Uso de videoanalítica y modelos matemáticos para la detección y seguimiento de objetos para apoyar la gestión vial en la ciudad de Bogotá.

David Leonardo Moreno Bedoya

Universidad EAN

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

13 de junio de 2026

Uso de videoanalítica y modelos matemáticos para la detección y seguimiento de objetos para apoyar la gestión vial en la ciudad de Bogotá.

David Leonardo Moreno Bedoya

Trabajo de grado presentado como requisito para optar al título de:

Magíster en Ciencia de Datos

Director (a):

Miguel Ángel Zúñiga Gutierrez

Modalidad:

Misión Académica Internacional

Universidad EAN

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

13 de junio de 2026

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá, 13 de junio de 2026

"Vision is the art of seeing what is invisible to others."

— Jonathan Swift

.

Agradecimientos

Agradezco a la Universidad EAN y al programa de Maestría en Ciencia de Datos por la formación recibida. A mi director de trabajo de grado, Miguel Ángel Zúñiga Gutierrez, por su orientación y acompañamiento durante el desarrollo de esta investigación. A la Secretaría Distrital de Movilidad de Bogotá por el trabajo tan interesante que me ha sido encargado. A mi madre, Graciela Bedoya, a Alejandra María mi esposa y a mis hijos María José y Juan David por su apoyo incondicional.

RESUMEN

Bogotá enfrenta una alta accidentalidad vial, con fallecidos que aumentan progresivamente (508 en 2021, 665 en 2024), siendo el exceso de velocidad uno de los principales factores contribuyentes. Los sistemas actuales de videoanalítica de la Secretaría de Movilidad alcanzan una precisión cercana al 80 %, insuficiente para los modelos de calibración de congestión que requieren más del 85 %.

Este proyecto propone un sistema de seguimiento multi-objeto (MOT) que integra detección YOLO con un Filtro de Kalman de 8 dimensiones, velocidades estimadas mediante flujo óptico de Lucas-Kanade y descriptores de apariencia OSNet. La asociación óptima se resuelve mediante el algoritmo Húngaro, combinando IoU y similitud coseno.

La metodología cuantitativa empleó tres datasets: Waymo Open Dataset (990 frames, cámara móvil), MOT17 (5316 frames, cámara estática) y un dataset local de Bogotá (500 frames, cámara fija), con un estudio de ablación de 14 configuraciones.

Los resultados muestran que el tracker propuesto alcanza 85.26 % MOTA y el mejor IDF1 (87.64 %) en el dataset de Bogotá, superando marginalmente el umbral del 85 %, aunque con MOTA inferior a ByteTrack (89.15 %) y SORT (89.63 %). Su principal ventaja reside en la preservación de identidades (32 ID switches vs 103 de SORT), dimensión crítica para estimación de trayectorias. El estudio de ablación revela que el ajuste de hiperparámetros (`min_hits=1`: +9.14 % MOTA) tiene mayor impacto que los componentes algorítmicos (flujo óptico: +3.45 %, embeddings: +2.76 %).

Se concluye que el sistema ofrece evidencia preliminar de viabilidad para la Secretaría de Movilidad en el escenario evaluado, con configuraciones adaptables según las necesidades de procesamiento. La validación en condiciones más diversas es necesaria antes de un despliegue operativo.

Palabras clave: videoanalítica, seguimiento multi-objeto, Filtro de Kalman, flujo óptico, YOLO, seguridad vial, Bogotá.

ABSTRACT

Bogotá faces high road accident rates, with fatalities increasing progressively (508 in 2021, 665 in 2024), speeding being one of the main contributing factors. Current video analytics systems at the Mobility Secretariat achieve approximately 80 % accuracy, insufficient for congestion calibration models requiring above 85 %.

This project proposes a multi-object tracking (MOT) system integrating YOLO-based detection with an 8-dimensional Kalman filter, velocities estimated via Lucas-Kanade optical flow, and OSNet appearance descriptors. Optimal association is solved using the Hungarian algorithm, combining IoU and cosine similarity.

The quantitative methodology employed three datasets: Waymo Open Dataset (990 frames, mobile camera), MOT17 (5316 frames, static camera), and a local Bogotá dataset (500 frames, fixed camera), with an ablation study of 14 configurations.

Results show that the proposed tracker achieves 85.26 % MOTA and the best IDF1 (87.64 %) on the Bogotá dataset, marginally exceeding the 85 % threshold, though with lower MOTA than ByteTrack (89.15 %) and SORT (89.63 %). Its main advantage lies in identity preservation (32 ID switches vs. 103 for SORT), a critical dimension for trajectory estimation. The ablation study reveals that hyperparameter tuning (`min_hits=1`: +9.14 % MOTA) has greater impact than algorithmic components (optical flow: +3.45 %, embeddings: +2.76 %).

We conclude that the system provides preliminary evidence of viability for the Mobility Secretariat in the evaluated scenario, with configurations adaptable to processing requirements. Validation across more diverse conditions is necessary before operational deployment.

Keywords: video analytics, multi-object tracking, Kalman Filter, optical flow, YOLO, road safety, Bogotá.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1
INTRODUCCIÓN

1.1 Contexto y Antecedentes

La accidentalidad vial constituye un problema de salud pública a nivel mundial, con aproximadamente 1.19 millones de muertes anuales según la Organización Mundial de la Salud (World Health Organization, 2023). En el contexto colombiano, la ciudad de Bogotá presenta un número elevado de fallecidos en vía: 508 en el año 2021, 630 en 2022, 629 en 2023 y 665 en 2024 (Secretaría Distrital de Movilidad, 2023). Estas muertes se asocian a diversos comportamientos de los actores viales, siendo los más relevantes desobedecer las señales de tránsito (129 casos en 2024) y el exceso de velocidad (61 casos en 2024).

El proyecto Bloomberg Philanthropies Initiative for Global Road Safety Bloomberg Philanthropies, 2004 identifica cuatro comportamientos riesgosos asociados a lesiones graves o muertes: exceso de velocidad, conducción bajo efectos del alcohol, no uso de cinturones de seguridad y uso inapropiado de cascos en motocicleta. Un estudio de la Universidad de los Andes en conjunto con la Universidad Johns Hopkins identifica el exceso de velocidad como la conducta que más contribuye al aumento de la accidentalidad y mortalidad vial.

Como parte de los esfuerzos para abordar esta problemática, la Secretaría de Movilidad está implementando modelos de videoanalítica para el seguimiento de actores viales en las vías (Secretaría Distrital de Movilidad, 2023). Estos sistemas, utilizados inicialmente para el conteo de vehículos en jornadas como el día sin carro, permiten capturar, extraer y analizar información de imágenes y videos para identificar comportamientos de conductores en tiempo real mediante las cámaras instaladas en la ciudad. El seguimiento multi-objeto (MOT, *Multiple Object Tracking*) es un área activa de investigación en visión por computador que aborda precisamente este tipo de tareas (Luo et al., 2021; Zhang et al., 2024).

1.2 Planteamiento del Problema

El problema principal está enmarcado en la precisión limitada de los sistemas actuales basados en YOLO (Jocher et al., 2023; Redmon et al., 2016). Los algoritmos empleados para conteo vehicular y cálculo de velocidades alcanzan una precisión que no supera el 80 %, mientras que los

modelos de calibración de congestión y análisis de flujos vehiculares requieren de una precisión superior al 85 % para generar estimaciones confiables (Secretaría Distrital de Movilidad, 2023).

Adicionalmente, los modelos de detección utilizan pesos pre-entrenados con datos de Estados Unidos y Europa, cuyo parque automotor difiere significativamente del colombiano (Lin et al., 2014). Los sistemas actuales presentan deficiencias en la persistencia de identidades a través del tiempo, el seguimiento de trayectorias completas, la estimación robusta de velocidades y el manejo adecuado de las oclusiones (Ciaparrone et al., 2020; Luo et al., 2021).

1.3 Declaración de Contribución

La precisión limitada de los sistemas actuales ($\sim$80 %) impide una adecuada calibración de los modelos de congestión pues estos requieren precisiones mayores al 85 % para que puedan ser de utilidad. Esta brecha técnica representa una barrera significativa para implementar una gestión eficiente del tráfico en Bogotá.

El problema de fondo no es solo detectar vehículos, sino estimar sus distancias y velocidades reales a partir de las imágenes de video. Bajo el modelo de cámara *pinhole*, la distancia de un objeto a la cámara puede calcularse a partir de la altura de su detección en la imagen mediante la relación $d = fH/h$, donde f es la distancia focal, H la altura real del vehículo y h la altura en píxeles de la caja delimitadora. Sin embargo, las detecciones que producen los modelos CNN como YOLO no son perfectamente consistentes: la caja delimitadora de un mismo vehículo varía en tamaño y posición entre frames consecutivos debido al ruido inherente del detector. Estas oscilaciones, aunque pequeñas a nivel visual, se amplifican cuadráticamente al aplicar la transformación perspectiva —dado que $d \propto 1/h$, un error relativo $\delta h/h$ en la altura se traduce directamente en un error proporcional $\delta d/d$ en la distancia estimada—. Para la estimación de velocidad, que depende de la derivada temporal de la distancia, el problema es aún más severo: el ruido de las detecciones genera oscilaciones espurias en la velocidad estimada que pueden enmascarar el movimiento real del vehículo.

A partir de esta observación, y del conocimiento profundo del autor sobre las capacidades del filtro de Kalman como estimador óptimo de estados en presencia de ruido, se concibió de forma independiente un tracker cuyo objetivo primario es suavizar y estabilizar las detecciones del modelo CNN para habilitar una estimación confiable de distancias y velocidades mediante pro-

yección perspectiva. El sistema integra tres componentes complementarios: el Filtro de Kalman de 8 dimensiones, que estima y predice la posición y dimensiones de cada objeto filtrando el ruido frame-a-frame de las detecciones; el flujo óptico de Lucas-Kanade, que proporciona una estimación independiente de la velocidad a partir del cambio de intensidad de los píxeles entre frames consecutivos; y embeddings de apariencia extraídos con OSNet, que permiten la re-identificación de objetos tras oclusiones parciales. Una vez implementado el sistema, se identificó que enfoques similares —como DeepSORT (Wojke et al., 2017), que combina filtro de Kalman con embeddings de apariencia— ya existían en la literatura, lo que valida la intuición detrás del diseño propuesto. No obstante, el tracker difiere de DeepSORT en la parametrización del estado (8D con (a, h) en lugar de 7D con (s, r)), la integración de flujo óptico y la estrategia de fusión de información. La elección de (a, h) sobre (s, r) tiene una ventaja física concreta: bajo el modelo de cámara *pinhole*, la altura h es inversamente proporcional a la distancia ($h \propto 1/d$), una relación más lineal que la del área $s \propto 1/d^2$, lo que reduce el error de predicción del filtro de Kalman y produce estimaciones de distancia y velocidad más estables (Capítulo 2).

Como contribución adicional, se construyó un dataset de evaluación con video de tráfico capturado en la malla vial de Bogotá (500 frames, cámara fija), etiquetado mediante auto-anotación con BoTSORT+ReID y verificación manual. Este dataset constituye, hasta donde se tiene conocimiento, uno de los primeros benchmarks de tracking vehicular con datos de la malla vial colombiana.

El aporte esperado es un sistema que no solo supere el 85 % de precisión en el seguimiento, sino que fundamentalmente mejore la consistencia de las detecciones para habilitar estimaciones confiables de distancia y velocidad —la información que efectivamente necesitan los modelos de gestión de tráfico de la Secretaría de Movilidad.

El impacto institucional se centra en dos áreas clave de la Secretaría de Movilidad: por un lado, la Dirección de Inteligencia para la Movilidad se beneficiaría de una mejora en la calidad de datos para modelos predictivos y de gestión que apoyen la toma de decisiones; por otro, la Dirección de Seguridad Vial podría contar con la identificación automatizada de comportamientos de riesgo asociados a exceso de velocidad.

1.4 Objetivos

1.4.1 Objetivo General

Implementar una metodología basada en arquitecturas de seguimiento (trackers) como SORT (Bewley et al., 2016) y similares, que permita identificar objetos mejorando la caracterización de sus propiedades espaciales y dinámicas. Esta metodología integra herramientas como el flujo óptico, la detección de objetos mediante YOLO (Redmon et al., 2016) y el uso de redes neuronales convolucionales (CNN) que permite mejorar su precisión y optimizar el desempeño y eficiencia del sistema de detección.

1.4.2 Objetivos Específicos

1. Desarrollar un modelo de videoanalítica basado en el detector de objetos YOLO que permita identificar con precisión los actores viales que circulan por las vías de Bogotá (vehículos, peatones, ciclistas, camiones, patinetas, buses, camionetas y otros).

2. Calcular e integrar el flujo óptico asociado al movimiento de los objetos entre una imagen y otra al Filtro de Kalman para determinar la velocidad y la dirección del movimiento de los objetos detectados, mejorando la estimación de su trayectoria.

3. Utilizar redes neuronales convolucionales (CNN) para extraer embeddings de apariencia que permitan identificar de manera única cada objeto durante el seguimiento, incluso en condiciones complejas (e.g., iluminación variable, oclusión parcial).

1.5 Pregunta de Investigación

¿Es posible mejorar la precisión (actualmente cercana al 80 %) para la detección y el seguimiento de objetos, a través de la incorporación de una mayor cantidad de datos a los algoritmos actuales, al usar modelos como el flujo óptico y representaciones vectoriales de imágenes provenientes de modelos de redes neuronales convolucionales?

