Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Estimación de Posición de una Cámara Monocular Utilizando Aprendizaje Continuo

Tesis presentada para obtener el grado de

Doctor en Ingeniería del Lenguaje y del Conocimiento

Presenta:

M.C. Aldrich Alfredo Cabrera Ponce

Director de Tesis:

Dr. Manuel Isidrio Martin Ortíz, BUAP

Co-director de Tesis:

Dr. José Martínez Carranza, INAOE

Asesores de Tesis:

Dr. José Arturo Olvera López, BUAP
Dr. Roberto Olmos Pimentel, BUAP
Dr. Hiram Eredín Ponce Espinosa, UP

Puebla, Enero de 2026

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Quiero expresar mi agradecimiento al Dr. Manuel Isidro Martín Ortíz por haberme aceptado como su estudiante de doctorado, permitiéndome formar parte de la BUAP y ofreciéndome su confianza y apoyo continuo. Agradezco también las facilidades brindadas para utilizar los recursos del supercómputo nacional que hicieron posible el desarrollo de esta investigación. Asimismo, extiendo un enorme agradecimiento a mi coasesor, el Dr. José Martínez Carranza, por haberme aceptado como estudiante y por todo su respaldo académico y personal. Sus consejos me encaminaron hacia el mundo de la robótica, y su invitación a integrarme al equipo QuetzalC++ me permitió participar activamente en proyectos de investigación que fortaleció mi formación. Le agradezco profundamente su paciencia, orientación y la oportunidad de colaborar en la organización del Taller de Drones 2023 y 2024, MICAI 2024, MEXCIR 2025 e IMAV 2025.

Agradezco sinceramente a la Benemérita Universidad Autónoma de Puebla (BUAP) por brindarme acceso a sus instalaciones y por el apoyo financiero otorgado para mi participación en conferencias nacionales e internacionales.

Extiendo también mi agradecimiento al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) por recibirme durante mi estancia de investigación y permitirme continuar mis estudios utilizando sus laboratorios y espacios de trabajo.

Esta investigación fue posible gracias al apoyo financiero de la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), a través de la beca No. 802791, CVU 924462, cuyo respaldo agradezco profundamente.

De igual manera, agradezco a los miembros de mi jurado: Dr. José Arturo Olvera López, Dr. Roberto Olmos Pimentel y Dr. Hiram Eredín Ponce Espinosa, por sus valiosos comentarios, observaciones y retroalimentación constructiva a lo largo de estos años, los cuales enriquecieron enormemente este trabajo.

Finalmente, quiero agradecer eternamente a mi mamá, por su apoyo incondicional durante todo el transcurso de mis estudios, por su paciencia, comprensión y por motivarme siempre a seguir adelante y no rendirme.

Resumen

Esta tesis presenta el desarrollo de estrategias de aprendizaje continuo para la localizacion visual de una cámara monocular a bordo de un dron en situaciones de pérdida de señal GPS. El objetivo principal fue diseñar una metodología de localización de respaldo capaz de proporcionar una posición aproximada a la real, mientras incorpora de manera dinámica nueva información del entorno. Para lograr este objetivo, se propusieron tres esquemas de localización basados en aprendizaje continuo: localización topológica, localización jerárquica y localización progresiva utilizando la percepción visual del escenario mediante imágenes aéreas. Además, se implementaron dos estrategias de aprendizaje continuo y un mecanismo de búsqueda para asociar las imágenes de prueba con su posición correspondiente.

Los experimentos se realizaron en escenarios controlados con datos reales capturados en vuelo, evaluando el desempeño de las propuestas frente a trayectorias discontinuas y diversos entornos. Los resultados demostraron que las metodologías permiten recuperar la localización aproximada a la velocidad de procesamiento de la imagen, demostrando su uso como un sistema de respaldo ante fallas de GPS. A diferencia de enfoques tradicionales que requieren largos tiempos de entrenamiento y grandes conjuntos de datos, en este trabajo se exploró el uso de redes ligeras y arquitecturas binarias capaces de estimar la posición en tiempo cercano a la frecuencia de captura de la cámara.

Las principales contribuciones de esta tesis incluyen: 1) El desarrollo de estrategias de aprendizaje continuo especializadas a la localización visual; 2) La creación de una metodología con múltiples modelos que permite asignar posiciones a lo largo de la trayectoria; 3) El uso de redes ligeras como modelos de Máquinas de Soporte Vectorial para Regresión y redes binarias que pueden competir con arquitecturas profundas en términos de precisión y eficiencia. Estos avances ofrecen una base para el desarrollo de futuros sistemas de localización visual de respaldo, aplicados especialmente en entornos donde no se dispone de un sistema de localización basado en posicionamiento externo.

Abstract

This thesis presents the development of continual learning strategies for visual localisation of a monocular camera onboard a drone in situations of GPS signal loss. The main objective was to design a backup localisation methodology capable of providing an approximate position while dynamically incorporating new information. To achieve this, three localisation schemes based on continual learning were proposed: topological localisation, hierarchical localisation, and progressive localisation, all using visual perception of the environment through aerial images. In addition, two continual learning strategies and a search mechanism were implemented to associate test images with their corresponding positions.

The experiments were conducted in controlled scenarios with real flight data, evaluating the performance of the proposed methods against discontinuous trajectories and diverse environments. The results demonstrated that the methodologies are able to recover approximate localisation at the image processing rate, confirming their usefulness as a backup system in the event of GPS failure. Unlike traditional approaches that require long training times and large datasets, this work explored the use of lightweight networks and binary architectures, capable of estimating the position in near real-time, at a rate close to the camera's capture frequency.

The main contributions of this thesis include: (1) the development of continual learning strategies specialised for visual localisation; (2) the creation of a multi-model methodology that assigns positions along the flight trajectory; and (3) the use of lightweight models such as Support Vector Regression and Binary Networks, which can compete with deep architectures in terms of accuracy and efficiency. These advances provide a benchmark for the development of future visual backup localisation systems, particularly in environments where no external positioning infrastructure is available.

Capítulo 1

Introducción

La estimación de posición a partir de imágenes monoculares representa uno de los retos más importantes en robótica aérea, donde se busca localizar un dron dentro de un entorno. A diferencia de enfoques tradicionales que requieren sistemas de posicionamiento global (GPS, por sus siglas en inglés), sensores adicionales o mapas previos, este método utiliza únicamente la información visual capturada por una cámara monocular para estimar la posición tridimensional. Esto es inspirado por la capacidad humana de interpretar entornos complejos a partir de la visión, buscando replicar esa habilidad en sistemas de localización.

Métodos visuales para tareas de localización utilizando drones han desempeñado un gran avance en la robótica, permitiendo extraer y procesar las características visuales presentes en una escena para su interpretación. Por ejemplo, el uso de descriptores visuales y sistemas de mapeo han demostrado su capacidad de reconocimiento en entornos tanto interiores como exteriores. Esto permite determinar la ubicación de un dron utilizando únicamente su cámara en entornos donde la señal GPS es débil o inexistente. Sin embargo, este proceso es funcional en escenarios altamente texturizados donde las características son clave para el reconocimiento del entorno.

Esta investigación se enfoca en el desarrollo de métodos para estimar la posición y localización de drones en entornos complejos. Estos métodos aprovechan la percepción visual obtenida a partir de imágenes aéreas capturadas por una cámara monocular a bordo del dron, procesando las características del entorno para su interpretación. Este estudio define entornos complejos como escenarios dinámicos y cambiantes, donde existen variaciones de iluminación, aparición de nuevos elementos, presencia de nuevas estructuras u objetos y desaparición de elementos anteriores. Los objetivos específicos de esta investigación son, por un lado, lograr la estimación de la posición tridimensional del dron durante el vuelo, y por otro, permitir que el sistema aprenda y se adapta visualmente a los cambios del entorno.

Para alcanzar estos objetivos en entornos cambiantes, esta investigación se centra en el uso de estrategias de aprendizaje continuo aplicadas a modelos de percepción visual. El estudio desarrolló métodos para permitir que un modelo de aprendizaje profundo se actualice de forma continua con nueva información, sin olvidar el conocimiento previamente adquirido. Este enfoque asegura que la localización pueda adaptarse a las variaciones del entorno, manteniendo la posición del dron.

La fase experimental de este estudio consistió en pruebas en entornos reales al aire libre, utilizando distintas trayectorias de vuelo para evaluar la capacidad del sistema de localización visual. El dron cuenta únicamente con una cámara monocular a bordo para la captura de imágenes aéreas, las cuales se utilizaron como datos de entrada para entrenar y evaluar los modelos de aprendizaje continuo. Los experimentos permitieron medir métricas como el error de estimación de posición, el tiempo de inferencia, la precisión visual en la localización y la velocidad de operación, facilitando el análisis del rendimiento del sistema durante las trayectorias ejecutadas.

Los resultados obtenidos en los experimentos demostraron mejoras en la capacidad del sistema para aprender de manera continua la información visual del entorno, manteniendo al mismo tiempo el conocimiento previamente adquirido. Asimismo, se observaron avances en la velocidad de procesamiento y en los tiempos de inferencia para estimar la localización del dron. Si bien la precisión en la estimación de la posición presenta un compromiso esencial con la velocidad de operación, los hallazgos destacan la viabilidad de utilizar únicamente la percepción visual como una alternativa confiable de localización, especialmente en escenarios donde la señal GPS se pierde o es inestable.

