizaje profundo para la asociación a fin de mantener la consistencia de la identidad. Además, Du et al. (2023) introdujeron StrongSORT, que optimiza la asignación de objetos mediante una métrica de asociación refinada, reduciendo errores en escenarios con alta densidad de objetos. Su (2024) comparó varios métodos de seguimiento basados en aprendizaje profundo, enfatizando el beneficio de combinar detección y segmentación para una mayor precisión. Lohani et al. (2022) revisaron los sistemas de videovigilancia para la detección de intrusiones, destacando la importancia de técnicas de seguimiento robustas en aplicaciones de seguridad.Investigadores han explorado el análisis de la densidad de multitudes con técnicas destinadas a mejorar la precisión del conteo de personas. Su (2024) comparó varios métodos de seguimiento basados en Aprendizaje Profundo, enfatizando el beneficio de combinar detección y segmentación para una mayor precisión. Lohani et al. (2022) revisaron los sistemas de videovigilancia para la detección de intrusiones, destacando la importancia de técnicas de seguimiento robustas en aplicaciones de seguridad.

A pesar de los recientes avances en esta área, la detección de objetos únicos y múltiples en entornos poblados sigue siendo un desafío debido a las condiciones críticas de los individuos y el hardware estándar, como luces, ángulos, posiciones, oclusiones y movimientos, entre otros. Esto sugiere la necesidad de desarrollar modelos más eficientes y adaptables para diversos entornos. Por lo tanto, en este trabajo, se propone la exploración de tres versiones diferentes de modelos YOLO para maximizar la detección de individuos en escenarios concurridos. La principal contribución detrás de este objetivo se relaciona con la implementación de un pipeline de detección con aprendizaje transferido de modelos YOLO de vanguardia.

El resto del documento se organiza de la siguiente manera: la sección de Materiales y Métodos describe la base de datos de video empleada, el método propuesto con el esquema de transferencia y el fine tuning del pipeline propuesto, y la configuración experimental diseñada para entrenar, validar y probar el método propuesto utilizando diferentes métricas de detección. La sección de Resultados y Discusión presenta los resultados más relevantes en términos de rendimiento de detección, la selección del mejor modelo YOLO y su validación en el conjunto de prueba externo para medir su poder de generalización en datos no vistos. Finalmente, se presentan las conclusiones, destacando los logros más importantes del método desarrollado y sus futuras extensiones.

2. DESARROLLO DEL TEMA

2.1. Materiales y Métodos

2.1.1 Base de datos

Se utilizó una base de datos privada, que consta de 1729 imágenes extraídas de diez videos de vigilancia aleatorios de diferentes cámaras ubicadas en todo el campus universitario. Los videos tienen una duración que oscila entre 0 y 150 segundos, representando los escenarios más críticos en un entorno concurrido.

2.1.2 Método Propuesto

El método propuesto explora tres versiones pre-entrenadas del *framework* de detección de objetos YOLO utilizando la arquitectura medium: YOLOv10, YOLOv11 y YOLOv12. Estos modelos son ampliamente reconocidos por su rendimiento en tiempo real y su alta precisión en la detección de múltiples objetos dentro de una sola pasada hacia adelante. YOLO formula la detección de objetos como un problema de regresión, prediciendo directamente los cuadros delimitadores y las probabilidades de clase a partir de la imagen completa utilizando una única red neuronal convolucional (Ultralytics, 2024). La arquitectura medium se seleccionó sobre las variantes small, large y extra-large debido a su equilibrio óptimo entre eficiencia computacional y precisión de detección. Si bien la versión small es más rápida y ligera, a menudo tiene un rendimiento inferior en escenas complejas con objetos pequeños o superpuestos. Por el contrario, las versiones large y extra-large ofrecen una mayor precisión, pero incurren en una sobrecarga computacional significativa, lo que puede dificultar el procesamiento en tiempo real en hardware GPU estándar. La configuración medium ofrece así un compromiso práctico: es computacionalmente tratable para su implementación en entornos del mundo real, al tiempo que mantiene capacidades de detección robustas. A continuación, se proporciona una breve descripción de los modelos seleccionados:

- YOLOv10 introduce un cambio de paradigma en la detección de objetos al eliminar la etapa tradicional de Supresión No Máxima (NMS) y adoptar una estrategia de asignación dual para la selección de cuadros delimitadores. Este reemplazo simplifica el *pipeline* de post-procesamiento y reduce significativamente la latencia de inferencia, lo que hace que el modelo sea más adecuado para aplicaciones en tiempo real en dispositivos *edge*. Además, YOLOv10 incorpora *classification heads* (cabezales de clasificación) ligeras, que reducen la sobrecarga computacional sin comprometer el rendimiento de la detección. La arquitectura también aplica técnicas para minimizar la pérdida de información durante la extracción y agregación de características, preservando la calidad de las representaciones de objetos en diferentes escalas. En general, YOLOv10 ofrece un enfoque más eficiente y optimizado para la detección de objetos al desacoplar el post-procesamiento complejo, manteniendo una alta precisión (Ultralytics, 2024).
- YOLOv11 se basa en el diseño ligero de YOLOv10, pero enfatiza las mejoras arquitectónicas que optimizan tanto la velocidad como la precisión. Introduce nuevos bloques de red que capturan mejor las jerarquías espaciales y las características semánticas, lo que lleva a una detección mejorada de objetos en diferentes escalas. Estos nuevos módulos están optimizados para mantener un bajo costo computacional al tiempo que mejoran la capacidad de representación del modelo. YOLOv11 también refina el proceso de fusión de características en múltiples niveles, lo que resulta en una mejor localización y clasificación de objetos en escenas desafiantes. El equilibrio logrado entre la eficiencia de la inferencia y la precisión hace que YOLOv11 sea particularmente efectivo para tareas de visión de alto rendimiento donde la consistencia del rendimiento es crítica (Ultralytics, 2024).