1.6 Hipótesis

Un sistema de seguimiento multi-objeto que integre un filtro de Kalman de 8 dimensiones con parametrización (a, h), velocidades estimadas mediante flujo óptico de Lucas-Kanade y

embeddings de apariencia OSNet, alcanzará una precisión (MOTA) superior al 85 % y producirá trayectorias más estables que los trackers de referencia (SORT, ByteTrack, OC-SORT), medido mediante IDF1 y la variabilidad de la altura estimada ($\sigma(\Delta h)$), en escenarios de tráfico vehicular urbano.

1.7 Estructura del Documento

Este artículo se organiza de la siguiente manera: el Capítulo 2 presenta el marco teórico con los fundamentos de los algoritmos empleados; el Capítulo 3 describe el diseño metodológico y los datasets utilizados; el Capítulo 4 presenta los resultados experimentales y la propuesta de solución; y el Capítulo 5 discute los hallazgos y presenta las conclusiones.

CAPÍTULO 2
MARCO TEÓRICO

El presente proyecto se enmarca en el desarrollo de un sistema de seguimiento de objetos múltiples (MOT, *Multiple Object Tracking*) para aplicaciones de movilidad urbana, combinando técnicas clásicas de procesamiento de señales con modelos modernos de aprendizaje profundo. El campo del MOT ha experimentado avances significativos en la última década, transitando desde enfoques basados exclusivamente en geometría hacia sistemas que integran información de apariencia y movimiento (Ciaparrone et al., 2020; Zhang et al., 2024). A continuación se describen los fundamentos teóricos de los componentes principales del sistema.

2.1 Filtro de Kalman para Seguimiento de Objetos

El filtro de Kalman es un algoritmo markoviano, recursivo, óptimo, con respecto al método de los mínimos cuadrados, que estima el estado de un sistema dinámico lineal a partir de mediciones con ruido (Kalman, 1960). En el sistema propuesto, cada objeto se modela mediante un **estado de 8 dimensiones**:

Cada objeto que detecta la red neuronal YOLO (detector), arroja, además de la clase que identifica el tipo del objeto, la posición de su detección dentro de cada imagen. Esta posición incluye un centroide (x, y) y usualmente, el ancho (w) y alto (h) de dicha detección. Es a esta posición a la que se le hace seguimiento y por eso se propone la definición del siguiente estado que nos va a ayudar a identificar y corregir su posición actual usando los datos encontrados por el detector, y a hacer una predicción de su posición indicando la ubicación en donde se podría encontrar en la imagen siguiente:

$$\mathbf{x} = \begin{bmatrix} x & y & a & h & v_x & v_y & v_a & v_h \end{bmatrix}^T \tag{2.1}$$

donde (x, y) son las coordenadas del centroide, $a = w/h$ es la relación de aspecto, y (v_x, v_y, v_a, v_h) son las velocidades asociadas. Nótese que también se involucran los datos de velocidad obtenidos en el flujo óptico para poder condicionar mejor el modelo.

El filtro opera en dos fases recursivas. La **fase de predicción** proyecta el estado al siguiente paso temporal:

$$\mathbf{x}_{k|k-1} = \mathbf{F}_k \mathbf{x}_{k-1|k-1}, \quad \mathbf{P}_{k|k-1} = \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{Q}_k \tag{2.2}$$

La **fase de actualización** incorpora observaciones mediante la ganancia de Kalman (la cual es obtenida de forma que hace óptima la aproximación de un sistema lineal, como el seguimiento de la posición de un objeto basado en su posición y velocidad):

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{H}^T (\mathbf{H} \mathbf{P}_{k|k-1} \mathbf{H}^T + \mathbf{R})^{-1} \tag{2.3}$$

$$\mathbf{x}_{k|k} = \mathbf{x}_{k|k-1} + \mathbf{K}_k (\mathbf{z}_k - \mathbf{H} \mathbf{x}_{k|k-1}) \tag{2.4}$$

donde $\mathbf{x}_{k|k-1}$ es el estado predicho en el paso k dado las observaciones hasta $k-1$, $\mathbf{F}_k$ es la matriz de transición de estado que modela la dinámica del sistema (velocidad constante), $\mathbf{P}_{k|k-1}$ es la covarianza del error de predicción, $\mathbf{Q}_k$ es la covarianza del ruido del proceso, $\mathbf{K}_k$ es la ganancia de Kalman, $\mathbf{H}$ es la matriz de observación que relaciona el estado con las mediciones, $\mathbf{R}$ es la covarianza del ruido de medición, y $\mathbf{z}_k$ es el vector de observación obtenido del detector en el paso k.

El filtro de Kalman lineal es óptimo bajo supuestos de linealidad y ruido gaussiano (Welch & Bishop, 2006a, 2006b), reduciendo el ruido en trayectorias y manejando oclusiones temporales mediante propagación de estados no observados.

2.2 Flujo Óptico de Lucas-Kanade

El flujo óptico estima el movimiento aparente de objetos entre dos imágenes consecutivas bajo el supuesto de **conservación de intensidad** (Lucas & Kanade, 1981). Existen dos enfoques principales: métodos dispersos como Lucas-Kanade, que calculan el flujo en puntos de interés, y métodos densos como el de Farnebäck (2003), que estiman el flujo para cada píxel. En este trabajo se emplea Lucas-Kanade por su menor costo computacional. El método asume flujo constante en una vecindad espacial, descrita en la ecuación:

$$I_x v_x + I_y v_y + I_t = 0 \tag{2.5}$$

donde I_x, I_y son los gradientes espaciales e I_t es la diferencia temporal entre fotogramas.

Para resolver el sistema subdeterminado, Lucas-Kanade impone consistencia local mediante mínimos cuadrados en una ventana $w \times w$ (típicamente 15×15 píxeles):

$$\begin{bmatrix} \sum_w I_x^2 & \sum_w I_x I_y \\ \sum_w I_x I_y & \sum_w I_y^2 \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix} = - \begin{bmatrix} \sum_w I_x I_t \\ \sum_w I_y I_t \end{bmatrix} \tag{2.6}$$

La velocidad estimada por flujo óptico se incorpora al estado mediante una actualización de Kalman dedicada. Se construye un filtro temporal de 2 dimensiones sobre los estados de velocidad (v_x, v_y), donde la medición es la velocidad del flujo óptico $\mathbf{v}_{\text{flow}}$:

$$\mathbf{v}_{k|k} = \mathbf{v}_{k|k-1} + \mathbf{K}_v(\mathbf{v}_{\text{flow}} - \mathbf{v}_{k|k-1}) \tag{2.7}$$

donde $\mathbf{K}_v = \mathbf{P}_v(\mathbf{P}_v + \mathbf{R}_v)^{-1}$ es la ganancia de Kalman para los estados de velocidad, $\mathbf{P}_v$ es la covarianza de velocidad extraída del filtro principal y $\mathbf{R}_v$ es la covarianza del ruido de medición del flujo óptico. De esta forma, los pesos de fusión no son constantes fijos, sino que se determinan de manera óptima según la incertidumbre relativa entre la predicción del Kalman y la medición del flujo óptico.

Ego-movimiento y compensación: En el contexto de conducción autónoma, el *vehículo ego* es aquel que porta la cámara y los sensores. Cuando este vehículo se desplaza, el flujo óptico captura tanto el movimiento real de los objetos como el movimiento aparente causado por el desplazamiento de la propia cámara (*ego-movimiento*). Si no se compensa, el ego-movimiento contamina la estimación de velocidad de los objetos rastreados. En la implementación propuesta, se estima el flujo global de la escena (mediana del flujo óptico de todos los píxeles) y se resta a cada medición individual, aislando así el movimiento propio de cada objeto. En cámaras estáticas (como las de tráfico en Bogotá), este problema no existe y el flujo óptico refleja directamente el movimiento de los objetos.

2.3 Redes Neuronales para Embeddings de Apariencia

Las redes neuronales convolucionales (CNN) son modelos de aprendizaje profundo que aplican filtros espaciales aprendidos sobre imágenes, extrayendo jerárquicamente características desde bordes y texturas hasta patrones semánticos de alto nivel. En el contexto de MOT, se emplean

para obtener representaciones vectoriales (*embeddings*) que codifican la apariencia visual de cada objeto detectado (Zhou et al., 2019):

$$\phi(I) = f_{\text{CNN}}(I) \in \mathbb{R}^d \tag{2.8}$$

donde d es la dimensionalidad del embedding (512 para OSNet). La similitud entre objetos se calcula mediante similitud coseno:

$$s(i,j) = \frac{\phi(I_i) \cdot \phi(I_j)}{\|\phi(I_i)\|\|\phi(I_j)\|} \tag{2.9}$$

El sistema utiliza **OSNet (Omni-Scale Network)** (Zhou et al., 2019) como modelo principal, seleccionado por su arquitectura ligera (2.2M parámetros) y su capacidad de capturar características a múltiples escalas, validada en benchmarks de re-identificación de personas (Zheng et al., 2015). OSNet supera a modelos de propósito general como ResNet18 (He et al., 2016) (11.7M parámetros), EfficientNet-B0 (Tan & Le, 2019) (5.3M parámetros) y MobileNetV2 (Sandler et al., 2018) (3.4M parámetros) en tareas de re-identificación, los cuales fueron también probados inicialmente.

2.4 Métrica Intersección sobre Unión (IoU)

La Intersección sobre Unión mide el grado de superposición espacial entre dos cajas delimitadoras A y B:

$$\text{IoU}(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{2.10}$$

donde $|A \cap B|$ es el área de intersección entre las dos cajas, y $|A|$, $|B|$ son sus áreas respectivas. Un valor de IoU cercano a 1 indica alta superposición espacial, mientras que IoU = 0 indica cajas disjuntas. Extensiones como GIoU (Rezatofighi et al., 2019) generalizan esta métrica para cajas no superpuestas, aunque en este trabajo se emplea la formulación clásica. En el sistema propuesto, se utiliza $1 - \text{IoU}$ como componente de la matriz de costos para la asociación entre detecciones y tracks predichos.

2.5 Algoritmo Húngaro para Asignación Óptima

En cada fotograma, el sistema debe asociar m detecciones nuevas con n tracks existentes. Este problema se formula como una asignación lineal con matriz de costos $\mathbf{C} \in \mathbb{R}^{m \times n}$ (Kuhn, 1955):

$$\text{mín} \sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} X_{ij} \quad \text{sujeto a:} \quad \sum_{j} X_{ij} \leq 1, \quad \sum_{i} X_{ij} \leq 1, \quad X_{ij} \in \{0, 1\} \tag{2.11}$$

donde C_{ij} es el costo de asignar la detección i al track j (combinación ponderada de $1 - \text{IoU}$ y distancia de apariencia), y $X_{ij} = 1$ indica que la detección i se asigna al track j. Las restricciones garantizan que cada detección se asigne a lo sumo a un track y viceversa. El algoritmo Húngaro (Munkres, 1957) resuelve este problema en tiempo polinomial $O(n^3)$, garantizando la asignación óptima global.

2.6 Algoritmos de Seguimiento Multi-Objeto (MOT)

El paradigma de seguimiento por detección (*tracking-by-detection*) domina el campo del MOT (Luo et al., 2021; Zhang et al., 2024). A continuación se describen los algoritmos de seguimiento, *trackers*, de referencia usados comúnmente y contra los cuales se compara la propuesta.

2.6.1 SORT: Simple Online Realtime Tracking

SORT (Bewley et al., 2016) combina el filtro de Kalman con el algoritmo Húngaro usando un vector de estado de 7 dimensiones. Es muy rápido ($\sim$150 FPS) pero carece de información de apariencia, presentando dificultades con oclusiones.

2.6.2 ByteTrack: Asociación en Cascada

ByteTrack (Zhang et al., 2022) mejora SORT mediante asociación en dos etapas: primero empareja detecciones cuya confianza supera un umbral alto (típicamente 0.5), y luego utiliza las detecciones de baja confianza (por encima de 0.1) para intentar asociar tracks que quedaron sin emparejar en la primera etapa. De esta forma recupera tracks usando detecciones que otros métodos descartarían.

2.6.3 OC-SORT: SORT Centrado en Observaciones

OC-SORT (Cao et al., 2023) introduce re-actualización centrada en observaciones (ORU) para corregir errores del Kalman filter después de oclusiones, y trayectorias virtuales para mejorar re-asociación. Añade consistencia de dirección de velocidad a la matriz de costos.

2.6.4 Tracker Propuesto: IoU + Embeddings de Apariencia

El tracker propuesto sigue el paradigma introducido por DeepSORT (Wojke et al., 2017), que combina filtro de Kalman, embeddings de apariencia y algoritmo Húngaro. Sin embargo, a diferencia de DeepSORT —que usa el mismo vector de estado 7D de SORT con parametrización (s, r) (área y aspect ratio, donde solo el área tiene velocidad)—, el tracker propuesto emplea una parametrización (a, h) (aspect ratio y altura) con velocidades para ambas componentes, resultando en un estado de 8 dimensiones (ver Tabla 2.1):

El tracker propuesto integra tres componentes fundamentales. Como modelo de movimiento emplea un filtro de Kalman de 8 dimensiones que, a diferencia de los 7 estados de SORT, incluye velocidades tanto para la relación de aspecto como para la altura. Para la re-identificación visual, el sistema utiliza un modelo de apariencia basado en OSNet que genera embeddings de 512 dimensiones para cada detección. Finalmente, la función de costo para la asignación combina ambas fuentes de información mediante una ponderación:

$$C(i, j) = w_{iou} \cdot (1 - \text{IoU}(i, j)) + w_{app} \cdot (1 - \text{sim\_cos}(\mathbf{e}_i, \mathbf{e}_j)) \tag{2.12}$$

donde los pesos $w_{iou} = 0{,}9$ y $w_{app} = 0{,}1$ fueron determinados empíricamente mediante el estudio de ablación descrito en el Capítulo 4. Esta distribución asimétrica refleja que, en seguimiento vehicular, la predicción espacial del filtro de Kalman es altamente informativa y la apariencia actúa como factor de desempate en casos de ambigüedad.