Las principales contribuciones de esta investigación incluyen el desarrollo de una metodología de estimación de posición basada en estrategias de aprendizaje continuo, permitiendo al sistema adaptar su conocimiento visual durante la misión de vuelo. Además, se propone un esquema de múltiples modelos y sub-mapas que facilita el entrenamiento progresivo sin olvidar la información previamente adquirida. Adicionalmente, se exploraron otras arquitecturas de aprendizaje, validando la eficacia y velocidad del sistema en trayectorias aéreas reales donde la adquisición de la localización del dron es esencial. A través de estas contribuciones, esta tesis busca avanzar en el diseño de soluciones prácticas y eficientes para la localización visual de drones, brindando un sistema alternativo que garantice su operación en entornos desafiantes.

1.1. Motivación

En aplicaciones reales de la robótica aérea, uno de los principales factores que causan la pérdida de drones durante misiones de vuelo es la interrupción o pérdida total de la señal GPS. Esto provoca que el vehículo continúe navegando sin una referencia de su ubicación, aumentando el riesgo de pérdida o accidente. Dado que estas fallas pueden ocurrir en distintos escenarios, surge la necesidad de desarrollar sistemas de localización alternativos que permitan mantener el vuelo del dron incluso ante la ausencia de señal GPS y guiar al vehículo hacia zonas previamente conocidas (Figura 1.1).

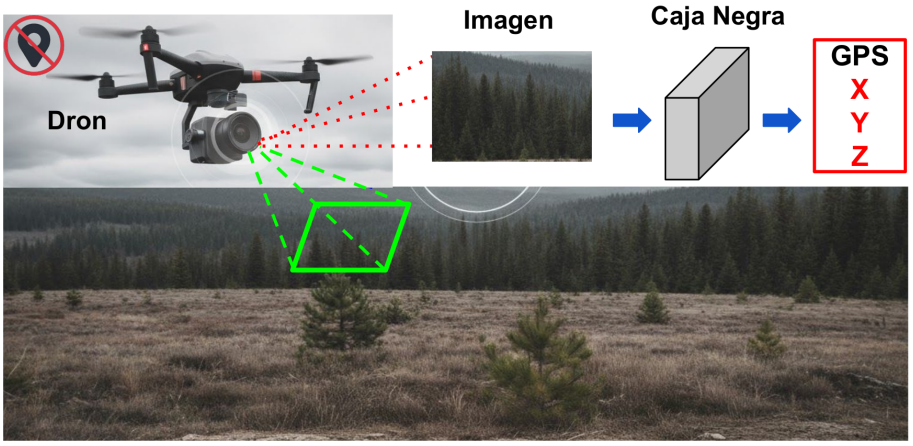

Figura 1.1: Representación del problema: cuando el dron pierde la señal GPS, una imagen aérea capturada por la cámara se procesa mediante un sistema (caja negra) para obtener una posición estimada, permitiendo continuar con la navegación.

La motivación principal de esta investigación radica en el uso de las cámaras monoculares para capturar información visual del entorno. A través de esta información, es posible extraer patrones visuales de forma similar a la percepción humana, permitiendo interpretar y comprender el escenario. Esta capacidad de reconocimiento basada en la visión plantea la oportunidad de utilizar imágenes aéreas no solo para estimar la posición de un dron, sino también para que el modelo aprenda y se adapte durante el vuelo utilizando esquemas de aprendizaje continuo.

Incorporar esta idea en un sistema de robótica aérea representa la oportunidad de crear un método de localización alternativo, capaz de operar en entornos desafiantes. Además, implementar un sistema alternativo podría mejorar la seguridad en las misiones aéreas y podría ser transferido a otras plataformas robóticas que requieran navegación autónoma en ausencia de señales satelitales.

1.2. Justificación

La estimación de la posición es un elemento clave en la autonomía de los sistemas robóticos aéreos como los drones. La posibilidad de obtener esta información con cámaras monoculares ha motivado a la comunidad científica a desarrollar sistemas de localización visual que permitan a los drones navegar en su entorno y ejecutar tareas de forma eficiente. Esto resulta relevante en escenarios donde tecnologías tradicionales, como el GPS, presentan limitaciones.

No obstante, los métodos actuales basados en aprendizaje profundo para estimar la posición utilizando imágenes aéreas enfrentan dificultades en entornos dinámicos o con frecuentes cambios tales como: aparición de nuevos objetos u obstáculos, cambios de iluminación, y modificaciones en el entorno que pueden afectar la precisión de los modelos entrenados previamente. Esto limita la capacidad de generalización y adaptabilidad durante misiones reales, donde se espera que el modelo de estimación continúe funcionando incluso ante condiciones inesperadas.

El uso de estrategias de aprendizaje continuo ofrece una solución a estos retos, permitiendo que el modelo se actualice progresivamente conforme el dron recorre nuevas zonas. De esta manera, se pueden diseñar sistemas más robustos que mantengan el conocimiento previamente adquirido y se adapten a escenarios cambiantes. Así, la implementación de un sistema de localización alternativo basado en visión, no solo incrementa la autonomía y seguridad de los drones, sino que representa una contribución significativa para el desarrollo de soluciones robustas en la robótica aérea.

1.3. Preguntas de Investigación e Hipótesis

La estimación de posición de una cámara monocular ha llevado al desarrollo de métodos de aprendizaje profundo utilizando redes neuronales convolucionales. Sin embargo, aún está lejos de ser autosuficiente ante los constantes cambios en el escenario, sin contar el alto consumo de recursos para obtener un modelo de entrenamiento. Por ello, formulamos las siguientes preguntas de investigación con sus respectivas hipótesis.

Pregunta de investigación 1:

¿Qué estrategia de aprendizaje continuo en conjunto con redes neuronales permite que una cámara pueda localizarse y el modelo resultante aprenda de manera dinámica a una velocidad cercana a la frecuencia de captura de la cámara?

Hipótesis 1:

La integración de estrategias de aprendizaje continuo basados en memorias externas junto con redes neuronales convolucionales, permite entrenar un modelo de localización de manera dinámica utilizando imágenes aéreas, manteniendo el conocimiento previamente adquirido mientras la cámara se desplaza a lo largo de una trayectoria con una velocidad de aprendizaje cercana a la frecuencia de captura.

Pregunta de investigación 2:

¿Qué tipo de esquema de localización puede implementarse a bordo de un dron para que, al recorrer una trayectoria, el sistema mantenga el aprendizaje previamente adquirido mientras incorpora de manera continua nueva información sin incurrir en el olvido?

Hipótesis 2:

La implementación de un esquema de múltiples modelos basado en aprendizaje continuo permite preservar el conocimiento previamente adquirido mientras se incorpora nueva información, facilitando la actualización progresiva de la localización aérea. Asimismo, el uso de una estrategia de sub-mapeo permite que el sistema de localización continúe aprendiendo nuevas posiciones durante el recorrido de una trayectoria en un entorno determinado.

1.4. Objetivo General

Desarrollar una metodología para estimar la posición de una cámara monocular utilizando imágenes aéreas y un esquema de aprendizaje continuo que permita localizar un dron a lo largo de una trayectoria, preservando el conocimiento previamente adquirido mientras se adquiere nueva información durante el recorrido.

1.4.1. Objetivos Específicos

1. Investigar arquitecturas de redes neuronales convolucionales para la estimación de posición a partir de imágenes aéreas.

2. Investigar estrategias de aprendizaje continuo aplicables al entrenamiento dinámico de modelos de estimación.

3. Proponer una metodología de estimación de posición que integre el aprendizaje continuo durante una misión de vuelo.

4. Proponer un esquema de localización aérea que permita actualizar el modelo de manera continua.

5. Investigar y proponer un esquema de localización aérea a bordo de un dron para estimar la posición de la cámara durante el desplazamiento del dron.

6. Evaluar la metodología mediante trayectorias aéreas y secuencias de imágenes capturadas en vuelo.

1.5. Fundamento Metodológico

Para validar las hipótesis de investigación y cumplir con los objetivos de esta tesis, se plantea la siguiente metodología enfocada en la estimación de la posición de una cámara monocular mediante aprendizaje continuo, así como el diseño de un sistema de localización para drones.

1. Revisar el estado del arte en localización visual, redes neuronales convolucionales y aprendizaje continuo.

2. Investigar arquitecturas de redes neuronales aplicadas a la estimación de posición a partir de imágenes aéreas.

3. Investigar enfoques de aprendizaje continuo que permitan actualizar modelos sin olvidar conocimientos previos.

4. Proponer un esquema de aprendizaje continuo basado en múltiples modelos para la estimación y actualización de la posición.

5. Proponer una estrategia de localización basada en submapas para facilitar el aprendizaje progresivo durante el vuelo.

6. Evaluar la metodología propuesta en trayectorias aéreas reales.

7. Desarrollar un sistema de localización visual alternativo para drones capaz de operar en escenarios desafiantes.

Asimismo, se presenta un diagrama general de la metodología propuesta para la localización visual mediante aprendizaje continuo, mostrando el flujo desde la entrada hasta la salida en la Figura 1.2.