- YOLOv12 se centra en la incorporación de mecanismos de atención para mejorar la capacidad del modelo de comprender tanto los contextos locales como globales dentro de una imagen. Esta versión integra módulos de atención avanzados que guían a la red para que se centre dinámicamente en las regiones más informativas de la entrada, lo que permite una detección de objetos superior en entornos desordenados o parcialmente ocluidos. YOLOv12 también optimiza la agregación de características multiescala a través de estrategias de fusión eficientes, mejorando la adaptabilidad del modelo a entradas visuales complejas. En comparación con sus predecesores, YOLOv12 demuestra mejoras significativas en la precisión, particularmente en escenarios con objetos pequeños o densamente empaquetados. Su equilibrio inteligente entre la eficiencia computacional y la robustez de la detección lo posiciona como una solución de vanguardia para las aplicaciones modernas de visión por computadora (Ultralytics, 2024).

Estos modelos se transfirieron al método propuesto para evitar el significativo costo computacional y el tiempo asociado con el entrenamiento de detectores de objetos profundos desde cero. El entrenamiento de modelos de alto rendimiento como YOLOv10–12 desde cero generalmente requiere conjuntos de datos anotados masivos y amplios recursos de GPU, lo que puede no ser factible en muchos escenarios prácticos. Al aprovechar el aprendizaje transferido, explotamos el conocimiento previo codificado en modelos pre-entrenados en el conjunto de datos COCO, que es ampliamente considerado como uno de los *benchmarks* de detección de objetos más completos debido a su diversidad en categorías de objetos, oclusiones, contextos y escalas. Este pre-entrenamiento proporciona una representación de características sólida y generalizable, lo que mejora significativamente la inicialización del modelo y el comportamiento de convergencia durante el ajuste fino. Posteriormente, ajustamos los modelos

transferidos en nuestro propio conjunto de datos específico del dominio para adaptarlos y especializarlos para el problema en estudio, como se ilustra en la Figura 1.

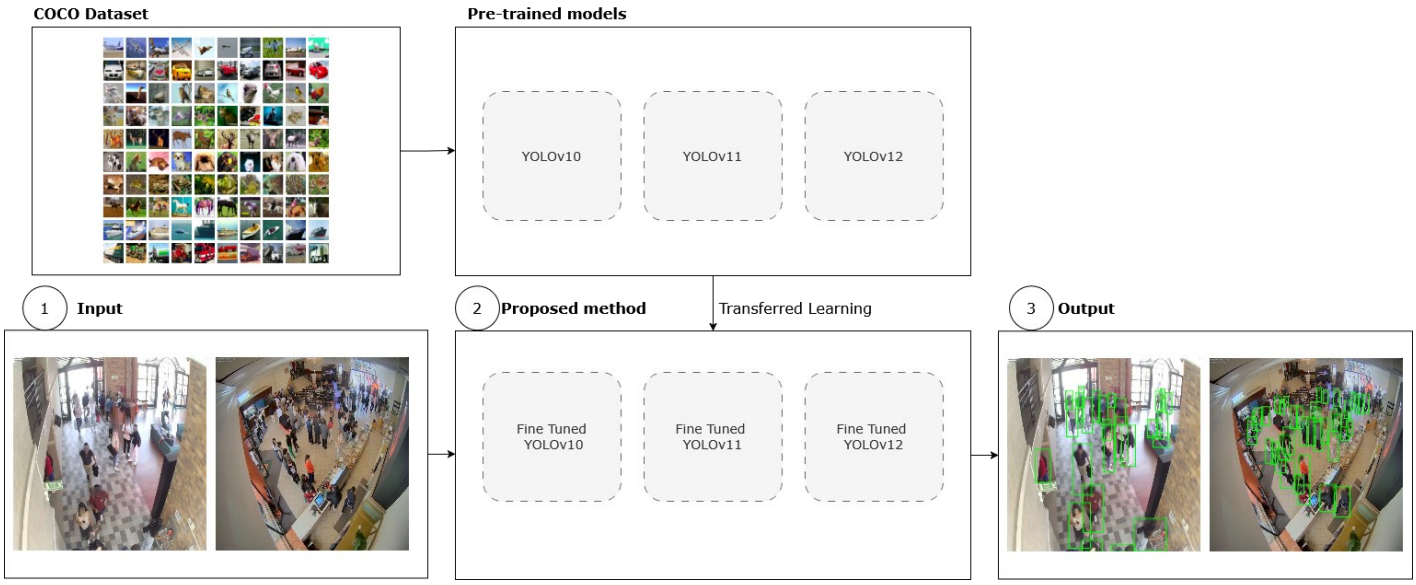

Figura 1. Flujo de trabajo del método propuesto

2.2. Configuración experimental

Esta sección detalla la configuración experimental empleada para llevar a cabo la investigación, describiendo los pasos de procesamiento y etiquetado de datos, las configuraciones del modelo, las métricas de evaluación, los criterios de selección y la implementación.