Tabla 2.1
Comparación de vectores de estado entre trackers

Tracker	Dim	Vector de Estado	Apariencia
SORT	7D	$[x, y, s, r, \dot{x}, \dot{y}, \dot{s}]$	No
DeepSORT	7D	$[x, y, s, r, \dot{x}, \dot{y}, \dot{s}]$	Sí (CNN)
ByteTrack	7D	$[x, y, s, r, \dot{x}, \dot{y}, \dot{s}]$	No
OC-SORT	7D	$[x, y, s, r, \dot{x}, \dot{y}, \dot{s}]$	No
Propuesto	8D	$[x, y, a, h, \dot{x}, \dot{y}, \dot{a}, \dot{h}]$	Sí (OSNet)

donde $s = w \cdot h$ es el área de la caja delimitadora, $r = w/h$ es la relación de aspecto y h es la altura. Las variables r y a representan la misma cantidad (w/h), pero se mantiene la notación diferenciada para respetar las convenciones de cada sistema original. Nótese que los trackers de referencia no estiman velocidad para r, mientras que el tracker propuesto incluye $\dot{a}$ y $\dot{h}$.

Justificación de la parametrización (a, h) y su relación con la estimación de distancia. La elección de (a, h) sobre (s, r) se fundamenta en dos argumentos complementarios: la compatibilidad con el supuesto de linealidad del filtro de Kalman y la conexión directa con la estimación de distancia y velocidad mediante proyección perspectiva.

Bajo el modelo de cámara *pinhole*, la altura proyectada de un objeto a distancia d es:

$$h = \frac{fH}{d} \quad \Longleftrightarrow \quad d = \frac{fH}{h} \tag{2.13}$$

donde f es la distancia focal y H la altura real del objeto. Esta relación establece una correspondencia directa entre la altura en píxeles h y la distancia física d: si se dispone de una estimación estable de h, se puede calcular la distancia (y, por derivación temporal, la velocidad) del vehículo. Sin embargo, las detecciones del modelo CNN introducen ruido en h; un error δh se propaga a la estimación de distancia como $\delta d/d = \delta h/h$, de modo que oscilaciones de pocos píxeles en la altura de la caja delimitadora generan fluctuaciones proporcionales en la distancia estimada. El filtro de Kalman, al modelar explícitamente h y su velocidad $\dot{h}$ como estados, filtra este ruido y produce estimaciones de h considerablemente más estables que las detecciones crudas del detector.

Para un objeto que se desplaza a velocidad constante ($d = d_0 + v_d t$), la derivada temporal

de h es:

$$\dot{h} = -\frac{fHv_d}{(d_0 + v_d t)^2} \tag{2.14}$$

que varía suavemente entre fotogramas consecutivos. En contraste, el área $s = w \cdot h \propto 1/d^2$ tiene derivada $\dot{s} \propto -2/d^3$, cuya tasa de cambio es más pronunciada y no lineal. Para intervalos cortos entre fotogramas (Δt pequeño), la aproximación lineal $h_{k+1} \approx h_k + \dot{h}\Delta t$ introduce menor error de predicción que $s_{k+1} \approx s_k + \dot{s}\Delta t$, ya que $\ddot{h}/\dot{h} < \ddot{s}/\dot{s}$. Adicionalmente, incluir $\dot{a}$ permite modelar cambios en la relación de aspecto causados por giros o cambios de perspectiva, que SORT asume constantes. En resumen, la parametrización (a, h) no solo mejora la predicción del filtro de Kalman, sino que produce estimaciones de altura más estables que se traducen directamente en cálculos de distancia y velocidad más confiables mediante la Ecuación 2.13.

CAPÍTULO 3
METODOLOGÍA

3.1 Tipo de Investigación

Siguiendo la clasificación de Hernández Sampieri y Mendoza Torres (2018), la presente investigación se clasifica como **aplicada y cuantitativa**, con un alcance **descriptivo-explicativo**. Es aplicada porque se orienta a resolver un problema práctico mediante el desarrollo de un sistema de videoanalítica para la detección y seguimiento de actores viales. El enfoque cuantitativo se evidencia en el tratamiento de datos numéricos: métricas de precisión (MOTA, MOTP, IDF1), conteos de objetos y estimación de velocidades. El alcance descriptivo-explicativo se manifiesta en la caracterización del rendimiento de los trackers (descripción) y en el análisis de las causas de las diferencias de rendimiento mediante el estudio de ablación (explicación).

Tabla 3.1

Clasificación metodológica del estudio según Hernández Sampieri y Mendoza Torres (2018)

Dimensión	**Caracterización**
Propósito	Aplicada: Desarrollo de solución concreta para la Secretaría de Movilidad
Alcance	Descriptivo-Explicativo: Caracteriza rendimiento y analiza causas
Enfoque	Cuantitativo: Métricas de precisión, conteos vehiculares
Inferencia	Deductiva: Parte de teorías establecidas (Kalman, Lucas-Kanade)
Temporalidad	Transversal con elementos longitudinales

3.1.1 Variables del Estudio

La Tabla 3.2 presenta las variables del estudio, clasificadas en independientes (factores manipulados en el experimento) y dependientes (métricas de rendimiento medidas).

Tabla 3.2
Variables del estudio

Tipo	**Variable**	**Descripción**
Independientes	Configuración del tracker	Tipo de tracker (SORT, ByteTrack, OC-SORT, DeepSORT, Propuesto)
	Flujo óptico	Activación/desactivación de la estimación de velocidad por Lucas-Kanade
	Embeddings de apariencia	Activación/desactivación de descriptores OSNet
	Hiperparámetros	`min_hits`, `iou_thresh`, `max_age`
Dependientes	MOTA	Precisión global del seguimiento (frame a frame)
	IDF1	Consistencia de identidades a lo largo del tiempo
	ID Switches	Número de cambios incorrectos de identidad
	FPS	Velocidad de procesamiento (frames por segundo)

3.1.2 Población y Muestra

En el contexto de la videoanalítica, la **población** está constituida por el universo de secuencias de video de tráfico urbano disponibles en benchmarks internacionales y en la infraestructura de cámaras de la Secretaría de Movilidad de Bogotá. La **muestra** se seleccionó de forma intencional (no probabilística) a partir de tres fuentes complementarias, buscando cubrir escenarios diversos en tipo de cámara, tipo de objeto y condiciones de detección (Hernández Sampieri & Mendoza Torres, 2018):

- **Waymo Open Dataset**: 5 segmentos de entrenamiento (990 frames), seleccionados del conjunto de 1150 segmentos disponibles. Criterio de selección: diversidad de densidad vehicular y condiciones de iluminación.
- **MOT17**: 7 secuencias de evaluación (5316 frames), que constituyen el conjunto completo de entrenamiento con ground truth público del benchmark.
- **Dataset de Bogotá**: 500 frames seleccionados de un video de 5 minutos (9000 frames) capturado en la intersección Av. Circunvalar por Calle 85. Criterio de selección: representatividad

de flujo vehicular típico en horario diurno.

La muestra total comprende 6806 frames con 12 806 trayectorias anotadas, evaluadas bajo 14 configuraciones experimentales del tracker propuesto y 4 trackers de referencia.

3.2 Conjuntos de Datos

Para validar el sistema de seguimiento propuesto y compararlo con algoritmos de referencia, se utilizan tres conjuntos de datos estándar en la comunidad de visión por computador.

3.2.1 Waymo Open Dataset

El Waymo Open Dataset (Sun et al., 2020) es un conjunto de datos de conducción autónoma que presenta desafíos únicos para el seguimiento multi-objeto. Comprende escenas urbanas con vehículos, peatones y ciclistas capturadas desde una cámara móvil montada en el vehículo ego. Para este estudio se seleccionaron 5 segmentos de entrenamiento (990 frames en total), utilizando las cajas de ground truth como detecciones (modo oracle) para aislar el rendimiento del tracker del rendimiento del detector.

Desafío principal: La cámara móvil introduce ego-movimiento que contamina las estimaciones de flujo óptico, requiriendo compensación o desactivación de esta característica.

3.2.2 MOT17 (Multiple Object Tracking 2017)

MOT17 (Milan et al., 2016) es el benchmark estándar para evaluación de seguimiento de peatones. Contiene exclusivamente secuencias de peatones capturadas con cámaras estáticas de vigilancia. Se evaluaron 7 secuencias (5316 frames) bajo dos modalidades de detección: oracle (cajas de ground truth) y pública (detector SDP con $\sim$65 % recall), lo que permite analizar el rendimiento del tracker tanto en condiciones ideales como con detecciones ruidosas.

3.2.3 Dataset de Tráfico de Bogotá

Para validar el sistema en condiciones operativas locales, se construyó un dataset utilizando video capturado en la malla vial de Bogotá:

El video fue capturado a 1920×1080 píxeles y 30 FPS (5 minutos, 9000 frames) desde una cámara fija, similar a las instaladas en la red de la Secretaría de Movilidad. El etiquetado se realizó

mediante auto-anotación con BoTSORT+ReID seguida de verificación manual, y se seleccionaron 500 frames representativos para la evaluación.

Tabla 3.3
Resumen de datasets utilizados

Dataset	Frames	Cámara	Objetos	Detecciones
Waymo	990	Móvil	Vehículos	Oracle (100 % recall)
MOT17	5316	Estática	Peatones	Oracle / SDP (65 %)
Bogotá	500	Fija	Vehículos	YOLOv8 (Jocher et al., 2023)

3.3 Instrumentos de Medición

El instrumento de medición consiste en un sistema automatizado que integra detección de objetos (YOLO), seguimiento (Filtro de Kalman), y métricas de evaluación estándar del área. Las métricas empleadas están validadas por el protocolo *CLEAR MOT* (*Classification of Events, Activities and Relationships for Multi-Object Tracking*) (Bernardin & Stiefelhagen, 2008), el cual establece un marco estandarizado para la evaluación de sistemas de seguimiento multi-objeto. El protocolo define cómo asociar las predicciones del tracker con el *ground truth* (anotaciones manuales de referencia) en cada fotograma, utilizando un umbral de IoU para determinar correspondencias correctas, y a partir de estas asociaciones calcula las métricas descritas a continuación.

3.3.1 Métricas de Seguimiento

La métrica principal es **MOTA (Multiple Object Tracking Accuracy)** (Bernardin & Stiefelhagen, 2008), definida como:

$$\text{MOTA} = 1 - \frac{FN + FP + IDSW}{GT}$$

donde FN es el número de falsos negativos (objetos reales no detectados), FP es el número de falsos positivos (detecciones sin objeto real correspondiente), $IDSW$ es el número de cambios incorrectos de identidad, y GT es el número total de objetos anotados en el ground truth. Un MOTA de 1 indica seguimiento perfecto. MOTA captura la precisión frame-a-frame del tracker, pero no refleja directamente la coherencia temporal de las identidades asignadas.

Para evaluar esta coherencia se emplea **IDF1 (Identity F1-Score)** (Ristani et al., 2016):

$$\text{IDF1} = \frac{2 \cdot \text{IDTP}}{2 \cdot \text{IDTP} + \text{IDFP} + \text{IDFN}}$$

donde $IDTP$ son los verdaderos positivos de identidad (detecciones correctamente asignadas a su identidad real), $IDFP$ son los falsos positivos de identidad e $IDFN$ son los falsos negativos de identidad. IDF1 resulta particularmente relevante para aplicaciones de tráfico donde la continuidad de las trayectorias determina la utilidad del sistema —por ejemplo, en estimación de tiempos de viaje o matrices origen-destino, un tracker con alto MOTA pero bajo IDF1 produce trayectorias fragmentadas que invalidan el análisis posterior.

De forma complementaria, **MOTP (Multiple Object Tracking Precision)** mide el promedio de IoU entre detecciones emparejadas y ground truth, reflejando la calidad de localización espacial (Padilla et al., 2020). Por último, el conteo de **ID switches** cuantifica directamente los cambios incorrectos de identidad, siendo un indicador crítico de la robustez de la asociación. Métricas más recientes como HOTA (Luiten et al., 2021) buscan unificar la evaluación de detección y asociación en una sola métrica; no obstante, en este trabajo se emplean MOTA e IDF1 por ser las métricas estándar del benchmark MOTChallenge (Dendorfer et al., 2021) y las más ampliamente reportadas en la literatura.

3.3.2 Validación del Instrumento

Las métricas MOTA e IDF1 son estándares reconocidos internacionalmente, utilizados en los benchmarks MOT Challenge. Su validez está respaldada por la comunidad científica de visión por computadora (Bernardin & Stiefelhagen, 2008; Ristani et al., 2016).

3.4 Arquitectura del Sistema Propuesto

La Figura 3.1 presenta la arquitectura general del sistema de seguimiento multi-objeto propuesto. El pipeline procesa cada fotograma en cinco etapas: (1) el detector YOLO identifica objetos en la imagen, (2) el filtro de Kalman predice las posiciones de los tracks existentes, (3) el flujo óptico de Lucas-Kanade estima velocidades a partir del movimiento aparente entre fotogramas consecutivos, (4) los embeddings CNN (OSNet) extraen descriptores de apariencia, y (5) la asignación Húngara asocia detecciones con tracks usando una función de costo que combina IoU y similitud

de apariencia.