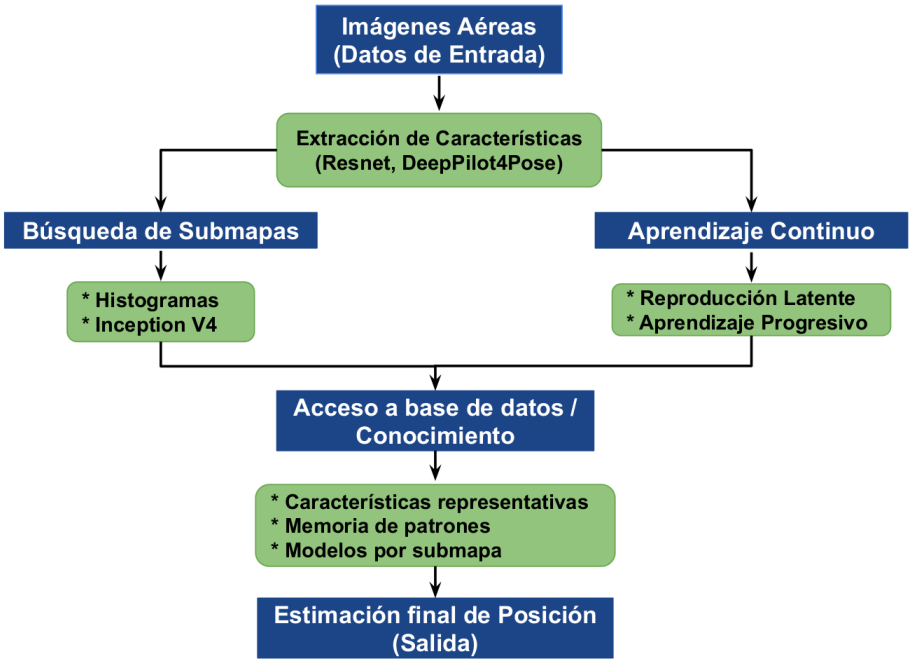

Figura 1.2: Diagrama general de la metodología propuesta para la localización visual, mostrando el flujo desde la entrada de imágenes aéreas, módulos de procesamiento, aprendizaje continuo y la estimación final de la posición.

1.5.1. Tipo de Investigación

La presente investigación adopta un enfoque cuantitativo, ya que se fundamenta en el análisis de datos numéricos mediante métricas, evaluaciones comparativas y resultados experimentales, con el propósito de validar las hipótesis planteadas. Este enfoque permite medir de forma objetiva el desempeño del sistema de localización propuesto, así como analizar su comportamiento bajo diferentes condiciones. Asimismo, la investigación se clasifica de tipo experimental, dado que se diseñan e implementan pruebas utilizando trayectorias de un dron, evaluando su desempeño en entornos reales (Pita Fernández & Pértegas Díaz, 2002; Rus Arias, 2021). Este tipo de investigación permite comprobar la viabilidad de la metodología propuesta, ofreciendo evidencia para sustentar las conclusiones de esta tesis.

1.5.2. Antecedentes

En investigaciones previas, se han explorado diversas arquitecturas de redes neuronales para abordar el problema de la estimación de posición a partir de imágenes aéreas. Estas redes, debido a su capacidad para aprender representaciones abstractas, permiten extraer características relevantes del entorno visual que pueden asociarse a coordenadas geográficas o espaciales. En una etapa inicial de esta línea de investigación, se desarrolló un sistema de geolocalización utilizando imágenes aéreas y una red neuronal convolucional, el cual mostró resultados prometedores en la predicción de posiciones geográficas a partir de una cámara monocular (Cabrera-Ponce & Martinez-Carranza, 2019).

Este avance motivó una exploración más profunda hacia el diseño e implementación de redes neuronales más compactas y eficientes, con el objetivo de reducir los tiempos de inferencia durante misiones reales. Esta adaptación se justifica en la necesidad de que un dron pueda estimar su localización de manera rápida, incluso en escenarios donde la señal GPS fuera limitada (Cabrera-Ponce et al., 2022). La experiencia adquirida a partir de estos trabajos confirma el potencial de las redes neuronales como herramientas efectivas para la localización visual. Sin embargo, también se encontró una limitación en la incapacidad de que los modelos puedan adaptarse dinámicamente a cambios en el entorno sin requerir un reentrenamiento completo.

A partir de esta necesidad, se identificó al aprendizaje continuo como una alternativa para enfrentar dicha limitación. Este enfoque busca que los modelos puedan incorporar nueva información de forma progresiva sin olvidar el conocimiento previamente adquirido. En nuestros primeros pasos de esta investigación, se exploró la integración del aprendizaje continuo con redes neuronales convolucionales en contextos de localización, permitiendo el desarrollo de esquemas capaces de actualizar su aprendizaje mientras estiman la posición de una cámara en entornos dinámicos (Cabrera-Ponce et al., 2021). Estos antecedentes son las bases para la presente investigación, orientada a combinar estrategias de aprendizaje continuo con información visual para desarrollar un sistema de localización aérea.

1.5.3. Herramientas Utilizadas

Para la recolección, procesamiento y análisis de los datos en esta investigación, se utilizaron las siguientes herramientas: Sistema Operativo de Robot (ROS) (Quigley et al., 2009), versión Noetic Kame, OpenCV (Bradski & Kaehler, 2000), CUDA 12.1, PyTorch 2.3.0. Estas herramientas nos permitieron procesar las imágenes aéreas para la extracción de características y análisis de datos visuales relevantes para el modelo de localización. Toda

la infraestructura experimental fue implementada en los laboratorios de drones y robótica del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), donde se contó con el equipo técnico y computacional necesario para el desarrollo del sistema propuesto. Además del uso y recursos del Laboratorio Nacional de Supercómputo del Sureste de México, de la Benemérita Universidad Autónoma de Puebla (BUAP).

1.5.4. Estancia

Como parte del desarrollo de esta investigación, se realizó una estancia de investigación en el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), llevada a cabo del 1 de febrero al 31 de julio de 2023. Esta estancia tuvo como propósito principal la ejecución de experimentos y validaciones del sistema de localización propuesto, aprovechando los recursos tecnológicos, el equipo y los espacios disponibles en los laboratorios del INAOE. Como resultado de esta estancia, se generaron avances que dieron lugar a la producción de artículos científicos relacionados con el trabajo presentado en esta tesis.

1.5.5. Estructura de la Tesis

La presente tesis se encuentra organizada en siete capítulos, comenzando con el Capítulo 1 donde se presentan los fundamentos de la investigación, la motivación, los objetivos, las hipótesis planteadas, y las contribuciones principales. El Capítulo 2 aborda los conceptos teóricos necesarios de aprendizaje profundo, redes neuronales convolucionales, aprendizaje continuo y otras arquitecturas utilizadas en esta investigación. El Capítulo 3 presenta la revisión de la literatura, analizando trabajos en localización visual, redes profundas y enfoques de aprendizaje continuo. El Capítulo 4 describe la metodología general de la tesis, así como las variables de estudio y recursos metodológicos. Los Capítulos del 5 al 7 describen las contribuciones implementadas, desde la localización topológica, localización jerárquica y localización progresiva, incluyendo los experimentos realizados, la integración de esquemas de múltiples modelos, submapas, así como otras arquitecturas evaluadas durante el trabajo. Finalmente, el Capítulo 8 presenta las conclusiones de la investigación, las limitaciones encontradas, el trabajo a futuro y posibles líneas de trabajo derivadas de esta investigación.

1.6. Contribuciones

Las principales contribuciones de esta tesis se resumen en los siguientes puntos:

1. Diseño de una metodología de estimación de posición basada en aprendizaje continuo, utilizando imágenes aéreas capturadas por cámaras monoculares en drones.

2. Propuesta de un sistema de localización aérea que integra redes neuronales convolucionales y estrategias de aprendizaje continuo, permitiendo al modelo adaptarse dinámicamente mientras el dron recorre nuevas trayectorias sin olvidar información previamente aprendida.

3. Desarrollo de un esquema de múltiples modelos con submapas, dividiendo el conocimiento para facilitar el entrenamiento progresivo del sistema de localización.

4. Implementación de técnicas de aprendizaje continuo con arquitecturas ligeras y modelos binarizados, orientadas a reducir los tiempos de entrenamiento e inferencia.

5. Revisión comprensiva de distintos métodos de entrenamiento incluyendo memorias externas, máquinas de vectores de soporte para regresión, aprendizaje incremental, y arquitecturas binarias para la estimación de posición de una cámara monocular.

6. Desarrollo de un sistema alternativo de localización visual para drones, proporcionando una solución para misiones de vuelo en las que la señal GPS sea inestable o inaccesible.

1.7. Publicaciones

Revistas indexadas:

1. **A. A. Cabrera-Ponce**, M. I. Martin-Ortiz, J. Martinez-Carranza. "Continual Learning via Multiple Support Vectors Models for Localisation with a Single Aerial Image". Unmanned Systems. 2025. Doi: https://doi.org/10.1142/S2301385026500172.

2. **A. A. Cabrera-Ponce**, M. I. Martin-Ortiz, J. Martinez-Carranza. "Continual Learning for Topological Geo-localisation", Journal of Intelligent & Fuzzy Systems. Pre-press, pp. 1-13, April 5, 2023. Doi: https://doi.org/10.3233/JIFS-223627.

Artículos indexados en otros índices:

1. **A. A. Cabrera-Ponce**, M. Martín-Ortiz, J. Martinez-Carranza. "Localización de una Cámara Monocular utilizando Métodos de Visión y Aprendizaje Profundo: Una Descripción General", United Academic Journals (UA Journals), 2022. Link: https://issuu.com/uajournals/docs/000015.

Capítulos de libros y conferencias internacionales:

1. **A. A. Cabrera-Ponce**, M. I. Martin-Ortiz, and J. Martinez-Carranza. "A Review on Binary Networks for 6D Aerial Pose Estimation", Handbook of Intelligent Robots: Theory, Methods and Applications, Chapter. Taylor & Francis CRC Press, 2025. (Aceptado)

2. **A. A. Cabrera-Ponce**, M. I. Martin-Ortiz, and J. Martinez-Carranza. "Continual Learning for Camera Localisation", Machine Learning for Complex and Unmanned Systems, Chapter. CRC Press, 2024. Doi: https://doi.org/10.1201/9781003385615.