2.2.1 Procesamiento y etiquetado de datos

Cada video de la base de datos se separó en 60 fotogramas por segundo para formar un total de 1729 imágenes. Luego, todas las imágenes se sometieron a un proceso de etiquetado en el que se anotaron manualmente cuadros delimitadores alrededor de los individuos presentes en cada fotograma. Para acelerar esta anotación, se realizó un paso inicial de autoetiquetado utilizando el modelo Florence 2, que generó automáticamente las coordenadas de los cuadros

delimitadores. Estas etiquetas autogeneradas fueron posteriormente refinadas y corregidas mediante una revisión manual utilizando la herramienta Roboflow para asegurar la precisión y consistencia de las anotaciones.

2.2.2 Conjuntos de entrenamiento, validación y prueba

El conjunto de datos preparado se dividió en conjuntos de entrenamiento y prueba utilizando una división del 85%-15% para facilitar el desarrollo y la evaluación del modelo. La partición de entrenamiento alimenta un método de validación cruzada de diez pliegues para garantizar particiones de entrenamiento y validación disjuntas durante la etapa de entrenamiento del método propuesto. El uso de la estrategia de validación cruzada asegura que los modelos se entrenen con una porción sustancial de los datos, al tiempo que se reserva un conjunto de prueba independiente para evaluar su rendimiento de generalización en datos no vistos.

2.2.3 Configuración del modelo

Todos los modelos YOLO se entrenaron con un conjunto consistente de parámetros para aislar el efecto de las diferencias arquitectónicas. El entrenamiento abarcó 200 épocas, donde una época representa un pase completo del conjunto de datos de entrenamiento a través del modelo. Se consideró que este número de épocas era suficiente para que los modelos aprendieran los patrones subyacentes en los datos. Se utilizó el optimizador AdamW para actualizar los pesos del modelo, aprovechando su eficiencia y efectividad en tareas de aprendizaje profundo. Para mitigar el sobreajuste, se empleó una estrategia de parada temprana, monitoreando el rendimiento de validación y deteniendo el entrenamiento si no se observaba mejora durante 20 épocas consecutivas. El tamaño del lote se estableció en 16 para equilibrar la eficiencia computacional y las limitaciones de memoria. Además, validamos el rendimiento del modelo en un rango de umbrales de confianza, que varían de 0.5 a 0.9. El umbral de

confianza determina la probabilidad mínima que una detección debe tener para ser considerada válida, y la evaluación en un rango de umbrales permitió el análisis del compromiso precisión-recall.

2.2.4 Métricas de evaluación y selección del modelo

Calculamos la media de mAP50, mAP, precisión (PRE) y recall (REC) a través de diferentes umbrales de confianza durante el proceso de entrenamiento. Sin embargo, la media de mAP50 se seleccionó como la medida principal porque ofrece un estándar transparente e interpretable para la precisión de la detección de objetos, recompensando las predicciones que logran al menos una superposición moderada (IoU $\geq$ 0.5) con la verdad fundamental. A diferencia del PRE o REC brutos, que reflejan aspectos aislados del rendimiento, la media de mAP50 integra tanto la calidad de la localización como la corrección de la detección. Además, se calculó la prueba de rangos con signo de Wilcoxon con un alfa de 0.05 para comparar estadísticamente el rendimiento de la detección entre modelos.

El criterio de selección se basó en la siguiente regla: el modelo que maximiza la media de la puntuación mAP50 se tomó como el pivote de comparación estadística. Si otros modelos muestran un rendimiento estadísticamente significativo similar al pivote, pasan a la siguiente fase. En la fase posterior, se priorizaron las configuraciones con umbrales de confianza más bajos para favorecer el recall, asumiendo que la precisión se mantenía aceptable. Si varios

candidatos cumplían estos criterios, se seleccionó el modelo con el menor costo computacional como la versión óptima.

3. RESULTADOS Y DISCUSIÓN MATERIALES Y MÉTODOS

Esta sección presenta y discute los resultados obtenidos de la experimentación de 15 modelos de detección utilizando una estrategia de validación cruzada de diez pliegues. El rendimiento de detección de todos los modelos basado en las métricas calculadas se muestra en la Tabla 1.

3.1. Rendimiento en la Fase de Entrenamiento

La fase de entrenamiento implicó un procedimiento de validación cruzada de 10 pliegues, para evaluar el rendimiento de los modelos YOLO (v10, v11 y v12) en diferentes configuraciones. Los valores promedio de las métricas de evaluación (mAP@0.5, Precisión y Recall) para cada modelo y umbral de confianza se resumen en la Tabla 1. Esta tabla proporciona una visión general completa del rendimiento de los modelos durante el entrenamiento, permitiendo una comparación detallada de sus fortalezas y debilidades.