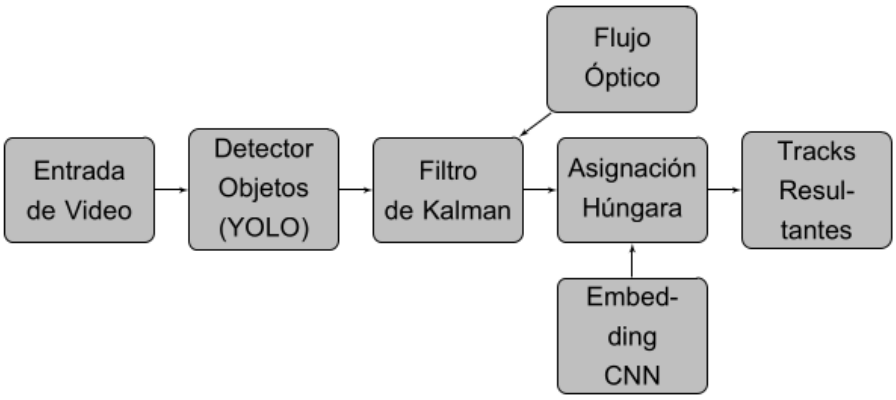

Figura 3.1
Arquitectura general del sistema de seguimiento multi-objeto propuesto

3.5 Configuraciones Experimentales

Se implementaron 14 configuraciones del tracker propuesto para el estudio de ablación, evaluando el impacto de cada componente e hiperparámetro. La Tabla 3.4 resume las 5 configuraciones principales. Cada fila activa o desactiva un componente respecto al baseline, permitiendo aislar su contribución individual:

Los componentes evaluados incluyen el **flujo óptico**, que estima velocidades mediante Lucas-Kanade y las fusiona con el filtro de Kalman, y los **embeddings de apariencia**, descriptores OSNet de 512 dimensiones utilizados para calcular similitud coseno entre detecciones y tracks. En cuanto a los hiperparámetros, `min_hits` establece el número mínimo de detecciones consecutivas antes de confirmar un track: con el valor por defecto (3), un objeto nuevo solo se incluye en la salida tras haber sido asociado exitosamente durante 3 frames seguidos, filtrando detecciones espurias (reflejos, sombras, artefactos), mientras que un valor de 1 reporta el track desde su primera aparición, mejorando el recall a costa de incrementar los falsos positivos. Por su parte, `iou_thresh` define el umbral mínimo de IoU para aceptar una asociación entre detección y track; un valor bajo (0.2) permite asociaciones con menor superposición, útil cuando los objetos se mueven rápido entre frames.

Tabla 3.4
Configuraciones principales del estudio de ablación

Configuración	Flujo Óptico	Embeddings	min_hits	iou_thresh
Baseline (solo IoU)	No	No	3	0.35
+ Flujo óptico	Sí	No	3	0.35
+ Embeddings	No	Sí	3	0.35
Completa	Sí	Sí	3	0.35
Óptima	Sí	Sí	1	0.20

3.6 Técnicas de Análisis

3.6.1 Estudio de Ablación

Un estudio de ablación consiste en remover o desactivar componentes individuales de un sistema para medir su contribución al rendimiento global. El término, tomado de la medicina (donde ablación significa la extirpación de tejido), es análogo a un **estudio de sensibilidad**: en ambos casos se varía un factor a la vez mientras los demás permanecen constantes, para aislar el efecto de cada uno. En este trabajo, el estudio de ablación evalúa sistemáticamente la contribución de cada componente (flujo óptico, embeddings de apariencia) y cada hiperparámetro (`min_hits`, `iou_thresh`, `max_age`) al rendimiento final del sistema.

3.6.2 Evaluación Comparativa

Se comparan cuatro trackers del estado del arte (SORT, ByteTrack, OC-SORT, DeepSORT) contra el tracker propuesto bajo condiciones controladas en los tres datasets.

3.6.3 Análisis del Techo de Detección

Se analiza la calidad de las detecciones para determinar el techo teórico de MOTA alcanzable, separando las pérdidas atribuibles al detector de las atribuibles al algoritmo de tracking.

3.7 Plataforma de Implementación

Los experimentos se ejecutaron sobre una estación de trabajo equipada con GPU NVIDIA compatible con CUDA, 16 GB de RAM y almacenamiento en SSD. El stack de software se com-

pone de Python 3.9+, PyTorch 2.0+ para la inferencia de modelos neuronales (YOLO, OSNet), OpenCV 4.8+ para el cálculo de flujo óptico, FilterPy 1.4+ para la implementación del filtro de Kalman y SciPy para la resolución del algoritmo Húngaro.

3.8 Conexión Metodología-Objetivos

El diseño metodológico permite evaluar sistemáticamente cada objetivo específico del proyecto. La evaluación en el dataset de Bogotá con detecciones YOLOv8 valida la precisión del detector en condiciones locales (OE1), mientras que el estudio de ablación con 14 configuraciones cuantifica el impacto del flujo óptico en la estimación de velocidad (OE2). La comparación de cuatro modelos de apariencia —OSNet, ResNet18, EfficientNet y MobileNet— identifica la arquitectura más adecuada para re-identificación vehicular (OE3). La evaluación multi-dataset (Waymo, MOT17, Bogotá) busca evaluar la generalización de los hallazgos, aunque el alcance limitado del dataset local (una intersección, condiciones diurnas) restringe las conclusiones que pueden extraerse sobre el desempeño en el conjunto de la malla vial de Bogotá.

CAPÍTULO 4
RESULTADOS

Este capítulo presenta los resultados experimentales organizados según la estructura metodológica: primero la presentación objetiva de los datos, seguida del análisis e interpretación, y finalmente la propuesta de solución derivada de los hallazgos.

4.1 Presentación de Resultados

4.1.1 Resultados en Waymo Open Dataset

Se evaluaron cuatro trackers en 5 secuencias del Waymo Open Dataset (990 frames, cámara móvil, vehículos). DeepSORT utiliza el mismo modelo OSNet que el tracker propuesto para una comparación justa.

Tabla 4.1
Resultados en Waymo Open Dataset (5 secuencias, 990 frames)

Tracker	MOTA	IDF1	ID Switches	FPS
ByteTrack	**88.48 %**	79.67 %	876	78.2
DeepSORT	82.99 %	75.31 %	919	2.4
OC-SORT	81.37 %	**81.43 %**	**407**	57.6
Propuesto (config. estándar)	80.95 %	73.50 %	817	1.9
Propuesto (min_hits=1)	**87.78 %**	65.88 %	929	2.5

4.1.2 Estudio de Ablación: Waymo

Tabla 4.2
Estudio de ablación de componentes (Waymo, 3 secuencias)

Configuración	MOTA	IDF1	ID Sw	ΔMOTA
Componentes principales				
Baseline (solo IoU)	78.64 %	65.04 %	534	–
+ Flujo óptico (OF)	82.09 %	70.08 %	439	+3.45 %
+ Embeddings OSNet	81.40 %	68.41 %	432	+2.76 %
+ OF + Embeddings	82.29 %	69.08 %	428	+3.65 %
Hiperparámetros				
iou_thresh=0.2	85.32 %	72.73 %	384	+6.68 %
iou_thresh=0.5	67.92 %	59.77 %	582	-10.72 %
min_hits=1	**87.78 %**	65.88 %	929	**+9.14 %**
min_hits=5	76.26 %	69.20 %	295	-2.38 %

4.1.3 Estudio de Ablación: MOT17

Tabla 4.3
Estudio de ablación de componentes (MOT17, 3 secuencias)

Configuración	MOTA	IDF1	ID Sw	ΔMOTA
Baseline (solo IoU)	56.61 %	61.23 %	607	–
+ Flujo óptico (OF)	58.14 %	63.17 %	464	+1.53 %
+ Embeddings OSNet	57.95 %	63.32 %	474	+1.34 %
+ OF + Embeddings	58.20 %	63.29 %	459	+1.59 %
min_hits=1	**59.66 %**	63.48 %	672	**+3.05 %**
min_hits=5	56.31 %	62.83 %	**355**	-0.30 %

La tendencia observada en Waymo se replica en MOT17: `min_hits=5` reduce los ID switches un 47 % respecto al baseline (355 vs 672 con `min_hits=1`), con una pérdida marginal

de MOTA (-0.30pp). Esto confirma que la relación entre persistencia de confirmación y estabilidad de identidades es consistente entre dominios (vehículos y peatones) y no un artefacto del dataset.

4.1.4 Comparación de Modelos de Apariencia

Tabla 4.4
Comparación de modelos de apariencia (Waymo, pesos fijos IoU=0.9 / Apariencia=0.1)

Modelo	Dim.	MOTA	IDF1	ID Sw	FPS	Parámetros
OSNet-x1.0	512	**75.13 %**	**55.3 %**	**153**	2.08	2.2M
MobileNetV3	960	70.29 %	48.4 %	159	1.89	5.4M
EfficientNet-B0	1280	69.77 %	51.1 %	162	1.82	5.3M
ResNet18	512	68.83 %	45.9 %	170	2.06	11.7M

4.1.5 Resultados en MOT17

Tabla 4.5
Resultados en MOT17 (7 secuencias, 5316 frames, todos los trackers con `min_hits=3`*)*

Tracker	MOTA	IDF1	ID Switches	FPS
DeepSORT	**64.41 %**	63.07 %	**540**	4.2
ByteTrack	64.33 %	**67.66 %**	742	220.6
Propuesto	63.19 %	64.92 %	1151	2.8
OC-SORT	58.94 %	62.64 %	685	169.1

4.1.6 Resultados en Dataset de Bogotá

Tabla 4.6
Resultados en dataset de tráfico de Bogotá (500 frames, cámara fija)

Tracker	MOTA	IDF1	Precisión	ID Sw.	FPS
SORT	**89.63 %**	84.91 %	**93.53 %**	103	301.8
OC-SORT	89.50 %	84.97 %	93.48 %	106	62.9
ByteTrack	89.15 %	87.28 %	90.46 %	**28**	78.9
Propuesto (IoU + App)	85.29 %	87.26 %	87.68 %	33	1.2
Propuesto (Full)	85.26 %	**87.64 %**	87.65 %	32	0.7
Propuesto (solo IoU)	85.23 %	87.48 %	87.72 %	37	53.0
DeepSORT	78.99 %	80.55 %	83.27 %	55	1.9

Los resultados cuantitativos se presentan en la Tabla 4.6. La Figura 4.1 presenta una comparación visual del seguimiento en una intersección de Bogotá (Av. Circunvalar por Calle 85). Cada panel muestra los bounding boxes producidos por un tracker diferente sobre el mismo frame: Ground Truth (verde), SORT (azul), ByteTrack (rojo) y el tracker propuesto (cian). Se observa que SORT y ByteTrack generan cajas adicionales (falsos positivos) en zonas con oclusión parcial, mientras que el tracker propuesto mantiene correspondencia más cercana al ground truth en la asignación de identidades.

Figura 4.1

Comparación visual de trackers en el dataset de Bogotá (frame 250). Paneles: Ground Truth (verde, superior izq.), SORT (azul, superior der.), ByteTrack (rojo, inferior izq.), Propuesto IoU+App (cian, inferior der.). Los números sobre cada caja indican el ID asignado.

4.2 Análisis e Interpretación

4.2.1 Interpretación de Resultados en Waymo

En el dataset Waymo, ByteTrack obtiene el mejor MOTA (88.48 %), superando al tracker propuesto en configuración estándar (80.95 %) por 7.53 puntos porcentuales. Al ajustar el hiperparámetro `min_hits=1`, el tracker propuesto alcanza 87.78 % MOTA, reduciendo la brecha a 0.7 puntos porcentuales. Es importante contextualizar esta diferencia: si bien 0.7pp puede parecer pequeña en términos absolutos, esta reducción se logra a un costo significativo en consistencia de identidades —el IDF1 cae de 73.50 % a 65.88 % (una pérdida de 7.62pp) y los ID switches aumentan de 817 a 929—. Es decir, el tracker sacrifica la coherencia temporal de las trayectorias para ganar cobertura frame-a-frame. Este compromiso no es menor: para aplicaciones de estimación de velocidad o tiempos de viaje, donde cada cambio de identidad corrompe una trayectoria completa, la configuración con `min_hits=1` resultaría contraproducente pese a su MOTA superior. En consecuencia, la comparación más justa con ByteTrack es la configuración estándar del tracker pro-

puesto, donde la brecha real es de 7.53pp a favor de ByteTrack en MOTA, aunque con un balance diferente en ID switches.

La ventaja de ByteTrack radica en su asociación en cascada, que recupera detecciones de baja confianza descartadas por otros métodos. Este mecanismo resulta particularmente efectivo para seguimiento vehicular con cámara móvil, donde el movimiento del vehículo ego genera variaciones en la posición y escala de las detecciones. Resulta llamativo que los trackers con información de apariencia (DeepSORT, Propuesto) no superen a los basados exclusivamente en IoU en este escenario, lo cual sugiere que el ego-movimiento degrada la calidad de los embeddings visuales al introducir variaciones de perspectiva y escala entre frames consecutivos.

Por su parte, OC-SORT logra la mejor preservación de identidades con solo 407 ID switches, menos de la mitad que ByteTrack (876). Este resultado confirma la efectividad de su mecanismo de re-actualización centrado en observaciones (ORU), que corrige la deriva del filtro de Kalman tras períodos de oclusión. Para aplicaciones donde la continuidad de tracks es prioritaria sobre la cobertura instantánea, OC-SORT ofrece el mejor perfil en este dataset.