3. **A. A. Cabrera-Ponce**, L. O. Rojas-Pérez, M. I. Martin-Ortiz, and J. Martinez-Carranza. "Binary networks and continual learning for pose estimation from a single aerial image," in 15^{th} international micro air vehicle conference and competition (IMAV2024). Bristol, United Kingdom, September, 2024. Link: https://www.imavs.org/papers/2024/11.pdf

4. **A. A. Cabrera-Ponce**, M. Martín-Ortiz, J. Martinez-Carranza. "Hierarchical Continual Learning for Single Image Aerial Localisation", in 14^{th} international micro air vehicle conference and competition (IMAV2023). Aachen, Germany, September 11-15, 2023. Link: https://www.imavs.org/papers/2023/5.pdf

5. **A. A. Cabrera-Ponce**, M. Martín-Ortiz, J. Martinez-Carranza. "Multi-model continual learning for camera localisation from aerial images," in 13^{th} international micro air vehicle conference (IMAV2022), Delft, the Netherlands, 2022, p. 103–109. Link: https://www.imavs.org/papers/2022/12.pdf

Artículos de divulgación y otros informes en eventos internacionales:

1. **A. A. Cabrera-Ponce**, M. Martín-Ortiz, J. Martinez-Carranza. "Aprendizaje profundo para localización aérea", Komputer Sapiens. Enero-Abril, 2025.

2. **A. A. Cabrera-Ponce**, M. Martín-Ortiz, J. Martinez-Carranza. "Discrete Hierarchical Continual Learning for Single View Geo-Localisation", Computer Vision and Pattern Recognition Conference: LatinX in AI (LXAI) Research Workshop (CVPR2023). Vancouver, Canada, 2023. Doi: https://doi.org/10.52591/lxai202306189.

Capítulo 2

Marco Teórico

Este capítulo establece el marco teórico para la estimación de posición basada en imágenes aéreas, intregrando conceptos clave de localización, aprendizaje profundo y estrategias de aprendizaje continuo. Esta base respalda el diseño de sistemas de localización para drones capaces de operar en entornos complejos y cambiantes. A través de la integración de redes neuronales convolucionales, modelos de soporte vectorial, redes binarias y equemas de entrenamiento progresivo, este marco proporciona los fundamentos necesarios para el desarrollo de una metodología de localización visual para drones.

2.1. Posición

La noción de posición es fundamental en tareas de navegación y localización en sistemas autónomos, especialmente utilizando drones. En términos generales, la posición se refiere a la ubicación de un objeto dentro de un espacio, expresado mediante un sistema de coordenadas. Esta puede clasificarse en dos tipos: posición absoluta y posición relativa. La posición absoluta indica la ubicación de un objeto con respecto a un marco de referencia fijo y global, como el sistema de posicionamiento global (GPS), mediante coordenadas geográficas (latitud y longitud). Por otra parte, la posición relativa describe la ubicación de un objeto respecto a otro dentro del mismo entorno, comúnmente utilizando sistemas de coordenadas que no dependen de referencias externas (Craig, 2005; Thrun et al., 2005).
En el contexto de la robótica y la visión por computadora, la posición alude a su ubicación en el espacio (posición, orientación tridimensional) respecto a un marco de referencia. Esta posición representa la traslación (x, y, z) y rotación (roll, pitch, yaw), lo cual permite describir su estado espacial (Barfoot, 2017). Determinar la posición de un dron es esencial para tareas de control, navegación, o interacción con el entorno, ya que permite estimar su ubicación

actual. A partir de ella, se desarrollan métodos que utilizan referencias externas como el GPS, sensores inerciales, y sensores visuales para inferir la posición relativa del dron dentro de un entorno. Esto es clave para explorar técnicas de localización visual y estimación de posición basadas en imágenes.

2.2. Localización

La localización es un proceso esencial en los sistemas autónomos, ya que permite determinar la ubicación de un objeto dentro de un espacio físico o virtual (Barfoot, 2017; Thrun et al., 2005). En general, implica la capacidad de encontrar y situar algo en una posición específica dentro de un marco de referencia. En robótica aérea, la localización se refiere al proceso mediante el cual un dron estima su posición y orientación dentro de un entorno, con el objetivo de navegar y ejecutar tareas de forma precisa. Este proceso se ha basado en sistemas de navegación como el GPS, ofreciendo una estimación absoluta de la ubicación del dron. No obstante, en espacios interiores se recurre a métodos basados en sensores visuales, como cámaras monoculares.

En este contexto, la localización visual permite estimar la posición de un dron a partir de la información contenida en imágenes (Scaramuzza & Fraundorfer, 2011). Este proceso se basa en extraer características del entorno mediante visión permitiendo estimar la posición relativa comparando con un marco de referencia local. Además la localización visual puede estimar la posición absoluta reconociendo ubicaciones dentro de un mapa global. De esta manera, la localización no se limita a conocer la posición actual, sino, entender su entorno, reconocer patrones y adaptar la información visual en función de su ubicación estimada.

2.3. Configuración del Hardware

El sistema aéreo utilizado en este trabajo está compuesto por el dron DJI Matrice 100, una plataforma diseñada para tareas de investigación, desarrollo y aplicaciones comerciales. Este vehículo aéreo no tripulado cuenta con una estructura que permite la integración de diferentes sensores, sistemas de navegación y cámaras. Entre sus características destacan la controladora de vuelo, un módulo GPS de alta precisión, una batería de larga duración, y un control remoto (Matrice, 2017). Además, incorpora seis grados de libertad, permitiendo movimientos sobre los ejes *x, y, z*, y rotaciones en *roll, pitch, yaw*, lo cual es fundamental para tareas de vuelo autónomo y adquisición de datos en entornos tridimensionales. En la Figura 2.1 se muestra una representación visual del sistema aéreo utilizado para la captura

de imágenes aéreas, incluyendo la cámara monocular y el sistema de posicionamiento GPS.

Figura 2.1: Sistema aéreo utilizado para la estimación de posición. Este dron cuenta con una controladora de vuelo, un módulo GPS y una cámara monocular de alta resolución.

2.3.1. Cámara Monocular

Una parte esencial del sistema aéreo es la cámara monocular, permitiendo capturar imágenes desde una única perspectiva a una resolución de 1280×720. Esta calidad de imagen representa un nivel de detalle suficiente para la extracción de características visuales, lo cual resulta clave para las tareas de localización visual. La cámara se encuentra montada de manera fija en la estructura del dron, garantizando la estabilidad en la captura de datos y facilitando la sincronización con los sistemas de navegación y registro de posición GPS. Además, las imágenes capturadas con la cámara son en escala de color permitiendo una mejor representación visual del entorno durante los vuelos del dron. Asimismo, la frecuencia de operación es de 24 Hz, 30 Hz y hasta 60 Hz en su versión configurable, permitiendo la adquisición de la información en tiempo real.

2.4. Posición GPS y Conversión de Coordenadas

La información de posición obtenida con el GPS del dron y la sincronización con las imágenes aéreas están en el estándar WGS84 (*World Geodetic System 1984*). Este sistema geodésico es utilizado en aplicaciones de navegación, cartografía y sistemas marítimos, ya que representa la posición sobre la Tierra mediante coordenadas angulares: latitud

y longitud en grados decimales (Agency, 1987; Hofmann-Wellenhof et al., 2012). Sin embargo, la precisión de este sistema puede verse afectada al perder un decimal, por lo que resulta conveniente trabajar en un sistema métrico para una mejor interpretación de la localización. Para ello, las coordenadas WGS84 suelen transformarse al sistema UTM (*Universal Transverse Mercator*), el cual divide la superficie terrestre en 60 zonas de 6° de longitud, permitiendo representar ubicaciones en términos de distancias lineales (Hofmann-Wellenhof et al., 2012). Las zonas UTM se identifican mediante un número (longitud) y una letra (latitud) y para este trabajo, las posiciones se encuentran dentro de la zona 14Q, correspondientes al hemisferio norte (Figura 2.2).

Figura 2.2: Zona UTM 14Q en el hemisferio norte. Este sistema facilita la representación de posiciones en unidades métricas para la estimación de posición a partir de imágenes aéreas.

La conversión de las coordenadas de WGS84 a UTM implica varios pasos. A continuación, se presentan algunas de las ecuaciones que permiten transformar coordenadas geográficas a coordenadas métricas en el sistema UTM:

$$N = \frac{a}{\sqrt{1 - e^2 \sin^2(\phi)}} \tag{2.1}$$

donde a es el semieje mayor del elipsoide WGS84 (6378137.0 m), y e es su excentricidad. Después de determinar la zona UTM a la que pertenece la posición, se calcula la diferencia de longitud con respecto al meridiano central de la zona:

$$\lambda_0 = (\text{Zona} - 1) \cdot 6 - 180 + 3 \tag{2.2}$$

Después se convierten latitud y longitud a radianes, y se calculan los parámetros intermedios donde el meridiano es proyectado a una distancia aproximada desde el ecuador:

$$M = a\left[(1 - \frac{e^2}{4} - \frac{3e^4}{64} - \frac{5e^6}{256})\phi - (\frac{3e^2}{8} + \frac{3e^4}{32} + \frac{45e^6}{1024})\sin(2\phi) + ...\right] \quad (2.3)$$

Finalmente, se obtienen las coordenadas UTM (Este "E", Norte "N"), donde $A = (\lambda - \lambda_0) \cdot cos(\phi), k_0$ es el factor de escala (usualmente 0.9996), C y T son parámetros intermedios que dependen de la latitud.

$$E = k_0 N\left[A + \frac{(1 - T + C)A^3}{6} + \frac{(5 - 18T + T^2 + 72C - 58e^2)A^5}{120}\right] + 500000 \quad (2.4)$$

$$N = k_0\left[M + N\tan(\phi)\left(\frac{A^2}{2} + \frac{(5 - T + 9C + 4C^2)A^4}{24} + ...\right)\right] \quad (2.5)$$

Este proceso de conversión se implementa automáticamente en Python utilizando librerias de sistemas de información geográfica (GIS). Adicionalmente, se define como origen del sistema métrico la primera posición registrada en cada trayectoria de vuelo del dron. Por ejemplo, el punto inicial es definido como $(0, 0)$, correspondientes a una posición en UTM $(E : 572082.994, N : 2104515.563)$. Es importante destacar que el valor de altura utilizado como componente z de la posición representa a la información del altímetro del dron, obteniendo una medida más confiable. De esta forma, cada imagen capturada es asociada a coordenadas (x, y, z), donde x y y representan a las componentes UTM y z representa la altura relativa del dron respecto al suelo.