Model	Parameters	Threshold	mAP50 (SD)	Precision (SD)	Recall (SD)	Wilcoxon-Test
YOLOv10 medium	200 epochs AdamW batch 16 early stop 20	0.5	**0.9132 (0.007)**	**0.9131 (0.009)**	**0.8644 (0.013)**	**p=0.431**
		0.6	0.9032 (0.007)	0.9284 (0.007)	0.8403 (0.013)	p=0.013
		0.7	0.8849 (0.008)	0.9461 (0.006)	0.7979 (0.016)	p<0.05
		0.8	0.8407 (0.012)	0.9690 (0.005)	0.6997 (0.025)	p<0.05
		0.9	0.7441 (0.012)	0.9920 (0.003)	0.4941 (0.025)	p<0.05
YOLOv11 medium	200 epochs AdamW batch 16 early stop 20	0.5	0.9150 (0.008)*	0.9167 (0.007)	0.8679 (0.014)	-
		0.6	0.8971 (0.010)	0.9390 (0.008)	0.8256 (0.021)	p<0.05
		0.7	0.8608 (0.016)	0.9614 (0.007)	0.7444 (0.035)	p<0.05
		0.8	0.7847 (0.016)	0.9855 (0.003)	0.5801 (0.035)	p<0.05
		0.9	0.6687 (0.024)	0.9964 (0.002)	0.3411 (0.050)	p<0.05
YOLOv12 medium	200 epochs AdamW batch 16 early stop 20	0.5	0.9148 (0.007)	0.9213 (0.007)	0.8644 (0.013)	p=0.845
		0.6	0.8932 (0.011)	0.9446 (0.008)	0.8143 (0.025)	p<0.05
		0.7	0.8533 (0.018)	0.9674 (0.007)	0.7257 (0.039)	p<0.05
		0.8	0.7803 (0.018)	0.9866 (0.003)	0.5699 (0.036)	p<0.05
		0.9	0.6584 (0.031)	0.9975 (0.002)	0.3190 (0.062)	p<0.05

Tabla 1. Comparación de Rendimiento de YOLOv10, YOLOv11 y YOLOv12 a Través de Diferentes Umbrales de Confianza

La configuración que logró el mAP@0.5 más alto fue elegida como el modelo pivote para una comparación estadística adicional. Este fue YOLOv11 medium con una confianza de

0.5, alcanzando un mAP@0.5 de 0.915. Esta alta puntuación de mAP@0.5 indica que, en esta configuración, demostró sólidas capacidades de localización de objetos durante la fase de entrenamiento. La elección de mAP@0.5 como métrica principal refleja la importancia de una predicción precisa de los cuadros delimitadores para las tareas posteriores de seguimiento y estimación de densidad.

Utilizando la prueba de Wilcoxon, este modelo YOLOv11 se comparó estadísticamente con las configuraciones restantes para determinar si sus diferencias de rendimiento eran significativas. Los resultados indicaron que YOLOv12 (umbral de confianza 0.5) no fue significativamente diferente de YOLOv11, con un valor p de 0.841; YOLOv10 (umbral de confianza 0.5) tampoco mostró una diferencia estadísticamente significativa, con un valor p de 0.431; y finalmente, YOLOv10 (umbral de confianza 0.6) tampoco mostró una diferencia estadísticamente significativa. En este punto, según nuestra regla de selección, se descartaron las configuraciones con umbrales más altos, lo que significa que se excluyó YOLOv10 con un umbral de 0.6. Por lo tanto, todos los modelos con un umbral de 0.5 se consideraron estadísticamente equivalentes en rendimiento, lo que significa que sus diferencias en mAP@0.5 no fueron lo suficientemente grandes como para considerarse significativas. El hecho de que YOLOv12 lograra un mAP@0.5 comparable a YOLOv11 sugiere que sus mejoras arquitectónicas, como los mecanismos de atención y la agregación mejorada de características, no proporcionaron una ventaja estadísticamente significativa en términos de precisión de detección bruta en este conjunto de datos. De manera similar, la arquitectura optimizada de YOLOv10 también se desempeñó a la par de YOLOv11, lo que indica que las simplificaciones no afectaron negativamente sus capacidades de localización.

Con múltiples opciones estadísticamente equivalentes, el costo computacional se convirtió en el factor decisivo para seleccionar el modelo óptimo para la fase de prueba. Según

UltraLytics (2024), YOLOv10 con un umbral de 0.5 surgió como la opción óptima debido a su costo computacional significativamente menor. Si bien YOLOv12 incorpora módulos avanzados como Area Attention (A^2) y R-ELAN, que pueden ofrecer mejoras de precisión en ciertos contextos, es probable que estas mejoras aumenten la sobrecarga computacional debido a la complejidad añadida de los mecanismos de atención y la sofisticada agregación de características. Esta complejidad puede generar una mayor latencia y un mayor consumo de recursos. Por otro lado, la arquitectura de YOLOv10 simplifica la inferencia al introducir estrategias de submuestreo desacoplado y asignación dual, eliminando eficazmente NMS y reduciendo la redundancia computacional. Si bien aún pueden ser necesarias algunas optimizaciones menores para aprovechar al máximo sus ganancias de eficiencia, el diseño optimizado de YOLOv10 garantiza un menor consumo de recursos, lo que lo hace más adecuado para aplicaciones en tiempo real. En contraste, YOLOv11, aunque incorpora mecanismos de atención como C2PSA y *pooling* de contexto multiescala (SPPF), mantiene un diseño equilibrado pero sigue siendo menos eficiente que YOLOv10 en términos de costo computacional. La arquitectura de YOLOv10, con su *pipeline* de procesamiento más simple y eficiente, se posiciona como el claro ganador cuando se prioriza el costo computacional.

Además, el proceso de entrenamiento se monitoreó cuidadosamente para garantizar que los modelos no sobreajustaran los datos de entrenamiento. La Figura 2 muestra las curvas de pérdida de entrenamiento y validación para el modelo óptimo YOLOv10. La disminución constante tanto en la pérdida de entrenamiento como en la de validación, y la proximidad de las curvas de entrenamiento y validación en todo el modelo, indican que el modelo aprendió eficazmente sin memorizar los datos de entrenamiento. Esta observación respalda la fiabilidad del rendimiento del modelo y su potencial para una buena generalización en datos no vistos.