4.2.2 Interpretación del Estudio de Ablación

El estudio de ablación revela un hallazgo que merece reflexión cuidadosa: los hiperparámetros tienen mayor impacto cuantitativo que los componentes algorítmicos. El ajuste de `min_hits=1` constituye la mayor contribución individual (+9.14 % MOTA), seguido de `iou_thresh=0.2` (+6.68 % MOTA), mientras que los componentes algorítmicos aportan mejoras más modestas: flujo óptico +3.45 % y embeddings +2.76 %. Sin embargo, esta comparación directa requiere matices importantes.

Advertencia sobre `min_hits=1` y la relevancia de IDF1. El efecto de `min_hits=1` no refleja una mejora en la calidad del seguimiento, sino un cambio en el criterio de confirmación de tracks: el tracker reporta cada objeto desde su primera detección, sin esperar frames adicionales de verificación. Esto infla el MOTA al reducir falsos negativos, pero a un costo severo en la consistencia de identidades: el IDF1 cae de 73.50 % a 65.88 % (una pérdida de 7.62pp) y los ID switches se disparan de 817 a 929. Con `min_hits=1`, cada fragmentación de track genera inmediatamente un nuevo identificador confirmado, y el objeto que reaparece recibe una identidad distinta sin penalización en MOTA —pero con penalización completa en IDF1, que mide precisamente la co-

rrespondencia biunívoca entre identidades predichas y reales.

Este fenómeno tiene consecuencias directas en sistemas de producción. En aplicaciones de conteo vehicular como las desplegadas por la Secretaría Distrital de Movilidad, cada objeto se cuenta *una sola vez* al ingresar a una zona de interés (polígono). Si el tracker fragmenta un track antes de la zona de conteo, el mismo vehículo físico ingresa con un nuevo identificador y se contabiliza dos veces, inflando los conteos. Inversamente, si el tracker pierde un objeto antes del polígono, este nunca se cuenta. En ambos casos, el MOTA —que evalúa la cobertura frame a frame— no captura el problema porque la duplicación o pérdida ocurre entre frames, no dentro de uno. **El IDF1, al evaluar la consistencia de identidades a lo largo del tiempo, se convierte en la métrica más relevante para aplicaciones de conteo y estimación de trayectorias**, ya que un IDF1 alto garantiza que cada objeto físico mantiene una identidad única durante su tránsito por la escena.

Resulta revelador que la configuración opuesta, `min_hits=5`, produce el resultado más deseable para aplicaciones de aforo vehicular: solo 295 ID switches (la menor fragmentación de todo el estudio) y el mejor IDF1 (69.20 %), a costa de una caída modesta de 2.38pp en MOTA (76.26 %). Esta pérdida de MOTA corresponde a objetos que aparecen durante menos de 5 frames —típicamente vehículos en el borde del campo visual que nunca habrían alcanzado un polígono de conteo—. Es decir, `min_hits=5` descarta precisamente las detecciones efímeras que inflarían el conteo sin aportar información útil, mientras que preserva las trayectorias estables necesarias para un aforo confiable. La relación inversa entre `min_hits` y los ID switches ($929 \rightarrow 534 \rightarrow 295$ para valores 1, 3 y 5 respectivamente) confirma que exigir mayor persistencia temporal antes de confirmar un track actúa como filtro natural contra la fragmentación.

En contraste, las contribuciones del flujo óptico y los embeddings mejoran la calidad intrínseca de la asociación sin degradar otras métricas. Por tanto, afirmar que "los hiperparámetros importan más que los componentes" sería una simplificación: `min_hits` modifica el criterio de reporte mientras que los componentes algorítmicos mejoran la calidad del seguimiento mismo. La diferencia es análoga a bajar el umbral de aprobación de un examen frente a mejorar la preparación del estudiante: el primero infla la tasa de aprobación sin mejorar el aprendizaje. **Para sistemas de aforo en vía, la recomendación derivada de este análisis es clara: priorizar configuraciones con bajo ID switches y alto IDF1 (`min_hits` $\geq$ 3), aun si esto implica un MOTA inferior**, ya

que la precisión del conteo depende de la unicidad de las identidades, no de la cobertura frame a frame.

La contribución del flujo óptico con compensación de ego-movimiento es un resultado que contradice recomendaciones previas de desactivarlo en cámaras móviles. La implementación con compensación basada en la mediana del flujo global demuestra mejoras en ambos escenarios: +3.45 % MOTA en Waymo (cámara móvil) y +1.53 % en MOT17 (cámara estática). Que la mejora sea mayor en cámara móvil resulta coherente con la teoría: el ego-movimiento introduce incertidumbre adicional en las predicciones del filtro de Kalman, y el flujo óptico —una vez compensado— proporciona información de velocidad complementaria que reduce esa incertidumbre. No obstante, la magnitud modesta de estas mejoras (entre 1 y 4 puntos porcentuales) indica que el flujo óptico no es un componente transformador, sino un complemento útil cuya contribución debe ponderarse frente a su costo computacional.

Un resultado particularmente revelador es la no-aditividad de los componentes: la combinación de flujo óptico y embeddings (+3.65 % MOTA) apenas supera al flujo óptico solo (+3.45 %), con una diferencia de 0.2 puntos porcentuales que cae dentro del margen de variabilidad experimental. Esto sugiere que ambos componentes capturan información parcialmente redundante en la etapa de asociación: tanto la predicción de velocidad como la similitud visual contribuyen a resolver las mismas ambigüedades de asignación, y una vez que una de las dos fuentes resuelve un caso difícil, la otra no aporta información adicional. Este hallazgo tiene implicaciones prácticas: en escenarios con restricciones de cómputo, activar solo el flujo óptico ofrece casi el mismo beneficio que la combinación completa, con menor carga computacional.

4.2.3 Interpretación de Resultados en MOT17

Los resultados en MOT17 exponen tanto las fortalezas como las limitaciones del tracker propuesto en un dominio diferente al de su diseño original. DeepSORT obtiene el mejor MOTA (64.41 %), seguido de cerca por ByteTrack (64.33 %) y el tracker propuesto (63.19 %). Dado que el techo de MOTA está limitado por el recall del detector público SDP (*Static Detection Proposals*, $\sim$65 %), todos los trackers operan cerca del límite teórico y las diferencias de 1–2pp entre ellos deben interpretarse con cautela.

Sin embargo, la métrica más reveladora en MOT17 son los ID switches. El tracker pro-

puesto registra 1151 cambios de identidad, más del doble que DeepSORT (540) y 55 % más que ByteTrack (742). Este resultado constituye la principal debilidad identificada del sistema y merece una reflexión detenida. La causa probable es que el tracker fue diseñado y calibrado para seguimiento vehicular —objetos de mayor tamaño, menor densidad, movimiento más predecible— y sus hiperparámetros no se adaptan bien al seguimiento de peatones, donde las oclusiones son más frecuentes, los objetos son más pequeños y las trayectorias presentan cambios de dirección más abruptos. En particular, la estrategia de asociación que combina IoU (90 %) y apariencia (10 %) puede ser subóptima para peatones, donde la apariencia debería tener mayor peso dada la menor discriminabilidad de las cajas espaciales en escenas densas. La estrategia de cascada de DeepSORT, que prioriza la re-identificación por apariencia antes de recurrir a IoU, demuestra ser más robusta en este dominio.

4.2.4 Interpretación de Resultados en Bogotá

Los resultados en el dataset de Bogotá son los más relevantes para el contexto institucional del proyecto, aunque deben interpretarse considerando las limitaciones del dataset (una sola intersección, 500 frames, condiciones diurnas favorables).

SORT obtiene el mejor MOTA (89.63 %), seguido de OC-SORT (89.50 %) y ByteTrack (89.15 %), todos ellos superando al tracker propuesto (85.26 %) por 4–4.4 puntos porcentuales. Esta diferencia no es despreciable y merece una explicación honesta: los trackers basados exclusivamente en IoU se benefician de la escena favorable (cámara fija, vehículos grandes, densidad moderada) donde la predicción espacial del filtro de Kalman es suficiente para mantener asociaciones correctas. El tracker propuesto, al incorporar embeddings de apariencia, introduce un componente adicional de costo en la asociación que, en este escenario de baja ambigüedad, puede generar desasociaciones que los trackers más simples evitan. Esto se refleja en la menor precisión (87.65 % vs 93.53 % de SORT), indicando que el sistema genera más falsos positivos.

No obstante, cuando los resultados se interpretan desde la perspectiva de la aplicación objetivo —el conteo vehicular mediante polígonos de aforo como los desplegados por la Secretaría Distrital de Movilidad—, la jerarquía de rendimiento se invierte. En estos sistemas, cada objeto se cuenta una sola vez al ingresar a una zona de interés; por tanto, cada ID switch representa un vehículo potencialmente contado doble (si el track se fragmenta antes del polígono) o no contado

(si se pierde). Bajo este criterio, la métrica determinante no es el MOTA sino la combinación de IDF1 y número de ID switches (Tabla 4.7):

Tabla 4.7
Evaluación orientada a aforo vehicular (Bogotá, 500 frames)

Tracker	IDF1	ID Sw.	Configurable	Idoneidad para aforo
SORT	84.91 %	103	Sí (`min_hits`)	Baja — alto riesgo de doble conteo
OC-SORT	84.97 %	106	Sí (`min_hits`)	Baja — mismo problema
DeepSORT	82.15 %	55	Sí (`n_init`)	Media
ByteTrack	87.28 %	28	**No**	Alta — pero sin ajuste de persistencia
Propuesto	**87.64 %**	**32**	**Sí**	**Alta — mejor IDF1 + configurable**

El tracker propuesto obtiene el **mejor IDF1** (87.64 %) con solo 32 ID switches. Es importante aclarar que este resultado no implica una recomendación de desplegar el tracker propuesto en producción —su velocidad de 0.7 FPS lo descalifica para aplicaciones en tiempo real—. Su valor reside en establecer una **referencia de calidad de seguimiento**: demuestra que es alcanzable un IDF1 de 87.64 % con 32 ID switches en este escenario, y este estándar debe servir como criterio de evaluación para los trackers rápidos que se desplieguen operativamente.

En esta perspectiva, ByteTrack emerge como el candidato más cercano a este estándar en tiempo real (28 ID switches, IDF1 87.28 %, 78 FPS). Sin embargo, ByteTrack no expone un parámetro de persistencia mínima configurable: su lógica interna de confirmación de tracks requiere modificar el código fuente para ajustar cuántos frames debe persistir un objeto antes de ser contabilizado. SORT, OC-SORT, DeepOcSort (Maggiolino et al., 2023) y StrongSORT (Du et al., 2023) sí permiten esta calibración mediante el parámetro `min_hits`, lo que facilita ajustar el compromiso entre cobertura y precisión del conteo según las condiciones de cada intersección, sin reescribir el algoritmo.

La contribución central de este análisis es metodológica: **para aplicaciones de aforo vehicular, el criterio de evaluación debe priorizar IDF1 y número de ID switches sobre MOTA**. SORT y OC-SORT, pese a su MOTA superior (89.5–89.6 %), generan más de 100 cambios de identidad en 500 frames. En un sistema de aforo con polígonos, esto se traduce potencialmente en un error de conteo del orden del 20 %, una magnitud que invalidaría los modelos de calibración de

congestión que requieren precisión superior al 85 %. Los 4.4pp de ventaja en MOTA no compensan este riesgo operacional. Este hallazgo cuestiona la práctica habitual en la literatura de MOT de reportar MOTA como métrica principal, al menos para el dominio de gestión de tráfico urbano.

El rendimiento del tracker propuesto mejora en Bogotá (+4.31 % MOTA respecto a Waymo en configuración estándar), lo cual se explica por las características favorables de la escena —cámara fija que elimina la contaminación por ego-movimiento y objetos vehiculares de mayor tamaño que facilitan la asociación espacial— más que por una adaptación específica del sistema al contexto local.

Tabla 4.8
Comparación del tracker propuesto entre datasets

Dataset	MOTA	IDF1	ID Sw.	Características
Waymo	80.95 %	73.50 %	817	Cámara móvil, vehículos
MOT17	63.19 %	64.92 %	1151	Cámara fija, peatones
Bogotá	**85.26 %**	**87.64 %**	**32**	Cámara fija, vehículos

4.2.5 Análisis de OSNet como Modelo de Apariencia

La evaluación comparativa de modelos de apariencia sobre Waymo (con pesos fijos IoU = 0.9, Apariencia = 0.1) arroja un resultado que merece discusión. OSNet supera a todos los modelos evaluados por 4.8–6.3 % MOTA con 2.2M parámetros, frente a los 5–12M de las alternativas. Esta ventaja resulta contraintuitiva, ya que modelos más grandes como EfficientNet (1280 dimensiones) o ResNet18 (512 dimensiones, 11.7M parámetros) podrían esperarse más discriminativos. La explicación más probable es que OSNet fue diseñado específicamente para re-identificación —su arquitectura omni-escala captura patrones a múltiples resoluciones relevantes para distinguir instancias similares— mientras que los modelos de propósito general fueron entrenados para clasificación, tarea que prioriza invariancia inter-clase sobre discriminabilidad intra-clase. La calidad de los embeddings depende más de la tarea de pre-entrenamiento que de la capacidad bruta del modelo, lo cual valida la selección de OSNet.

No obstante, dado que el peso de apariencia es solo 0.1, la contribución de cualquier modelo de apariencia al rendimiento global es limitada. Una exploración del espacio de pesos

IoU/apariencia con cada modelo podría revelar perfiles de rendimiento diferentes.