2.5. Aprendizaje Profundo

El aprendizaje profundo es una rama del aprendizaje automático que utiliza redes neuronales con múltiples capas para aprender representaciones jerárquicas de los datos (Goodfellow et al., 2016; LeCun et al., 2015). Estas representaciones permiten modelar relaciones complejas y patrones abstractos a partir de información visual, textual o de señales. Dentro de esta rama existen diversas arquitecturas entre las que destacan las redes neuronales convolucionales (CNN, por sus siglas en inglés) para procesamiento de imágenes y redes neuronales recurrentes (RNN, por sus siglas en inglés) para secuencias temporales. De esta manera, el aprendizaje profundo tiene la capacidad de extraer características de los

datos de entrada mediante capas ocultas, funciones de activación no lineales y procesos de optimización. Esto crea modelos capaces de generalizar el conocimiento adquirido durante un proceso de entrenamiento para tareas como clasificación, segmentación o estimación.

2.5.1. Redes Neuronales Convolucionales

Las redes neuronales profundas se componen de múltiples capas ocultas las cuales permiten extraer representaciones jerárquicas de los datos de entrada. Esta información pasa de capa en capa donde las características se transforman desde patrones simples hasta representaciones más abstractas y complejas (Goodfellow et al., 2016; LeCun et al., 2015). Una de las arquitecturas más utilizadas para tareas de percepción visual son las redes neuronales convolucionales donde los datos se interpretan en forma de tensores.

Las redes convolucionales están diseñadas específicamente para procesar imágenes mediante operaciones de convolución utilizando filtros que detectan patrones tales como: bordes, texturas, colores y formas. Esta estructura aprovecha la correlación espacial de los píxeles, reduciendo el número de parámetros y mejorando la capacidad de generalización. Así mismo genera mapas de características que representan la información detectada de los patrones locales dentro de las imágenes (Krizhevsky et al., 2012).

Generalmente las redes convolucionales suelen combinarse con capas de activación no lineales tipo ReLU, capas de agrupación y capas totalmente conectadas para integrar la información extraída. Además, se emplean capas de normalización y técnicas de regularización para evitar el sobreajuste, manteniendo la capacidad de generalizar a nuevos datos. Esta combinación de capas convierte a estas arquitecturas en una herramienta clave para la extracción de características visuales, proporcionando la base para modelos de regresión y clasificación.

2.5.2. Entrenamiento

El entrenamiento de una red neuronal convolucional consiste en ajustar los parámetros internos para minimizar una función de pérdida que establece la diferencia entre las predicciones del modelo y valores reales. Este ajuste de parámetros se realiza mediante retropropagación del error y métodos de optimización como el gradiente descendente estocástico (Rumelhart et al., 1986), permitiendo encontrar mínimos locales que reduzcan el error del aprendizaje. El proceso de entrenamiento puede realizarse desde cero el cual implica inicializar la red con pesos aleatorios utilizando un conjunto de datos grande. Por otro lado, el entrenamiento por transferencia permite reutilizar redes previamente entrenadas en dominios similares ajustando las capas finales para adaptarlas a una nueva tarea.

Durante el entrenamiento es necesario controlar el ajuste de los pesos del modelo y la reducción del error utilizando parámetros como la tasa de aprendizaje la cual permite optimizar la búsqueda de mínimos locales. Esto evita el sobreajuste, reduce el costo computacional y acelera la convergencia en el entrenamiento de la red. Para tareas de estimación el entrenamiento de una CNN permite mapear patrones visuales extraídos de las imágenes hacia valores continuos, convirtiéndose en modelos de regresión capaces de aproximar coordenadas espaciales. Al finalizar el entrenamiento la relación entre los valores reales y los estimados proporciona una medida de la precisión del modelo, siendo un valor alto de buen desempeño mientras que uno bajo indica un entrenamiento erróneo.

2.6. Aprendizaje Continuo

El aprendizaje continuo es una rama del aprendizaje profundo inspirada en la neurociencia cuyo objetivo es permitir que un sistema aprenda de manera progresiva a partir de nuevas experiencias sin olvidar el conocimiento previamente adquirido. A diferencia del aprendizaje profundo tradicional, el aprendizaje continuo busca emular la capacidad del cerebro humano para integrar información de manera incremental (Parisi et al., 2019). Sin embargo, este enfoque enfrenta uno de los principales retos en los sistemas neuronales artificiales y biológicos: el dilema de plasticidad y estabilidad el cual consiste en equilibrar la adaptación a nuevos datos sin comprometer el conocimiento anterior.

Este dilema establece que el aprendizaje requiere plasticidad para incorporar nuevo conocimiento pero también estabilidad para prevenir el olvido del anterior (Mermillod et al., 2013). Por un lado, si un modelo conserva demasiada plasticidad puede adaptarse rápidamente a datos recientes pero con el riesgo de sufrir olvido catastrófico. Por el contrario, un modelo demasiado estable retiene el conocimiento anterior pero con limitaciones para integrar nuevas tareas. El aprendizaje continuo busca equilibrar estas dos propiedades mediante mecanismos que permitan a los modelos actualizarse de manera progresiva sin comprometer la información ya aprendida.

Para alcanzar este equilibrio, se han desarrollado diversas estrategias que permiten integrar nueva información mientras se preserva la ya adquirida. Entre las más utilizadas se encuentran los métodos de regularización que limitan modificaciones en los parámetros del modelo. Uso de memorias episódicas que almacenan ejemplos representativos de experiencias pasadas, y arquitecturas dinámicas que adaptan la estructura de la red en función del aprendizaje. Estas estrategias permiten optimizar los pesos del modelo y regular la función de pérdida, controlando la interferencia entre datos nuevos y anteriores sin necesidad de reentrenar desde cero, siendo útil en robótica donde la recolección de datos es continua.

2.6.1. Estrategias de Aprendizaje Continuo

En el aprendizaje continuo una de las estrategias más comunes es la regularización de parámetros el cual consiste en modificar la función de pérdida añadiendo penalizaciones cuando se producen cambios bruscos en los pesos. El objetivo principal es preservar el conocimiento adquirido previamente mientras se incorporan nuevos datos, evitando que el modelo olvide lo ya aprendido. Un enfoque representativo de esta técnica es utilizar una matriz que identifica los pesos más relevantes, permitiendo rastrear los parámetros esenciales durante el entrenamiento (Kirkpatrick et al., 2017). Con ello, los ajustes en los pesos se valoran en función del impacto en la salida del modelo, manteniendo la estabilidad del conocimiento previo. No obstante, este tipo de estrategias puede presentar limitaciones cuando se trabaja con secuencias muy extensas o altamente discontinuas.

Otra estrategia utilizada en aprendizaje continuo es la basada en memoria episódica, el cual reutiliza información pasada almacenada en memorias externas para reforzar el aprendizaje del modelo. Este enfoque conserva un conjunto reducido de muestras representativas de experiencias anteriores y las combina con los nuevos datos durante el entrenamiento, permitiendo que el modelo mantenga el conocimiento previo mientras incorpora información reciente. Un método representativo es la repetición latente de los datos propuesto en (Lomonaco & Maltoni, 2017), donde almacena representaciones latentes generadas por el modelo. Matemáticamente, si $z_i = f_0(x_i)$ representa la proyección latente de una muestra x_i a través de la red f_0, el conjunto de memoria se define como $M = \{z_1, z_2, ..., z_k\}$. Durante el entrenamiento con nuevos datos x_n las representaciones almacenadas en M se mezclan con las nuevas z_n, actualizando los pesos de manera que se refuercen las conexiones previas:

$$\mathcal{L}_{total} = \mathcal{L}(z_n) + \lambda \mathcal{L}(M) \tag{2.6}$$

donde λ controla la importancia de la memoria pasada frente a los datos nuevos. El método demuestra la capacidad de rejuvenecer los pesos del modelo, reforzando el conocimiento adquirido y reduciendo el olvido. Además, es útil en robótica y visión por computadora ya que disminuye el uso de memoria al almacenar representaciones compactas y acelera el entrenamiento en tareas de estimación.

Finalmente, las estrategias basadas en arquitectura se centran en adaptar la estructura del modelo para integrar nuevo conocimiento sin comprometer el ya adquirido. En contraste con la repetición latente, este enfoque modifica la red al expandir su arquitectura, incorporando módulos para nuevas tareas. Un método representativo de esta estrategia son las redes neuronales progresivas las cuales congelan los modelos entrenados y conectan nuevas redes para aprender tareas futuras (Rusu et al., 2016). De manera similar, esquemas maestro-

estudiante permiten que los modelos repliquen el comportamiento de las redes sin necesidad de acceder a los datos originales del entrenamiento.
En esta tesis se explora una variante de este enfoque mediante la construcción de múltiples modelos especializados en lugar de modificar directamente la arquitectura. Este esquema facilita la adaptación incremental del entorno y reduce la interferencia entre conocimientos, resultando útil para aplicaciones de localización aérea.