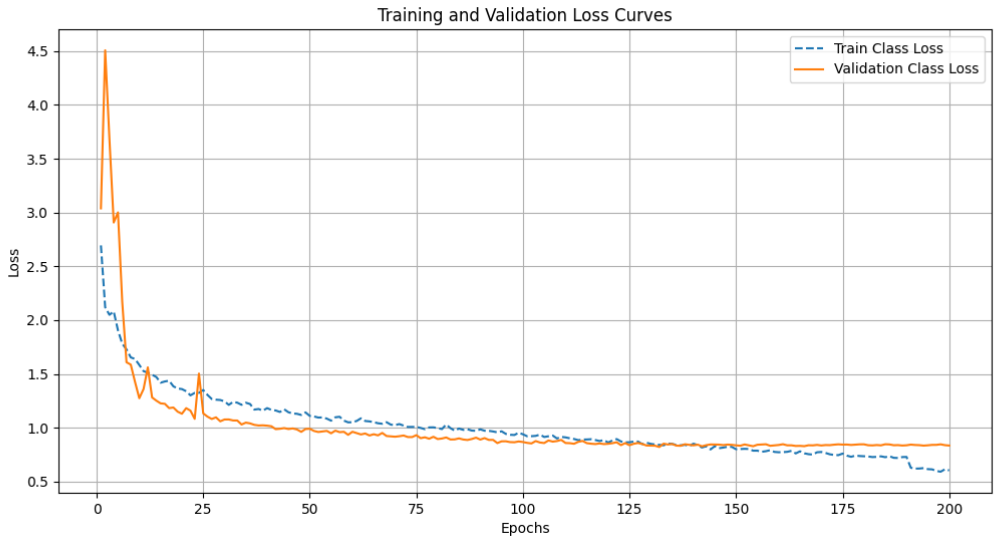

Figura 2. Curvas de Pérdida de Entrenamiento y Validación para YOLOv10.

3.2. Evaluación del rendimiento en la fase de prueba

El modelo YOLOv10 con un umbral de confianza de 0.5, seleccionado basándose en el análisis de la fase de entrenamiento, fue evaluado en un conjunto de prueba reservado para proporcionar una evaluación imparcial de su rendimiento de generalización. Esta evaluación es crucial para determinar la efectividad del modelo en la detección de individuos en datos de video no vistos, lo cual es representativo de escenarios de aplicación del mundo real. Los resultados muestran un rendimiento sólido, con un mAP@0.5 de **0.9267**, un mAP@0.5–0.95 de **0.7647**, una precisión de **0.9218** y un recall de **0.8844**.

Para evaluar aún más las capacidades del modelo, se realizó una evaluación cualitativa en escenarios de densidad de multitudes variables —que van de baja a alta— capturados desde diferentes perspectivas de cámara. Se muestran ejemplos representativos en la Figura 3.