4.2.6 Suavidad de Trayectorias y Validación de la Parametrización (a, h)

Para evaluar la calidad de las trayectorias estimadas, se calcularon métricas de suavidad sobre todos los tracks con al menos 10 frames. Se midieron: el **jerk medio** (norma de la tercera derivada de la posición, menor = más suave), la **varianza de la aceleración** (menor = movimiento más consistente), y la **estabilidad de la altura** $\sigma(\Delta h)$ (desviación estándar del cambio frame-a-frame). Adicionalmente, se evaluó la linealidad de h y s mediante el coeficiente R^2 de un ajuste lineal, lo cual valida la justificación teórica de la parametrización (a, h) presentada en la Sección 2. Los resultados se resumen en la Tabla 4.9.

Tabla 4.9
Comparación de suavidad de trayectorias (Waymo, tracks ≥ 10 frames)

Tracker	Jerk medio	Var. acel.	$\sigma(\Delta h)$	$R^2(h)$	$R^2(s)$
Propuesto (8D)	**0.42**	**1.84**	**0.89**	**0.816**	**0.715**
ByteTrack	5.77	59.76	5.94	0.711	0.623
OC-SORT	8.37	41.73	6.59	0.712	0.634

El tracker propuesto produce trayectorias considerablemente más suaves: su jerk medio (0.42) es 14 veces menor que ByteTrack (5.77) y 20 veces menor que OC-SORT (8.37), mientras que la varianza de aceleración es 32 veces menor. Estas diferencias de uno a dos órdenes de magnitud reflejan el efecto del filtro de Kalman 8D, que filtra las oscilaciones frame-a-frame en las cajas delimitadoras producidas por el detector.

La métrica más relevante para el objetivo de este trabajo es la estabilidad de la altura, $\sigma(\Delta h)$, ya que la altura h es la variable directamente conectada con la estimación de distancia mediante proyección perspectiva ($d = fH/h$, Ecuación 2.13 del Capítulo 2). El tracker propuesto alcanza $\sigma(\Delta h) = 0{,}89$ píxeles, frente a 5.94 de ByteTrack y 6.59 de OC-SORT —una reducción de 6.7 a 7.4 veces en la variabilidad frame-a-frame de la altura estimada—. Dado que el error relativo en la distancia se propaga como $\delta d/d = \delta h/h$, para un vehículo típico con $h \approx 100$ píxeles, el tracker propuesto introduce un error de distancia frame-a-frame del $\sim$0.9 %, comparado con $\sim$5.9 % de ByteTrack y $\sim$6.6 % de OC-SORT. Para la estimación de velocidad, que depende

de la diferencia Δd entre frames consecutivos, el efecto es aún más pronunciado: las oscilaciones de ByteTrack y OC-SORT generan velocidades espurias de hasta $\pm 12\,\%$ por frame, mientras que el tracker propuesto las mantiene por debajo de $\pm 2\,\%$. Este resultado valida el objetivo central del diseño del tracker: producir detecciones suficientemente estables para habilitar estimaciones confiables de distancia y velocidad a través de las ecuaciones de proyección perspectiva.

No obstante, es importante señalar que la suavidad no es un objetivo en sí mismo: un tracker que suaviza en exceso puede enmascarar cambios reales de velocidad o dirección. La validación de que esta suavización refleja movimiento real —y no sobresuavizado— requeriría ground truth de trayectorias a nivel sub-píxel, que no está disponible en los datasets empleados. Un indicio favorable es que el tracker mantiene MOTA competitivo (80.95 % en configuración estándar), lo cual sugiere que la suavización no sacrifica la capacidad de seguir objetos correctamente.

Los resultados de linealidad apoyan la justificación teórica de la parametrización (a, h): la altura h presenta un ajuste lineal superior ($R^2 = 0{,}816$) frente al área s ($R^2 = 0{,}715$) para el tracker propuesto, y esta diferencia es consistente en todos los trackers evaluados. Esto indica que h evoluciona de forma más compatible con el supuesto de velocidad constante del filtro de Kalman lineal, reduciendo el error de predicción en cada paso temporal. Sin embargo, esta ventaja teórica debe contrastarse con los resultados de MOTA: ByteTrack, que emplea parametrización (s, r), logra 88.48 % MOTA en Waymo frente al 80.95 % del tracker propuesto en configuración estándar. Esto indica que la parametrización del estado es solo uno de varios factores —junto con la estrategia de asociación, el manejo de oclusiones y los hiperparámetros— que determinan el rendimiento global, y que sus beneficios se manifiestan más en la calidad de las trayectorias que en las métricas de detección.

4.3 Propuesta de Solución

4.3.1 Diagnóstico de la Situación Actual

La Secretaría de Movilidad de Bogotá opera actualmente con sistemas de videoanalítica cuya precisión no supera el 80 %, complementados con procesamiento manual por parte de operadores que pueden analizar entre 2 y 3 horas de video diarias. Esta combinación presenta múltiples desafíos operativos interrelacionados. La escalabilidad es reducida porque el procesamiento manual no permite cubrir la totalidad de la red de cámaras instaladas en la ciudad. La precisión resulta

variable debido a la fatiga humana y a la sensibilidad de los detectores a condiciones de iluminación y oclusión. La reactividad ante incidentes es lenta porque depende de que un operador identifique la anomalía. Fundamentalmente, los sistemas actuales no mantienen persistencia de identidades ni generan trayectorias completas, lo que impide alimentar modelos de calibración de congestión que requieren precisiones superiores al 85 % y datos de flujo vehicular consistentes.

4.3.2 Oportunidades Identificadas a partir de los Resultados

Los resultados experimentales identifican oportunidades concretas para abordar estas limitaciones, aunque con matices importantes. El tracker propuesto alcanza 85.26 % MOTA en Bogotá, lo cual, si bien supera marginalmente el umbral del 85 %, debe interpretarse con cautela: este resultado proviene de una sola intersección en condiciones diurnas favorables, y no puede garantizarse que se mantenga en escenarios más exigentes (lluvia, noche, congestión severa). Los trackers basados en IoU (SORT, ByteTrack) alcanzan MOTA superior (89–90 %) con velocidades de procesamiento compatibles con tiempo real (78–302 FPS), aunque con menor consistencia de identidades. La disponibilidad de múltiples configuraciones con diferentes perfiles de rendimiento permite proponer un sistema adaptable a las necesidades operativas específicas de cada caso de uso.

4.3.3 Arquitectura de la Solución Propuesta

La solución propuesta consiste en un sistema modular de seguimiento multi-objeto diseñado para operar en dos modos complementarios, seleccionables según el caso de uso y las restricciones de infraestructura disponible.

El **modo de monitoreo en tiempo real** emplearía ByteTrack como tracker principal, operando a 78 FPS con un MOTA de 89.15 % y solo 28 ID switches en las condiciones evaluadas. Este modo es adecuado para conteo vehicular continuo, detección de congestión y alimentación de tableros de mando en centros de control. Su bajo costo computacional permite procesar múltiples streams de video simultáneamente en una sola GPU. SORT representa una alternativa aún más ligera (302 FPS) cuando la velocidad de procesamiento es la prioridad absoluta y los cambios de identidad son tolerables.

El **modo de análisis detallado** utilizaría el tracker propuesto con embeddings OSNet y, opcionalmente, flujo óptico, alcanzando un IDF1 de 87.64 % a una velocidad de 0.7–1.2 FPS. Es-

te modo está orientado a tareas que requieren trayectorias completas y coherentes: estimación de tiempos de viaje entre puntos de control, construcción de matrices origen-destino, análisis forense de incidentes y detección de patrones de conducción de riesgo. Al operar en modo offline (procesamiento posterior a la captura), la velocidad reducida no constituye una limitación operativa.

4.3.4 Consideraciones para la Implementación

La implementación del sistema requeriría abordar varios aspectos que exceden el alcance experimental de este trabajo pero que son determinantes para su viabilidad en producción.

En primer lugar, la **calibración geométrica** de cada cámara es un requisito previo para convertir las velocidades estimadas en píxeles/frame a unidades físicas (km/h). Esto implica establecer una transformación perspectiva entre el plano de la imagen y el plano del suelo, utilizando puntos de referencia conocidos en la escena (distancias entre líneas de carril, longitud de marcas viales). Sin esta calibración, el sistema puede generar trayectorias y conteos pero no estimaciones de velocidad útiles para la fiscalización.

En segundo lugar, el **fine-tuning del detector YOLO** con imágenes del parque automotor colombiano mejoraría la calidad de las detecciones, particularmente para categorías de vehículos sub-representadas en los datos de entrenamiento originales (bicitaxis, buses articulados, bicicletas de carga). Dado que la calidad del detector establece el techo de MOTA alcanzable, esta inversión tendría un impacto directo en el rendimiento del sistema completo.

En tercer lugar, la **integración con la infraestructura existente** de la Secretaría requiere definir el pipeline de datos desde las cámaras de la red hasta el sistema de tracking. Esto incluye la captura y almacenamiento de video, el procesamiento (en el borde o en un centro de datos centralizado), la persistencia de resultados y la visualización para los operadores. La elección entre procesamiento en el borde (edge computing, cercano a la cámara) y procesamiento centralizado depende de la latencia tolerable y del ancho de banda disponible en la red.

Finalmente, se recomienda una **estrategia de despliegue por fases**. En una fase piloto inicial, se desplegaría ByteTrack en modo tiempo real sobre 3–5 cámaras en intersecciones críticas, estableciendo una línea base de rendimiento en condiciones operativas reales y no controladas. En paralelo, el tracker propuesto se utilizaría en modo offline para análisis retrospectivos, generando reportes de movilidad que validen la utilidad de las trayectorias coherentes para la Secretaría. Los

resultados de esta fase piloto informarían la decisión de escalar el sistema a un mayor número de cámaras y la eventual inversión en infraestructura de GPU para habilitar procesamiento con embeddings en tiempo real.

Es importante señalar que los resultados presentados en este trabajo constituyen una prueba de concepto: la validación se realizó en un dataset limitado (500 frames, 1 intersección, condiciones diurnas) y la transición a un sistema en producción requiere evaluación en un espectro mucho más amplio de condiciones y ubicaciones.

4.3.5 Trabajo Futuro: Evaluación Comparativa con `min_hits=5`

Un experimento pendiente de alto valor práctico consiste en ejecutar *todos* los trackers con `min_hits=5` (o su equivalente: `n_init=5` en StrongSORT) y comparar los ID switches resultantes. Este experimento permitiría evaluar si la ventaja del tracker propuesto en preservación de identidades se mantiene cuando todos los competidores operan bajo el mismo criterio de persistencia. SORT, OC-SORT, DeepOcSort y StrongSORT admiten este parámetro directamente; ByteTrack, al no exponer `min_hits`, quedaría excluido o requeriría modificación de su código fuente. Los datos de aforo vehicular recopilados por la Secretaría de Movilidad mediante su sistema de videoanalítica (videos de intersecciones con conteos manuales como referencia) constituyen un banco de pruebas ideal para esta evaluación, al proporcionar ground truth de conteo en condiciones operativas reales de Bogotá. Este análisis formaría parte de un estudio teórico en preparación sobre variantes del filtro de Kalman en MOT, donde se formaliza la relación entre la persistencia de confirmación, la fragmentación de tracks y la precisión del conteo vehicular.

4.4 Tabla de Hallazgos Principales

Tabla 4.10
Hallazgos principales alineados con objetivos específicos

Objetivo	**Hallazgo**	**Fuente**
Obj. 1: Detector YOLO	Sistema de detección y seguimiento con 87.65 % precisión y 99.60 % recall en Bogotá	Experimento Bogotá
Obj. 2: Flujo óptico + Kalman	Flujo óptico aporta +3.45 % MOTA con compensación de ego-movimiento	Ablación Waymo
Obj. 3: Embeddings CNN	OSNet supera alternativas por +4.8–6.3 % MOTA con 5x menos parámetros (2.2M)	Comparación modelos
General: Precisión $>$85 %	Tracker propuesto alcanza 85.26 % MOTA y 87.64 % IDF1 en Bogotá	Experimento Bogotá
Hiperparámetros	`min_hits=1` infla MOTA (+9.14 %) pero degrada IDF1 y dispara ID switches (929); `min_hits=5` produce la menor fragmentación (295 ID Sw) con pérdida marginal de MOTA	Ablación Waymo/MOT17
Aforo vehicular	El tracker propuesto es el único con mejor IDF1 (87.64 %) *y* configurabilidad completa de `min_hits` — idóneo para conteo por polígonos	Bogotá + análisis

4.5 Compromiso Velocidad vs Precisión

Tabla 4.11
Recomendación de tracker según caso de uso

Tracker	MOTA	ID Sw.	FPS	Caso de uso recomendado
SORT	89.63 %	103	302	Conteo instantáneo (no requiere trayectorias)
ByteTrack	89.15 %	**28**	79	Tiempo real con preservación de identidades
OC-SORT	89.50 %	106	63	Tiempo real, buen balance MOTA/velocidad
Propuesto (Full)	85.26 %	32	0.7	Análisis offline: tiempos de viaje, trayectorias
DeepSORT	78.99 %	55	1.9	Peatones en cámara estática

Los ID switches determinan la idoneidad de cada tracker para aplicaciones específicas de tráfico. SORT y OC-SORT, pese a su alto MOTA, producen más de 100 cambios de identidad en 500 frames, lo que invalida el cálculo de trayectorias completas. ByteTrack y el tracker propuesto, con 28 y 32 ID switches respectivamente, son los únicos adecuados para aplicaciones que requieren continuidad de identidades como estimación de tiempos de viaje o matrices origen-destino.