2.7. Maquinas de Soporte Vectorial (SVM)

Las Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) son algoritmos de aprendizaje supervisado utilizados en tareas de clasificación binaria. Su objetivo principal es encontrar un hiperplano que separe los datos en distintas clases con un margen entre ellas (Cortes & Vapnik, 1995). Este margen representa la distancia entre el hiperplano y los puntos más cercano de cada clase, denominados vectores de soporte, los cuales definen la solución del modelo. De esta manera, dado un conjunto de datos etiquetados (x_i, y_i) donde $x_i \in \mathbb{R}^n$ y $y_i \in \{-1, +1\}$ el objetivo es encontrar un hiperplano:

$$w^T x + b = 0 \tag{2.7}$$

tal que:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 \textit{ sujeto a } \quad y_i(w^T x_i + b) \geq 1 \tag{2.8}$$

Esto permite maximizar el margen y garantizar la correcta separación de las clases, especialmente con el uso de datos no linealmente separables. Además, se introducen términos de suavizado y funciones kernel que proyectan los datos en espacios de mayor dimensión, haciendo que la separación sea posible. De esta manera, el entrenamiento de las SVM requiere de vectores de soporte para realizar predicciones nuevas, donde el modelo toma decisiones basado en la similitud de un nuevo dato con los vectores de soporte a través de su función kernel $K(x_i, x)$:

$$f(x) = \text{sign}\left(\sum_{i=1}^{Nsv} \alpha_i y_i K(x_i, x) + b\right) \tag{2.9}$$

donde α_i son los coeficientes de Lagrange asociados a los vectores de soporte x_i y $N_s v$ es el número total de los vectores de soporte. Además de su aplicación en clasificación

binaria, las SVM pueden adaptarse a tareas de regresión, conocidas como Máquinas de Soporte Vectorial para Regresión (SVR, por sus siglas en inglés). Las SVR son adecuadas para minimizar la suma del error entre las predicciones y los valores reales, controlando al mismo tiempo la complejidad del modelo. Además permiten manejar datos con relaciones lineales como no lineales entre las variables de entrada y su salida.

2.7.1. Máquinas de Soporte Vectorial para Regresión (SVR)

Las SVR son una extensión de las SVM orientada a problemas de regresión, donde el objetivo no es clasificar datos, sino estimar un valor continuo a partir de los datos aprendidos (Drucker et al., 1996). Este método es útil en tareas donde las salidas corresponden a variables continuas como la estimación de posición y localización visual, ya que permite predecir coordenadas espaciales que varían de forma continua. El objetivo de la SVR es encontrar una función $f(x)$ que mantenga las predicciones dentro de una desviación de un valor ε respecto a los valores reales y_i:

$$\min_{w,b,\xi,\xi^*} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.10}$$

sujeto a:

$$\begin{aligned} y_i - w^T x_i - b &\leq \varepsilon + \xi_i \\ w^T x_i + b - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned} \tag{2.11}$$

donde ε define una zona de tolerancia dentro de la cual las predicciones no se penalizan, y C es un parámetro que controla el equilibrio entre la complejidad del modelo y el error de predicción. Del mismo modo que las SVM, las predicciones se basan únicamente en los vectores de soporte, pero en este caso corresponden a los puntos que se encuentran fuera del margen ε. Estos puntos son aquellos que afectan a la función de regresión y cuya predicción de un nuevo dato se calcula de la siguiente manera:

$$f(x) = \sum_{i=1}^{Nsv}(\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{2.12}$$

donde α_i y α_i^* representan la influencia de cada vector de soporte sobre la predicción final. En esta tesis se explora utilizar las SVR para estimar coordenadas espaciales a partir de características visuales extraídas de imágenes aéreas, aproximando funciones y generando estimaciones numéricas cercanas a los valores reales. Esta capacidad de predicción las

convierten en un método útil para problemas de localización, ya que permiten interpretar los ejemplos que influyen en cada predicción. Además, puede ser llevado a su integración en esquemas de aprendizaje continuo mediante el uso de memorias externas o estrategias progresivas, extendiendo su uso en escenarios dinámicos de estimación de posición.

2.8. Redes Neuronales Binarias (BNN)

Las redes neuronales binarias (BNN, por sus siglas en inglés), son un tipo de red neuronal que surge como alternativa frente a las crecientes demandas computacionales y de memoria de los modelos de aprendizaje profundo. Este tipo de redes son diseñadas para arquitecturas de gran tamaño como los modelos de lenguaje, reduciendo la precisión numérica de los parámetros del modelo de valores flotantes a representaciones binarias. Así, el uso de representaciones binarias permiten optimizar el consumo de memoria y acelerar los cálculos en el proceso de entrenamiento de la red.

La conversión que se realiza en las redes binarias consiste en reemplazar los valores de punto flotante de 32 bits por representaciones binarias de 1 bit con rangos únicamente de $[-1, +1]$ (Courbariaux et al., 2016). Este proceso conocido como *binarización*, transforma los pesos y las activaciones sustituyendo las operaciones de multiplicación por sumas y restas, similar a operaciones lógicas como XNOR. De esta manera, se reduce la complejidad computacional y el número de operaciones durante el entrenamiento (Ma et al., 2024; Wang et al., 2023). Dado que las operaciones matriciales se implementan a cálculos XNOR, podemos utilizar conteo de bits para sustituir el costoso cálculo de multiplicación de matrices.

En las redes convolucionales la carga computacional proviene de las multiplicaciones de matrices durante la operación de convolución expresada como $Z = W * I$ donde W e I representan los pesos y las activaciones. Estas operaciones requieren grandes volúmenes de cálculos en 32 bits, afectando al rendimiento de la inferencia y alargando el proceso de entrenamiento. En contraste, las redes binarias aplican la binarización sobre W e I antes de la operación de convolución, reemplazando estas multiplicaciones por operaciones bit a bit mucho más ligeras. Para lograr esto se utiliza la función *sign* haciendo que los pesos binarizados sean cercanos a un valor 0 acelerando así las operaciones. La Figura 2.3 ilustra este proceso mostrando cómo las representaciones de punto flotante se convierten en representaciones de 1 bit con valores 0 y 1, haciendo un recuento de la suma obtenida y optimizando el procesamiento sin comprometer la estructura de la red.

En términos matemáticos la binarización de los pesos se realiza mediante una función de signo ajustado por un umbral α que representa el valor promedio de los pesos. Después se convierte cada peso $Wij \in \{-1, +1\}$ en valores binarios $\tilde{W}ij \in \{-1, +1\}$ definido como:

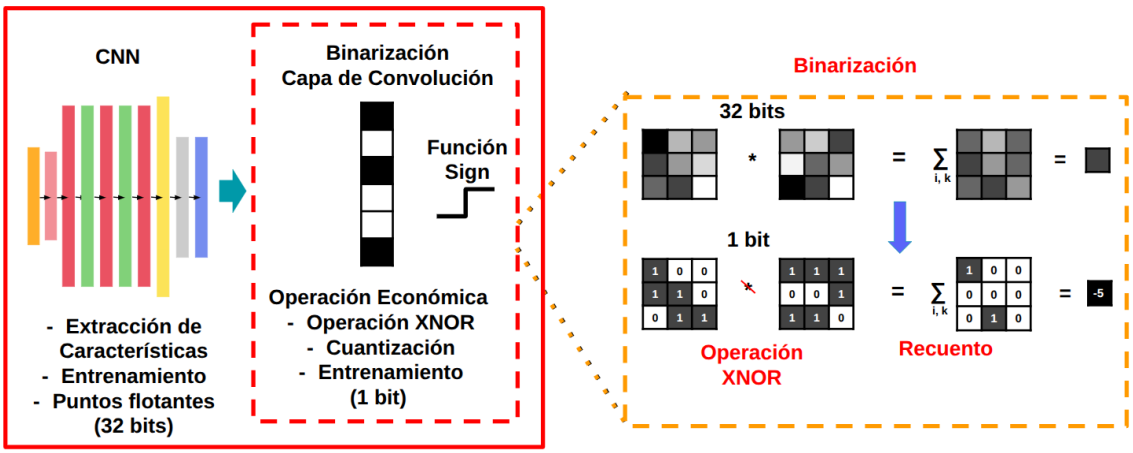

Figura 2.3: Proceso de binarización en una red convolucional. Las representaciones en los parámetros de los pesos se reducen a valores de 1 bit con valores 0 y 1 en lugar de representaciones abstractos de punto flotante de 32 bits.

$$\alpha = \frac{1}{nm} \sum_{i,j} W_{ij} \tag{2.13}$$

$$\tilde{W}ij = \text{Sign}(Wij - \alpha) = \begin{cases} +1, & \text{si } W_{ij} > \alpha - 1, \\ \text{en otro caso} \end{cases} \tag{2.14}$$