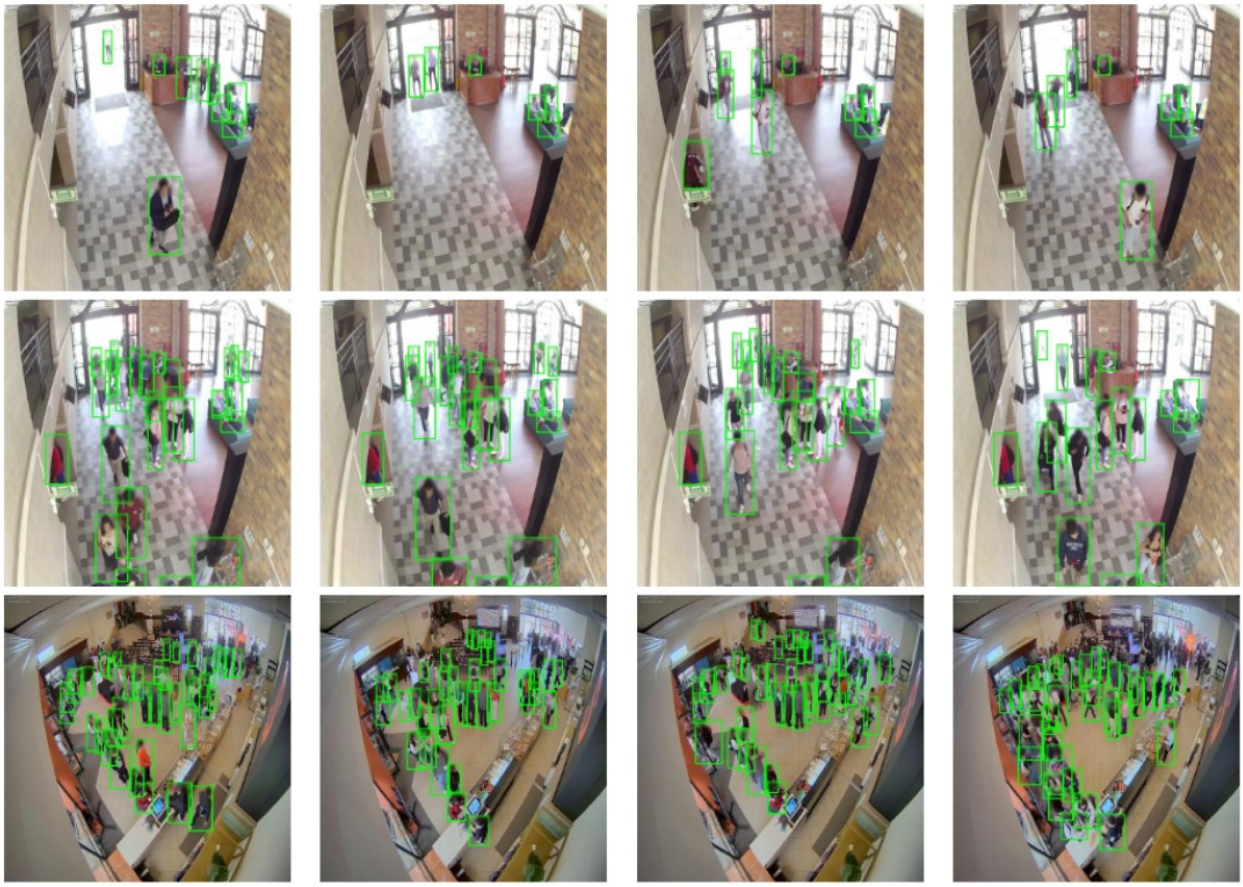

Figura 3. Ejemplos de diferentes escenarios de densidad: La fila superior muestra escenas de baja densidad, la fila central muestra escenas de alta densidad desde una cámara cercana, y la fila inferior muestra escenas de alta densidad desde una cámara distante.

3.3. Comparación basada en el estado del arte

Para contextualizar el rendimiento del sistema propuesto, podría ser interesante compararlo con los métodos existentes de vanguardia. En la introducción, destacamos los avances recientes en la detección de peatones, específicamente los modelos YOLO-ABD (Hua et al., 2024) y DP-YOLO (Jiao y Abdullah, 2024), que reportaron altas puntuaciones de mAP@0.5.

Hua et al. (2024) lograron un mAP@0.5 del 89.3 % utilizando YOLO-ABD, un método ligero que incorpora detección de objetos pequeños y *channel shuffling*, evaluado en el conjunto de datos IITB-Corridor. Jiao y Abdullah (2024) alcanzaron un mAP@0.5 del 89.7 % con DP-YOLO, que emplea convolución deformable y Varifocal Loss para mejorar la detección en escenas concurridas.

El modelo YOLOv10, seleccionado como el modelo óptimo en este estudio, logró un mAP@0.5 de 0.9267 en el conjunto de prueba. Este resultado demuestra que el componente de detección del sistema propuesto se desempeña favorablemente en comparación con el estado del arte reportado. La puntuación de mAP@0.5 más alta sugiere una precisión potencialmente mayor en la localización de individuos.

Sin embargo, es importante reconocer las diferencias en los conjuntos de datos y las metodologías de evaluación. Los modelos YOLO-ABD y DP-YOLO fueron evaluados en conjuntos de datos diferentes (IITB-Corridor y un conjunto de datos de escenas concurridas, respectivamente) al utilizado en este estudio, el cual fue creado específicamente para el entorno universitario. Por lo tanto, una comparación cuantitativa directa debe interpretarse con cautela.

A pesar de estas diferencias, la comparación proporciona una visión valiosa sobre la efectividad relativa del enfoque propuesto basado en YOLOv10. El sólido rendimiento de YOLOv10 indica su idoneidad para la detección precisa de personas dentro del contexto

específico de los videos de vigilancia universitaria y respalda su uso como base para el sistema de análisis de densidad de multitudes.

3.4. Prueba de Concepto: Aplicación a la Videovigilancia

Los métodos tradicionales, como el conteo manual y el uso de sensores físicos, enfrentan limitaciones significativas en cuanto a escalabilidad, adaptabilidad y precisión. Estos desafíos resaltan la necesidad de soluciones basadas en visión por computadora, particularmente aquellas impulsadas por el Aprendizaje Profundo, que ofrecen la capacidad de detectar y rastrear individuos en tiempo real, proporcionando una base para la toma de decisiones dinámica y automatizada.

Para demostrar la efectividad práctica del modelo YOLOv10 seleccionado, se desarrolló un sistema de prueba de concepto para el análisis automatizado de la saturación individual en grabaciones de video de vigilancia. Este sistema reemplaza el proceso de revisión manual con un *pipeline* completamente automatizado capaz de procesar grandes volúmenes de datos de video de manera eficiente y precisa. Su objetivo es extraer información detallada sobre los niveles de saturación de los individuos dentro de los espacios monitoreados, lo que permite tomar decisiones mejor informadas en áreas como el control de multitudes, la gestión de espacios públicos, la logística de eventos y la planificación de seguridad.

El *pipeline* se inicia escaneando y listando los archivos de video disponibles, desde los cuales los usuarios pueden seleccionar un video y definir una ventana de tiempo específica para el análisis. En su núcleo, el sistema emplea el modelo YOLOv10 ajustado para realizar la detección humana en tiempo real en cada fotograma del video. Se integra el seguimiento de múltiples objetos para preservar la consistencia de la identidad a través de los fotogramas, utilizando algoritmos como ByteTrack, BotSort y StrongSORT. La inclusión de StrongSORT

mejora significativamente la robustez del seguimiento mediante técnicas de reidentificación, que son especialmente valiosas en escenarios con oclusiones o individuos superpuestos.