CAPÍTULO 5
DISCUSIÓN Y CONCLUSIONES

5.1 Discusión

5.1.1 Génesis del Diseño y Relación con la Literatura

El tracker propuesto no surgió de una revisión sistemática de la literatura en MOT, sino de un razonamiento independiente a partir de dos observaciones. La primera es que las ecuaciones de proyección perspectiva ($d = fH/h$) permiten estimar distancias y velocidades reales de vehículos a partir de la altura de sus detecciones en la imagen, pero requieren que dichas detecciones sean temporalmente estables. La segunda es que el filtro de Kalman, como estimador óptimo lineal, tiene la capacidad de filtrar el ruido de las detecciones CNN y producir estimaciones suavizadas de posición y dimensiones. La combinación de ambas observaciones condujo al diseño de un tracker con estado de 8 dimensiones que incluye la altura h y su velocidad $\dot{h}$ como variables de estado explícitas, complementado con flujo óptico para estimar velocidades y embeddings de apariencia para re-identificación. Una vez implementado y evaluado el sistema, se identificó que la literatura ya contenía enfoques similares —DeepSORT (Wojke et al., 2017) combina filtro de Kalman con embeddings de apariencia— lo cual valida la intuición detrás del diseño. Sin embargo, las diferencias en parametrización del estado (8D con (a, h) en lugar de 7D con (s, r)), la integración de flujo óptico y el objetivo explícito de estabilizar las detecciones para proyección perspectiva distinguen la propuesta de los enfoques existentes.

En términos de rendimiento comparativo, los resultados son consistentes con la literatura existente. ByteTrack (Zhang et al., 2022) obtiene el mejor MOTA en Waymo (88.48 %), confirmando la efectividad de su asociación en cascada reportada por los autores originales. El tracker propuesto, en su configuración estándar, alcanza 80.95 % MOTA —7.53pp por debajo—, una diferencia que refleja la ventaja de la estrategia de doble umbral de ByteTrack para recuperar detecciones de baja confianza.

La configuración con `min_hits=1` reduce esta brecha a 0.7pp (87.78 % vs 88.48 %), pero este resultado debe interpretarse con honestidad: el ajuste de `min_hits` no mejora la calidad del tracking sino que relaja el criterio de confirmación de tracks, aceptando objetos desde su primera

detección a costa de un deterioro sustancial en IDF1 (de 73.50 % a 65.88 %) y un aumento en ID switches (de 817 a 929). Por tanto, la cercanía en MOTA no implica equivalencia en rendimiento global, y la comparación más informativa es la de configuración estándar, donde ByteTrack mantiene una ventaja clara.

Por otro lado, la superioridad de OC-SORT (Cao et al., 2023) en preservación de identidades (407 ID switches frente a 876 de ByteTrack) confirma la efectividad de su mecanismo de re-actualización centrado en observaciones (ORU). Este resultado es particularmente relevante para el contexto de este trabajo, dado que las aplicaciones de gestión vial priorizan trayectorias coherentes sobre cobertura instantánea.

5.1.2 Implicaciones Teóricas

Los resultados permiten extraer tres implicaciones teóricas, cada una con alcances y limitaciones que conviene explicitar.

En relación con la optimalidad del filtro de Kalman, los experimentos confirman que el filtro de Kalman lineal (Kalman, 1960) sigue siendo una base robusta para MOT cuando se combina con información complementaria. Sin embargo, esta conclusión debe matizarse: el filtro de Kalman es óptimo bajo supuestos de linealidad y ruido gaussiano que se cumplen solo aproximadamente en el seguimiento vehicular real. Las mejoras de suavidad observadas (jerk 14x menor) pueden reflejar tanto una estimación más precisa como un sobresuavizado que enmascara cambios reales de dinámica. La ausencia de ground truth de trayectorias sub-píxel impide distinguir entre ambos efectos.

Respecto al flujo óptico en cámaras móviles, los resultados contradicen la recomendación habitual de desactivarlo por contaminación de ego-movimiento. Con compensación basada en la mediana del flujo global, se obtienen mejoras de +3.45 % MOTA en Waymo (cámara móvil) y +1.53 % en MOT17 (cámara estática). Si bien estas mejoras son modestas en magnitud absoluta, resultan consistentes entre datasets y escenarios, lo que sugiere un efecto genuino y no un artefacto experimental. La mayor contribución en cámara móvil es coherente con la teoría, ya que el flujo óptico proporciona información de velocidad complementaria que resulta más valiosa cuando el ego-movimiento introduce incertidumbre adicional en las predicciones del Kalman. No obstante, la compensación por mediana global es una aproximación simplificada —asume ego-movimiento

traslacional uniforme— que podría degradarse en escenas con rotación significativa de la cámara.

En cuanto a la transferencia de modelos de apariencia, OSNet (Zhou et al., 2019), diseñado originalmente para re-identificación de peatones, transfiere efectivamente al dominio vehicular, superando modelos de propósito general (ResNet18, EfficientNet, MobileNet) con 5 veces menos parámetros. Este resultado es relevante para aplicaciones donde el cómputo es limitado, aunque se evaluó solo sobre vehículos en escenas urbanas con iluminación diurna; su rendimiento en condiciones de iluminación adversa o con categorías vehiculares atípicas queda por validar.

5.1.3 Implicaciones Prácticas para la Secretaría de Movilidad

La validación en el dataset de Bogotá ofrece evidencia preliminar sobre la viabilidad del sistema, aunque con alcance limitado. La precisión de 85.26 % MOTA supera marginalmente el umbral requerido del 85 %, pero este resultado proviene de una sola intersección en condiciones favorables (iluminación diurna, densidad vehicular moderada). Extrapolar este rendimiento a la totalidad de la malla vial de Bogotá —que incluye escenarios nocturnos, lluvia, congestión severa y tipos de cámara heterogéneos— requeriría una validación considerablemente más amplia.

Dicho esto, los resultados permiten identificar valor potencial en tres dimensiones. En primer lugar, el IDF1 de 87.64 % sugiere que el sistema puede generar trayectorias suficientemente coherentes para alimentar modelos de estimación de tiempos de viaje y matrices origen-destino, aplicaciones donde la consistencia de identidades es más importante que la cobertura instantánea. En segundo lugar —y quizá más relevante para las necesidades operativas de la Secretaría— la estabilidad de las detecciones producidas por el tracker propuesto ($\sigma(\Delta h) = 0{,}89$ píxeles, frente a 5.94 de ByteTrack) habilita la estimación de distancias y velocidades mediante proyección perspectiva con errores frame-a-frame del orden del 1 %, frente al 6 % de los trackers que no filtran el ruido del detector. Para los modelos de calibración de congestión que requieren datos de velocidad vehicular, esta estabilidad es potencialmente más valiosa que la diferencia de 4pp en MOTA. En tercer lugar, la disponibilidad de múltiples configuraciones —desde ByteTrack a 78 FPS para monitoreo continuo hasta el tracker propuesto a 0.7 FPS para análisis detallado— ofrece flexibilidad para adaptar el sistema a diferentes necesidades operativas sin cambiar la infraestructura subyacente. El hecho de que todos los componentes se basen en herramientas open-source (PyTorch, OpenCV, FilterPy) elimina costos de licenciamiento, aunque el costo de infraestructura de GPU y

el esfuerzo de integración con los sistemas existentes de la Secretaría no deben subestimarse.

5.1.4 Selección del Tracker según Características de la Escena

Los resultados experimentales revelan que no existe una configuración de tracker universalmente óptima; la elección de componentes depende de las características de la escena y los requisitos operativos. Esta conclusión, si bien puede parecer obvia, está respaldada cuantitativamente por los datos recopilados.

En escenarios con cámaras fijas de tráfico —como las que componen la red de la Secretaría de Movilidad de Bogotá— la compensación de ego-movimiento es innecesaria y trackers basados exclusivamente en IoU como ByteTrack ofrecen el mejor balance entre rendimiento (89.15 % MOTA en Bogotá) y velocidad de procesamiento (78 FPS). En estos escenarios, la adición de embeddings de apariencia y flujo óptico no compensa la pérdida de velocidad ni la reducción de MOTA observada con el tracker propuesto (85.26 %). En contraste, el escenario de cámara móvil en Waymo muestra que el flujo óptico con compensación de ego-movimiento aporta +3.45 % MOTA, un beneficio que justifica su costo computacional cuando la cámara se desplaza.

La densidad de oclusiones constituye un segundo eje de decisión. En escenas con oclusiones frecuentes —intersecciones congestionadas, hora pico— los embeddings de apariencia permiten reidentificar vehículos que reaparecen tras quedar ocultos, aportando +2.76 % MOTA según el estudio de ablación. Sin embargo, esta mejora se midió en el agregado del dataset; en escenas con flujo libre y baja oclusión, la asociación por IoU resulta suficiente y permite procesamiento en tiempo real. Esta dependencia entre configuración y escenario subraya la importancia de que la Secretaría de Movilidad no adopte un tracker único para toda su red, sino que seleccione componentes según las condiciones específicas de cada ubicación y el tipo de análisis requerido.

5.1.5 Limitaciones del Estudio

El estudio presenta varias limitaciones que condicionan el alcance de las conclusiones. La más fundamental es el compromiso entre MOTA e ID switches: la configuración con `min_hits=1` que alcanza 87.78 % MOTA en Waymo genera 929 cambios de identidad, más del doble que la configuración estándar (432). Esta tensión implica que no existe una configuración universalmente óptima, y la elección depende del caso de uso específico —un matiz que la presentación de

"resultados óptimos" puede oscurecer.

En cuanto a velocidad de procesamiento, la configuración completa (flujo óptico + embeddings) opera a 0.7–2 FPS, dos órdenes de magnitud por debajo de lo requerido para procesamiento en tiempo real. La extracción de embeddings OSNet constituye el principal cuello de botella, y aunque técnicas como el batching en GPU y el cálculo selectivo podrían reducir esta brecha, su implementación queda fuera del alcance de este trabajo. Esta limitación restringe el uso del tracker propuesto a escenarios de análisis offline.

La generalización de los resultados está acotada por las condiciones experimentales. La validación en Bogotá se realizó con un solo video de 500 frames en una intersección con iluminación diurna favorable, lo cual no cubre el espectro de condiciones operativas reales: lluvia, niebla, noche, congestión severa, cámaras con diferentes ángulos y resoluciones. Asimismo, el dataset se etiquetó mediante auto-anotación con BoTSORT+ReID seguida de verificación manual, lo que introduce un sesgo potencial si el tracker de anotación comparte limitaciones con los trackers evaluados.

Por último, los resultados carecen de análisis de significancia estadística. Todas las métricas reportadas provienen de ejecuciones únicas, sin intervalos de confianza ni pruebas de hipótesis que permitan distinguir diferencias genuinas de variabilidad experimental. Diferencias pequeñas como los 0.36pp en IDF1 entre el tracker propuesto y ByteTrack en Bogotá, o los 0.2pp de la no-aditividad de componentes, podrían no ser estadísticamente significativas. Trabajo futuro debe incluir múltiples ejecuciones con semillas aleatorias diferentes y, de ser posible, evaluación cruzada sobre múltiples escenas.

5.2 Conclusiones

Las conclusiones se organizan según los objetivos específicos planteados, seguidas de una conclusión general que integra los hallazgos del estudio.

5.2.1 Conclusión 1: Detección YOLO (OE1)

El sistema de detección y seguimiento basado en YOLO alcanza 87.65 % de precisión y 99.60 % de recall en el dataset de Bogotá, proporcionando una base adecuada para el tracking. Conviene señalar que la calidad de las detecciones establece un techo para el MOTA alcanzable por cualquier tracker, como se evidencia en MOT17 donde el detector SDP con $\sim$65 % recall limita a

todos los trackers a un rango estrecho de rendimiento. Por tanto, mejoras en el detector —como fine-tuning con datos del parque automotor colombiano— representan posiblemente la vía más directa para incrementar el rendimiento global del sistema.

5.2.2 Conclusión 2: Flujo Óptico + Kalman (OE2)

La integración del flujo óptico de Lucas-Kanade con el filtro de Kalman aporta mejoras consistentes pero modestas: +3.45 % MOTA en Waymo (cámara móvil) y +1.53 % en MOT17 (cámara estática), ambas con compensación de ego-movimiento. Estas magnitudes indican que el flujo óptico es un componente útil pero no transformador del sistema. El hallazgo de que su contribución es mayor en cámara móvil que en cámara estática resulta coherente con la teoría y sugiere su potencial para escenarios de conducción autónoma o drones, aunque la compensación empleada (mediana global) es una aproximación que debería refinarse para movimientos de cámara más complejos.

5.2.3 Conclusión 3: Embeddings CNN (OE3)

Entre los modelos de apariencia evaluados, OSNet obtiene los mejores resultados, superando a ResNet18, EfficientNet y MobileNet por +4.8–6.3 % MOTA con 5 veces menos parámetros (2.2M). La transferencia efectiva de un modelo diseñado para re-identificación de peatones al dominio vehicular es un resultado alentador, aunque debe notarse que la evaluación se realizó con un peso fijo IoU/apariencia de 0.9/0.1, lo que limita la contribución de cualquier modelo de apariencia. Una exploración más amplia del espacio de pesos podría revelar diferencias distintas entre modelos.

5.2.4 Conclusión General

El tracker propuesto alcanza 85.26 % MOTA y 87.64 % IDF1 en el dataset de Bogotá, lo que lo sitúa marginalmente por encima del umbral del 85 % establecido como requisito. Este resultado, si bien satisface el criterio numérico planteado, debe enmarcarse en su contexto: proviene de una sola intersección en condiciones favorables y sin intervalos de confianza que respalden la diferencia respecto al umbral. La viabilidad del sistema para la Secretaría de Movilidad queda condicionada a una validación más amplia.