No obstante, el proceso de binarización introduce una reducción en la precisión del aprendizaje, ya que los pesos binarizados no pueden optimizarse de manera directa utilizando el descenso de gradiente tradicional. Para resolver este problema, se aplica la técnica del Estimador Directo (STE, por sus siglas en inglés), que permite retropropagar pesos binarios a través de la función no diferenciable *Sign*. Con este método, los pesos reales de cada capa se mantienen durante el entrenamiento y se actualizan utilizando el estimador, preservando así una aproximación del gradiente original. Una vez finalizado el entrenamiento los pesos binarizados reemplazan a los pesos reales y son descartados. En cuanto a las activaciones, se utiliza un proceso de cuantización que normaliza los valores utilizando su norma infinita, ajustando el rango permitido al número de bits disponibles. Este proceso se define como:

$$\tilde{x} = \text{Clip}\left(x \cdot \frac{Q_b}{\gamma}, -Q_b + \epsilon, Q_b - \epsilon\right) \tag{2.15}$$

$$\text{Clip}(x, a, b) = \text{máx}(a, \text{mín}(b, x)), \quad \gamma = |x|_\infty \tag{2.16}$$

donde $Q_b = 2^b - 1$ representa el rango máximo de valores para b bits y $\epsilon = 1 \times 10^{-5}$ es un valor pequeño utilizado para evitar saturaciones numéricas. Esta cuantización permite mantener

una representación compacta y controlada de las activaciones, siendo adecuada para operaciones binarizadas. Con pesos y activaciones binarizadas, la operación de convolución en una capa puede expresarse de manera eficiente mediante operaciones bit a bit en lugar de multiplicaciones en punto flotante expresa como:

$$y = \tilde{W}\tilde{x} \tag{2.17}$$

donde y representa la salida de la capa calculada mediante operaciones bit a bit optimizadas. En esencia, la binarización y cuantización en redes neuronales reduce el uso de memoria, acelera tanto el proceso de entrenamiento como la inferencia y mejora la compatibilidad con arquitecturas tradicionales. Aunque las redes binarias se introdujeron inicialmente para tareas de procesamiento de lenguaje natural, su aplicabilidad a redes convolucionales ha demostrado un potencial en áreas de visión por computadora. Este avance abre la posibilidad de extender su uso hacia problemas de regresión y estimación de posición, donde la combinación de convoluciones binarizadas con salidas continuas resulta útil para la localización. Por lo tanto, en esta tesis se explora esta capacidad para la localización visual en escenarios reales, buscando un balance entre la precisión y la eficiencia del modelo.

2.9. Localización Aérea con Modelos de Aprendizaje

Los métodos de aprendizaje permiten desarrollar modelos de estimación capaces de predecir valores numéricos cercanos a los reales a partir de los datos aprendidos. En este contexto, la estimación de posición utilizando imágenes aéreas se ha convertido en un proceso esencial para aplicaciones en robótica y visión por computadora. Para este tipo de tareas, modelos basados en redes neuronales convolucionales, aprendizaje continuo, redes binarias y máquinas de soporte vectorial para regresión han demostrado ser adecuados. Su objetivo principal es mapear las características visuales extraídas de una imagen hacia una posición específica dentro de un entorno definido.

El proceso comienza con la extracción de características visuales a partir de imágenes utilizando redes convolucionales, las cuales identifican patrones complejos útiles para la estimación de posición. Después, estas características se asocian con coordenadas tridimensionales mediante modelos de regresión que generan predicciones basadas en ejemplos previamente aprendidos. Este proceso permite adaptar los modelos a nueva información, facilitando un entrenamiento progresivo que se actualiza continuamente con datos recientes. Además, la extracción de información visual permite establecer una relación entre la posición estimada y la perspectiva de la imagen, lo que permite la adaptabilidad

del sistema de localización y su implementación en diversas plataformas.
Este conjunto de técnicas ofrece múltiples ventajas tales como: eficiencia computacional en entornos con restricciones, procesamiento optimizado en plataformas con recursos limitados, inferencias rápidas, capacidad de adaptación a nueva información y entrenamiento con datos recientes. En resumen, el uso de modelos de aprendizaje para la estimación de posición permite determinar la localización de una cámara en un vehículo aéreo a partir de una sola imagen, abriendo posibilidades en navegación autónoma, mapeo aéreo, seguimiento de trayectorias y planificación, sin depender de sistemas de localización externos.

2.10. Sumario

El marco teórico reúne los principios clave que sustentan la investigación en estimación de posición utilizando cámaras monoculares mediante técnicas de aprendizaje. Así, se abordan los conceptos esenciales de localización, posición, configuración del hardware y el uso de coordenadas GPS. También, se introducen las bases del aprendizaje profundo y aprendizaje continuo enfocados en arquitecturas de redes neuronales convolucionales (CNN) para la extracción de características visuales. Conceptos complementarios en métodos de aprendizaje como las Máquinas de Soporte Vectorial para Regresión (SVR) y las Redes Neuronales Binarias (BNN) son integrados también.
De esta manera, este capítulo establece el sustento teórico necesario para comprender y diseñar un sistema de estimación de posición a partir de imágenes aéreas, integrando modelos de aprendizaje con distintos grados de precisión, adaptabilidad y eficiencia. La estructura del capítulo presenta los elementos técnicos y conceptuales para el desarrollo de esta investigación aplicable a la localización aérea utilizando imágenes capturadas con la cámara de un dron.

Capítulo 3

Estado del arte

La estimación de posición representa un importante área en visión por computadora, impulsada por la necesidad de obtener localizaciones a partir de imágenes capturadas por cámaras monoculares. En tareas de robótica aérea tales como navegación y vuelo autónomo con drones presenta un desafío que exige algoritmos de visión robustos capaces de operar en entornos complejos. Este capítulo presenta una revisión de los avances más recientes en estimación de posición y localización visual aplicada a drones, abarcando enfoques visuales tradicionales, métodos de aprendizaje profundo, aprendizaje continuo y otras técnicas de aprendizaje orientadas a la estimación de posición con cámaras monoculares.

3.1. Localización Visual

La localización visual utilizando imágenes aéreas capturadas con cámaras monoculares es un reto en robótica ya que requiere estimar la posición a partir de información visual. Este proceso consiste en extraer patrones y características representativas de la imagen para vincularlos a un marco de referencia espacial mediante correspondencia o regresión. A lo largo de los años, se han propuesto enfoques que abarcan diferentes métodos clásicos basados en características utilizados para conocer el entorno mediante representaciones visuales a coordenadas de posición. Por ejemplo, vecinos más cercanos sobre descriptores visuales han demostrado que es posible estimar la ubicación aprovechando las características visuales en escenas extensas (Meng et al., 2016).

Debido a que los datos influyen directamente a la estimación de posición, algunos trabajos exploran el uso de sensores como cámaras RGB (Li-Chee-Ming & Armenakis, 2018), cámaras infrarrojas (Su et al., 2017), sensores de profundidad, unidades de medición inercial (IMU, por sus siglas en inglés) y sensores láser (LiDAR) (Carrasco et al., 2021). Estos enfoques,

conocidos como fusión de múltiples sensores permiten combinar información complementaria para tareas de localización, mapeo, escaneo y estimación de posición en diferentes entornos (Abdi et al., 2016). Sin embargo, la integración de estos sensores presenta otros retos relacionados con la calibración, el movimiento, la sincronización y el desenfoque. Por ello, las cámaras son más ideales para trabajar con características y descriptores visuales ya que ofrecen una estructura de datos más fácil de interpretar.

Dado el rol central de las cámaras en la captura de información visual, un método popular consiste en la extracción y comparación de descriptores visuales. Estos descriptores se utilizan en tareas de localización y reconocimiento de lugares gracias a su invariancia a rotaciones y cambios de escala, haciéndolos robustos al trabajar con nube de puntos o imágenes con ruido en anotaciones (Chathuranga & Munasinghe, 2019; Wong et al., 2017). Entre los métodos populares está el emparejamiento de características ya que permite estimar la posición de una imagen de consulta al identificar similitudes con descriptores visuales conocidos (Figura 3.1). Otros trabajos extienden este método hacía vistas cruzadas utilizando imágenes aéreas y satelitales, encontrando correspondencias de posiciones en diferentes perspectivas y múltiples vistas (Dusmanu et al., 2021; Shetty & Gao, 2019). Sin embargo, el emparejamiento y la coincidencia de características requiere un proceso largo y costoso, ya que requieren cálculos adicionales para la comparación y su representación a coordenadas de posición.

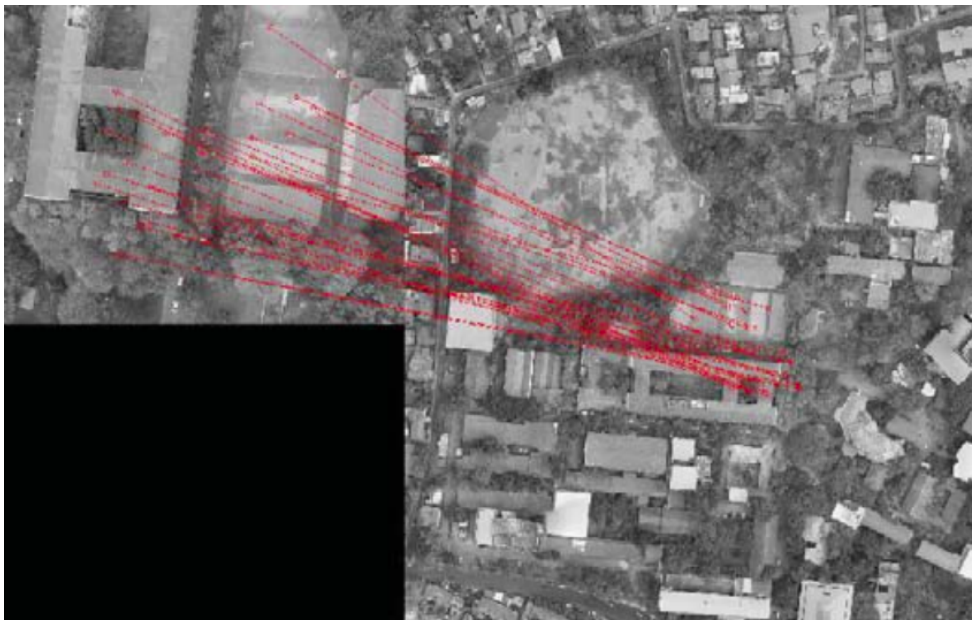

Figura 3.1: Emparejamiento de características de una imagen de consulta con una de referencia (Chathuranga & Munasinghe, 2019).