Tras el procesamiento, el sistema produce un informe exhaustivo que contiene resúmenes visuales y cuantitativos de la dinámica de saturación durante el intervalo seleccionado. Esto incluye una salida de video etiquetado con cuadros delimitadores y etiquetas de identidad únicas para cada persona detectada, junto con una línea de tiempo de saturación y métricas clave como la presencia individual promedio, máxima y mediana por segundo. La Figura 4 ilustra un ejemplo de salida del sistema, evidenciando su capacidad para generar información procesable a partir de la entrada de video sin procesar.

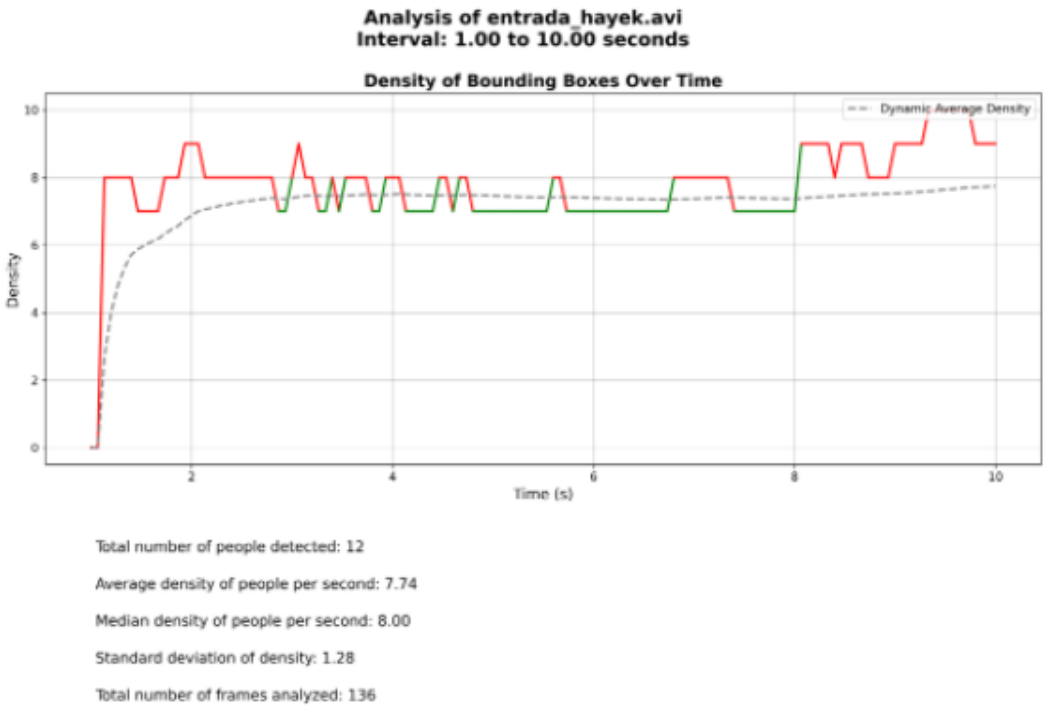

Figura 4. Análisis de Saturación

En este ejemplo particular, el sistema identificó 12 individuos únicos durante un intervalo de 10 segundos, analizando un total de 136 fotogramas. El número promedio de

individuos por segundo fue de 7.74, con una mediana de 8.00 y una baja desviación estándar de 1.28, lo que indica una alta consistencia temporal en la detección. El gráfico de análisis de saturación revela la progresión de la acumulación y dispersión individual dentro de la escena, mientras que las regiones resaltadas en rojo indican ráfagas cortas de aglomeración que podrían representar momentos de potencial congestión. Estos resultados resaltan el potencial del sistema para el monitoreo en tiempo real y su aplicabilidad en entornos prácticos que requieren sistemas de gestión espacial receptivos y sistemas de alerta temprana para la sobrepoblación o la saturación de multitudes.

4. CONCLUSIONES Y TRABAJO FUTURO

Esta investigación desarrolló y evaluó con éxito una metodología para el entrenamiento robusto y el análisis comparativo de modelos YOLO para la detección de personas en videos de vigilancia. El estudio demostró la eficacia de emplear una estrategia de validación cruzada de 10 pliegues, repetida tres veces, para asegurar una selección de modelo fiable y mitigar el sobreajuste. La evaluación comparativa de los modelos YOLOv10, v11 y v12 reveló que YOLOv10, con un umbral de confianza de 0.5, ofrecía un sólido equilibrio entre precisión y eficiencia computacional, siendo finalmente seleccionado como el modelo óptimo. Esta selección se realizó tras considerar la equivalencia estadística en mAP@0.5 con otros modelos y priorizar un menor costo computacional. Además, la implementación de la prueba de concepto de un sistema de análisis de saturación individual demostró el valor práctico del modelo YOLOv10 seleccionado para el monitoreo automatizado de multitudes, ofreciendo información procesable para la utilización del espacio y la planificación de seguridad. Aunque no se pueden realizar comparaciones directas debido a las diferencias en los conjuntos de datos y los contextos de evaluación, el rendimiento de YOLOv10 en este estudio proporciona un punto de referencia útil cuando se considera junto con modelos de detección de peatones de vanguardia como YOLO-ABD y DP-YOLO. Esto sugiere que el enfoque propuesto es efectivo para la implementación en el mundo real, particularmente en entornos dinámicos como los campus universitarios, y destaca su potencial para respaldar las decisiones operativas a través de un análisis de saturación preciso y escalable.