En la comparación con el estado del arte, los trackers basados exclusivamente en IoU (ByteTrack, SORT) superan al tracker propuesto en MOTA por 4–4.4pp en Bogotá. Sin embargo, esta comparación merece una reflexión sobre la validez de MOTA como métrica única de evaluación. MOTA penaliza por igual los falsos positivos, falsos negativos y los cambios de identidad, pero no distingue entre una detección correcta que mantiene la identidad del objeto y una que la pierde. Un tracker con MOTA alto pero muchos ID switches —como SORT (89.63 % MOTA, 103 ID switches)— está asociando correctamente una caja a un objeto en cada frame, pero asignándole identidades diferentes a lo largo del tiempo. Para conteo vehicular simple (cuántos vehículos hay en un frame), esto es aceptable. Pero para las aplicaciones que motivan este trabajo —estimación de velocidad mediante proyección perspectiva, cálculo de tiempos de viaje, matrices origen-destino— un cambio de identidad corrompe la trayectoria completa y hace que el MOTA "alto" sea engañoso: el tracker detecta el vehículo pero no puede rastrear su recorrido. El IDF1, que evalúa la consistencia de las identidades a lo largo del tiempo, captura mejor esta dimensión. En este sentido, el tracker propuesto (87.64 % IDF1, 32 ID switches) y ByteTrack (87.28 % IDF1, 28 ID switches) son los únicos adecuados para las aplicaciones objetivo, mientras que SORT y OC-SORT, pese a su MOTA superior, producen trayectorias fragmentadas que invalidan el análisis posterior.

Más allá de las métricas de tracking tradicionales (MOTA, IDF1), los resultados de suavidad de trayectorias revelan que la contribución más distintiva del tracker propuesto no se mide en puntos porcentuales de MOTA sino en la estabilidad de las detecciones: una reducción de 6.7 veces en la variabilidad de la altura estimada ($\sigma(\Delta h) = 0{,}89$ vs 5.94 de ByteTrack) que se traduce directamente en estimaciones de distancia y velocidad más confiables mediante proyección perspectiva. Mientras que en MOTA el tracker propuesto queda por debajo de ByteTrack y SORT, en la dimensión que motivó su diseño —la estabilidad para cálculo de magnitudes físicas— la ventaja es clara y cuantitativamente significativa.

El estudio de ablación aporta una contribución complementaria: una cuantificación sistemática del impacto de cada componente e hiperparámetro, que permite a un implementador tomar decisiones informadas sobre qué activar según su escenario específico. En particular, la evidencia de que los componentes (flujo óptico + embeddings) son parcialmente redundantes entre sí, y que el ajuste de hiperparámetros como `min_hits` tiene un efecto mayor que la adición de componentes algorítmicos —aunque operando en dimensiones diferentes del rendimiento—, constituye una guía

práctica para el despliegue del sistema. Los resultados confirman que las cámaras fijas de tráfico vehicular representan el escenario más favorable para el sistema propuesto.

5.3 Contribuciones Principales

La contribución central de este trabajo es el estudio de ablación sistemático con 14 configuraciones, que cuantifica el impacto individual de cada componente (flujo óptico, embeddings de apariencia) e hiperparámetro (`min_hits`, `iou_thresh`) sobre el rendimiento del tracker. Este tipo de análisis desagregado es poco frecuente en la literatura de MOT, donde los sistemas suelen presentarse como unidades monolíticas, y proporciona criterios concretos para la selección de componentes según las restricciones de cada escenario.

Como segunda contribución, se presenta una evaluación comparativa de trackers del estado del arte en condiciones de tráfico de Bogotá. Si bien el alcance es limitado (una intersección, 500 frames, condiciones diurnas), constituye —hasta donde se tiene conocimiento— uno de los primeros benchmarks de tracking vehicular con datos de la malla vial colombiana, proporcionando evidencia inicial para la toma de decisiones institucional en la Secretaría de Movilidad.

En tercer lugar, la estabilización de detecciones para proyección perspectiva constituye una contribución que trasciende las métricas tradicionales de MOT. La reducción de 6.7 veces en la variabilidad de la altura estimada ($\sigma(\Delta h) = 0{,}89$ vs 5.94 de ByteTrack) habilita el cálculo de distancias y velocidades mediante las ecuaciones de cámara *pinhole* con errores frame-a-frame inferiores al 1 %, un resultado que responde directamente a la necesidad operativa que motivó el diseño del tracker.

De forma complementaria, el trabajo ofrece guías de configuración con recomendaciones específicas para diferentes casos de uso, desde monitoreo en tiempo real con ByteTrack o SORT hasta análisis detallado de trayectorias con el tracker propuesto. Estas guías traducen los resultados experimentales en criterios accionables para un implementador. La demostración de que OSNet transfiere efectivamente del dominio de re-identificación peatonal al vehicular, con rendimiento superior a modelos de propósito general y menor costo computacional, es un resultado con aplicabilidad más allá del contexto específico de este estudio. Finalmente, la exploración comparativa realizada en este trabajo generó el conocimiento profundo que dio origen a un análisis teórico formal de las variantes del filtro de Kalman en MOT (Moreno Bedoya, 2026), actualmente en revisión,

ampliando el alcance de la contribución al ámbito de los fundamentos teóricos de estos sistemas.

5.4 Trabajo Futuro

En el ámbito de la optimización, resulta prioritario implementar la extracción de embeddings por lotes en GPU y el cálculo selectivo de flujo óptico para alcanzar velocidades de procesamiento compatibles con tiempo real. De manera complementaria, el fine-tuning de YOLO con imágenes del parque automotor colombiano permitiría mejorar la precisión del detector en condiciones locales.

En cuanto a la validación extendida, es necesario evaluar el rendimiento del sistema en condiciones adversas (lluvia, noche, congestión severa) y en múltiples ubicaciones de Bogotá. Asimismo, la implementación de calibración geométrica con transformación perspectiva permitiría estimar velocidades en km/h en lugar de px/frame, lo cual es un requisito para aplicaciones operativas de la Secretaría de Movilidad.

Respecto a capacidades avanzadas, el desarrollo de fusión multi-cámara con asociación de tracks entre cámaras habilitaría el seguimiento continuo a través de intersecciones.

De manera notable, la experiencia adquirida durante el desarrollo y evaluación comparativa del tracker propuesto —particularmente la comprensión profunda de las fortalezas y debilidades de cada representación de estado (7D vs 8D), estrategia de asociación y mecanismo de manejo de oclusión— condujo a un trabajo de investigación complementario (Moreno Bedoya, 2026), actualmente en revisión, que presenta un análisis teórico riguroso de las variantes del filtro de Kalman en trackers MOT modernos (SORT, DeepSORT, ByteTrack, OC-SORT, BoT-SORT, StrongSORT). Dicho trabajo formaliza en diez enunciados (teoremas, proposiciones y lemas) las condiciones bajo las cuales cada modificación es óptima y demuestra que estos resultados teóricos permiten seleccionar el tracker más adecuado para una escena dada a partir de sus características (tipo de objeto, frecuencia de oclusiones, movimiento de cámara). Validado experimentalmente en un corredor urbano de Bogotá (Calle 13), el análisis teórico predijo correctamente que BoT-SORT (Aharon et al., 2022) superaría a otros trackers para conteo vehicular en esa escena, con mejoras particularmente marcadas en las categorías de vehículos con relación de aspecto variable (motocicletas, bicicletas) —precisamente donde la teoría identifica la superioridad de la parametrización 8D—. Este resultado valida la hipótesis de que la exploración empírica realizada en el presente trabajo, lejos de ser

un ejercicio de benchmarking rutinario, generó el entendimiento profundo necesario para formular criterios teóricos de selección de tracker adaptados a cada escenario de despliegue.

5.5 Recomendaciones para Implementación

A partir de los resultados obtenidos, se recomienda una estrategia de implementación gradual que permita validar el sistema en condiciones operativas reales antes de un despliegue a escala. En una fase piloto inicial, se sugiere desplegar ByteTrack para monitoreo en tiempo real en un número reducido de cámaras críticas (3–5 intersecciones), dado que este tracker ofrece el mejor balance entre MOTA (89.15 %), ID switches (28) y velocidad (78 FPS) en el escenario evaluado. Los resultados de esta fase permitirían establecer una línea base de rendimiento en condiciones no controladas y evaluar la robustez del sistema ante variaciones de iluminación, clima y densidad vehicular que no fueron cubiertas en este estudio.

En paralelo, el tracker propuesto con embeddings OSNet podría emplearse para análisis offline: generación de reportes de movilidad, construcción de matrices origen-destino y análisis forense de incidentes, aplicaciones donde la velocidad de procesamiento no es una restricción y la preservación de identidades (IDF1 87.64 %) aporta valor diferencial. Para esta modalidad, la configuración recomendada es `min_hits=3` (no `min_hits=1`, que mejora MOTA a costa de IDF1), `iou_thresh=0.2` y pesos IoU/apariencia de 0.9/0.1, con una GPU NVIDIA de gama media o superior para la extracción de embeddings.

Es importante reiterar que estas recomendaciones se basan en resultados experimentales de alcance limitado, y que la decisión de despliegue institucional debería condicionarse a una validación más amplia que cubra múltiples ubicaciones, condiciones climáticas y horarios de operación.

REFERENCIAS

Aharon, N., Orfaig, R., & Bobrovsky, B.-Z. (2022). BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *arXiv preprint arXiv:2206.14651*.

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing*, *2008*, 1-10. https://doi.org/10.1155/2008/246309

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple Online and Realtime Tracking. *IEEE International Conference on Image Processing (ICIP)*, 3464-3468. https://doi.org/10.1109/ICIP.2016.7533003

Bloomberg Philanthropies. (2004). *Initiative for Global Road Safety* (Reporte sobre seguridad vial global). Bloomberg Philanthropies.

Cao, J., Weng, X., Khirodkar, R., Pang, J., & Kitani, K. (2023). Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9686-9696.

Ciaparrone, G., Luque Sánchez, F., Tabik, S., Troiano, L., Tagliaferri, R., & Herrera, F. (2020). Deep Learning in Video Multi-Object Tracking: A Survey. *Neurocomputing*, *381*, 61-88. https://doi.org/10.1016/j.neucom.2019.11.023

Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., & Leal-Taixé, L. (2021). MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. *International Journal of Computer Vision*, *129*, 845-881. https://doi.org/10.1007/s11263-020-01393-0

Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., & Meng, H. (2023). StrongSORT: Make DeepSORT Great Again. *IEEE Transactions on Multimedia*, *25*, 8725-8737. https://doi.org/10.1109/TMM.2023.3240881

Farnebäck, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. *Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA)*, *2749*, 363-370.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Hernández Sampieri, R., & Mendoza Torres, C. P. (2018). *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta* (7.ª ed.). McGraw-Hill.

Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, *82*(Series D), 35-45.

Kuhn, H. W. (1955). The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, *2*(1-2), 83-97. https://doi.org/10.1002/nav.3800020109

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision (ECCV)*, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

Lucas, B. D., & Kanade, T. (1981). An Iterative Image Registration Technique with an Application to Stereo Vision. *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 674-679.

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., & Leibe, B. (2021). HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking. *International Journal of Computer Vision*, *129*(2), 548-578. https://doi.org/10.1007/s11263-020-01375-2

Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., & Kim, T.-K. (2021). Multiple Object Tracking: A Literature Review. *Artificial Intelligence*, *293*, 103448. https://doi.org/10.1016/j.artint.2020.103448

Maggiolino, G., Ahmad, A., Cao, J., & Kitani, K. (2023). Deep OC-SORT: Multi-Pedestrian Tracking by Adaptive Re-Identification. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 3025-3029.

Milan, A., Leal-Taixe, L., Reid, I., Roth, S., & Schindler, K. (2016). MOT16: A Benchmark for Multi-Object Tracking. *arXiv preprint arXiv:1603.00831*.

Moreno Bedoya, D. L. (2026). *Understanding Motion Models in Multi-Object Tracking: A Rigorous Analysis from SORT to OC-SORT with Theory-Guided Tracker Selection* [Manuscript under review].

Munkres, J. (1957). Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*, *5*(1), 32-38.

Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). A Survey on Performance Metrics for Object-Detection Algorithms. *Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP)*, 237-242.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-788.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 658-666.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *European Conference on Computer Vision (ECCV) Workshops*, 17-35. https://doi.org/10.1007/978-3-319-48881-3_2

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510-4520.

Secretaría Distrital de Movilidad. (2023). Plan Distrital de Desarrollo 2024-2027: Bogotá Camina Segura.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., … Anguelov, D. (2020). Scalability in Perception for Autonomous Driving: Waymo Open Dataset. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2446-2454.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the International Conference on Machine Learning (ICML)*, 6105-6114.

Welch, G., & Bishop, G. (2006a). An Introduction to the Kalman Filter. *UNC-Chapel Hill, TR 95-041*.

Welch, G., & Bishop, G. (2006b). *An Introduction to the Kalman Filter* (inf. téc. N.º TR 95-041). University of North Carolina at Chapel Hill.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple Online and Realtime Tracking with a Deep Association Metric. *IEEE International Conference on Image Processing (ICIP)*, 3645-3649. https://doi.org/10.1109/ICIP.2017.8296962

World Health Organization. (2023). Global Status Report on Road Safety 2023.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., & Wang, X. (2022). ByteTrack: Multi-Object Tracking by Associating Every Detection Box. *European Conference on Computer Vision (ECCV)*, 1-21. https://doi.org/10.1007/978-3-031-20047-2_1

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2024). A Survey on Multi-Object Tracking: Methods, Datasets, and Metrics. *ACM Computing Surveys*, *56*(3), 1-36. https://doi.org/10.1145/3630104

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable Person Re-identification: A Benchmark. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1116-1124.

Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). Omni-Scale Feature Learning for Person Re-Identification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 3702-3712.