El trabajo futuro se centrará en varias áreas clave para mejorar las capacidades del sistema y abordar las limitaciones actuales. En primer lugar, los esfuerzos se dirigirán a mejorar la robustez del sistema en una gama más amplia de condiciones del mundo real, incluida la iluminación compleja, las oclusiones y las diferentes perspectivas de cámara.

Además, probar el sistema con secuencias de video más largas y heterogéneas en diferentes entornos ambientales proporcionaría una evaluación más completa de su rendimiento e identificaría áreas para un mayor refinamiento. Finalmente, si bien la prueba de concepto actual demuestra la viabilidad técnica del análisis automatizado de la saturación de individuos, el desarrollo de una aplicación completa para el usuario final sigue siendo un paso crucial. Esto implicaría la creación de una interfaz de usuario intuitiva, la integración del *backend* en una solución de software implementable y la garantía de la escalabilidad, el mantenimiento y la usabilidad para las partes interesadas del mundo real.

5. REFERENCIAS BIBLIOGRÁFICAS

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). *Simple online and realtime tracking*. En *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE. https://doi.org/10.1109/icip.2016.7533003

Cai, Y., Liu, J., Tang, J., & Wu, G. (2023). *Robust Object Modeling for Visual Tracking*. arXiv. https://arxiv.org/abs/2308.05140

Chen, S.-F., Chen, J.-C., Jhuo, I.-H., & Lin, Y.-Y. (2024). *Improving Visual Object Tracking through Visual Prompting*. arXiv. https://arxiv.org/abs/2409.18901

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., & Lu, H. (2021). *Transformer Tracking*. arXiv. https://arxiv.org/abs/2103.15436

Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., & Meng, H. (2023). *StrongSORT: Make DeepSORT Great Again*. arXiv. https://arxiv.org/abs/2202.13514

Fu, Z., Liu, Q., Fu, Z., & Wang, Y. (2021). *STMTrack: Template-free Visual Tracking with Space-time Memory Networks*. arXiv. https://arxiv.org/abs/2104.00324

Hua, C., Luo, K., Wu, Y., & Shi, R. (2024). YOLO-ABD: A Multi-Scale Detection Model for Pedestrian Anomaly Behavior Detection. *Symmetry*, *16*(8), 1003. https://doi.org/10.3390/sym16081003

Jiao, L., & Abdullah, M. I. (2024). *DP-YOLO: Enhancing Pedestrian Detection in Crowd Scenes with Deformable Convolution and Varifocal Loss*. En *Proceedings of the 2024 3rd International Conference on Cyber Security, Artificial Intelligence and Digital Economy* (págs. 226–231). Association for Computing Machinery. https://doi.org/10.1145/3672919.3672962

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment Anything*. arXiv. https://arxiv.org/abs/2304.02643

Lohani, D., Crispim-Junior, C., Barthélemy, Q., Bertrand, S., Robinault, L., & Tougne Rodet, L. (2022). Perimeter Intrusion Detection by Video Surveillance: A Survey. *Sensors*, *22*(9), 3601. https://doi.org/10.3390/s22093601

Qin, J., Wu, J., Yan, P., Li, M., Yuxi, R., Xiao, X., Wang, Y., Wang, R., Wen, S., Pan, X., & Wang, X. (2023). *FreeSeg: Unified, Universal and Open-Vocabulary Image Segmentation*. arXiv. https://arxiv.org/abs/2303.17225

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y.,

Girshick, R., Dollár, P., & Feichtenhofer, C. (2024). *SAM 2: Segment Anything in Images and Videos*. arXiv. https://arxiv.org/abs/2408.00714

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). *You Only Look Once: Unified, Real-Time Object Detection*. arXiv. https://arxiv.org/abs/1506.02640

Su, E. (2024). Visual Object Tracking Using Deep Learning Techniques: A Comparison. *Transactions on Computer Science and Intelligent Systems Research*, *7*, 619–626. https://doi.org/10.62051/7wag7n50

Ultralytics. (2024). *Ultralytics YOLO documentation*. Retrieved May 13, 2025, from https://docs.ultralytics.com/models/

Wojke, N., Bewley, A., & Paulus, D. (2017). *Simple Online and Realtime Tracking with a Deep Association Metric*. arXiv. https://arxiv.org/abs/1703.07402

Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., Krishnamoorthi, R., & Chandra, V. (2023). *EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything*. arXiv. https://arxiv.org/abs/2312.00863

Xu, T., Feng, Z.-H., Wu, X.-J., & Kittler, J. (2019). Learning Adaptive Discriminative Correlation Filters via Temporal Consistency Preserving Spatial Feature Selection for Robust Visual Object Tracking. *IEEE Transactions on Image Processing*, *28*(11), 5596–5609. https://doi.org/10.1109/tip.2019.2919201

Yan, B., Peng, H., Fu, J., Wang, D., & Lu, H. (2021). *Learning Spatio-Temporal Transformer for Visual Tracking*. arXiv. https://arxiv.org/abs/2103.17154

Yang, C.-Y., Huang, H.-W., Chai, W., Jiang, Z., & Hwang, J.-N. (2024). *SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory*. arXiv. https://arxiv.org/abs/2411.11922

Zhang, Y., Shen, Z., & Jiao, R. (2024). *Segment Anything Model for Medical Image Segmentation: Current Applications and Future Directions*. arXiv. https://arxiv.org/abs/2401.03495

Zhang, Z., & Peng, H. (2019). *Deeper and Wider Siamese Networks for Real-Time Visual Tracking*. arXiv. https://arxiv.org/abs/1901.01660