Otros trabajos han propuesto técnicas que se apoyan en vistas superpuestas y comparación

de mosaicos para estimar la posición global de un dron dentro del entorno (Samano et al., 2021; Zhao et al., 2017). Complementando el uso de descriptores visuales, técnicas más avanzadas llamadas sistemas de localización y mapeo simultáneo (SLAM, por sus siglas en inglés) y odometría visual, combinan la información visual con mapas 3D incluyendo datos LiDAR (Caselitz et al., 2016). Estos métodos permiten construir un mapa del escenario a partir de la información visual capturada con la cámara, determinando así la posición y orientación del sistema dentro del entorno. Entre los sistemas de mapeo más conocidos, ORB-SLAM2 (Mur-Artal & Tardós, 2017) se ha consolidado como una de las soluciones más utilizados en robótica y visión por computadora, al utilizar descriptores visuales para crear mapas capaces de funcionar en entornos tanto interiores como exteriores. El objetivo de los sistemas SLAM es construir y actualizar un mapa global del entorno mientras se estima la posición de la cámara.

A diferencia de estos métodos, la odometría visual calcula el desplazamiento relativo entre imágenes consecutivas sin necesidad de mantener un mapa completo. De esta manera, el método aprovecha los descriptores visuales para estimar el movimiento de la cámara y generar trayectorias locales, reduciendo la acumulación de error comúnmente encontrado en mapas extensos. Algunos trabajos han explorado la combinación de odometría visual con información LiDAR para mejorar la interpretación del escenario en entornos complejos y la estimación de posición (Chow et al., 2019; Qian et al., 2021; Yang et al., 2021). Con el uso de la odometría visual es posible obtener la localización de una cámara de manera rápida en múltiples escenarios incluyendo ambientes con sistemas GPS negado y exteriores (Mascaro et al., 2018).

En contraste con la odometría visual, el cual utiliza porciones del mapa para minimizar el error, los sistemas SLAM generan representaciones globales más completas del entorno. Esto ofrece resultados favorables en distintos escenarios gracias a los descriptores visuales y a su correspondencia con puntos 3D dentro del mapa. Por ejemplo, los sistemas SLAM son utilizados ampliamente en tareas de inspección donde permiten no solo estimar la localización sino planificar trayectorias para navegación autónoma (Benjumea et al., 2021). De igual forma, su aplicación en entornos interiores ha demostrado ser efectivo para recuperar la posición y determinar la localización en escenarios limitados (Martínez-Carranza et al., 2016). Por otra parte, desarrollos más recientes extienden este método hacía entorno densos y de múltiples escalas, mejorando la interpretabilidad de la escena mediante algoritmos de perspectiva-n-punto (PnP, por sus siglas en inglés) y consenso de muestras aleatorias (RANSAC, por sus siglas en inglés) para estimar la posición de la cámara (Rabiee & Biswas, 2021; Tang et al., 2021). Un ejemplo de SLAM se presenta en la Figura 3.2, donde se extraen descriptores visuales de la imagen para generar la trayectoria de una cámara.

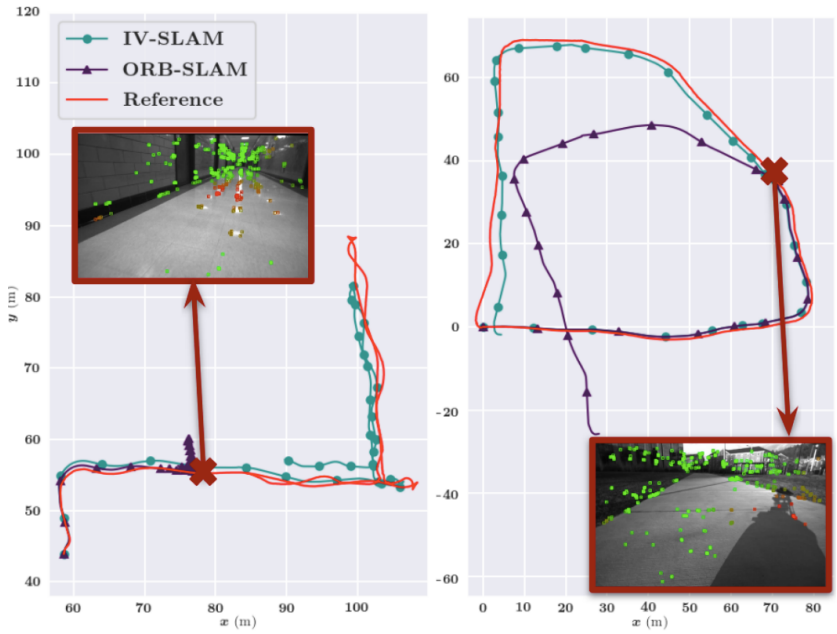

Figura 3.2: Ejemplo de un sistema SLAM siguiendo la referencia de trayectoria de una cámara (Rabiee & Biswas, 2021).

Con los trabajos previamente revisados, encontramos que los métodos se enfocan en la extracción de descriptores y su uso en técnicas clásicas de visión por computadora para estimar la posición de la cámara. Estos trabajos demuestran ser efectivos en múltiples escenarios gracias a la información visual que se obtiene de las imágenes. Sin embargo, cuando se enfrentan a entornos no estructurados y con baja textura tienden a fallar en su rendimiento debido a la dificultad para extraer información visual suficiente para la localización. Ante estas limitaciones, han surgido metodologías más recientes basadas en aprendizaje profundo utilizando redes neuronales convolucionales para aprender representaciones abstractas de las escenas. Estas técnicas demuestran una mayor capacidad para interpretar escenarios complejos y generalizar en diferentes condiciones, siendo ideales para tareas de localización y estimación de posición en aplicaciones de robótica aérea.

3.2. Aprendizaje Profundo para Estimación de Posición

En la última década, los métodos de aprendizaje profundo se han popularizado en áreas de visión por computadora y robótica gracias al poder computacional de las tarjetas gráficas y la eficacia de las redes neuronales convolucionales (CNN, por sus siglas en inglés) para extraer características robustas. Estas redes permiten aprender representaciones abstractas

y de alto nivel a partir de conjuntos de datos de imágenes, logrando generalizar y estimar valores continuos cercanos a los valores de referencia. De esta manera, uno de los trabajos pioneros en la estimación de posición de una cámara es PoseNet (Kendall et al., 2015), demostrando obtener la posición directamente a partir de una única imagen. A partir de ello, surgieron diversos trabajos que buscaban mejorar la precisión en la estimación, acelerar el entrenamiento y mejorar el tiempo de inferencia.

Algunos trabajos tempranos introducen modificaciones en la función de pérdida (Kendall & Cipolla, 2016) e incorporan restricciones geométricas para mejorar el resultado en la estimación (Kendall & Cipolla, 2017). Otros enfoques se centran en optimizar la arquitectura congelando capas intermedias, integrando módulos temporales, estructura de otras redes y métodos de suavizado para reducir la parametrización de los pesos y mejorar las predicciones (Müller et al., 2017; Wang et al., 2020; Zhang et al., 2018). Estas primeras versiones demostraron que PoseNet es ideal para aplicaciones tales como: localización en interiores (Acharya et al., 2019), uso de panoramas en escenarios de carretera (Cimarelli et al., 2019) y geolocalización con imágenes aéreas capturadas con drones (Cabrera-Ponce & Martinez-Carranza, 2019). Estos avances demostraron a PoseNet como un punto de partida para desarrollar modelos más especializados en diversos escenarios.

Tras las primeras modificaciones de PoseNet, surgieron enfoques más avanzados que exploraron su capacidad hacia escenarios más complejos. Algunos de ellos incorporan información temporal para mejorar la estimación en entornos urbanos, donde las condiciones del escenario afectan el rendimiento de los métodos clásicos (Li et al., 2021; Wattanacheep & Chitsobhuk, 2020). Otros estudios combinan PoseNet con el filtro Kalman para refinar las predicciones y reducir el ruido presente en las posiciones estimadas (Gu et al., 2021; Yang et al., 2021). Además, se han propuesto versiones compactas y destiladas que permiten acelerar la inferencia y disminuir el consumo computacional, resultando útil en aplicaciones de localización aérea tanto en interiores como en exteriores (Cabrera-Ponce et al., 2022; Rojas-Perez & Martinez-Carranza, 2023). La Figura 3.3 muestra un ejemplo de estas implementaciones, donde una versión ligera llamada *CompactPN* se usa en un dron para estimar su posición.

Como se ha observado, PoseNet y sus variantes han sido ampliamente empleadas para estimar la posición de una cámara a partir de una única imagen, demostrando su uso en múltiples escenarios, incluyendo entornos virtuales, interiores y exteriores. Algunos trabajos han aprovechado esta arquitectura dentro de estructuras 3D (Blanton et al., 2022) y han mostrado que compartir características visuales entre distintos entornos mejora la localización (Blanton et al., 2020). No obstante, las limitaciones de PoseNet en información temporal ha impulsado en el desarrollo de arquitecturas independientes combinando redes

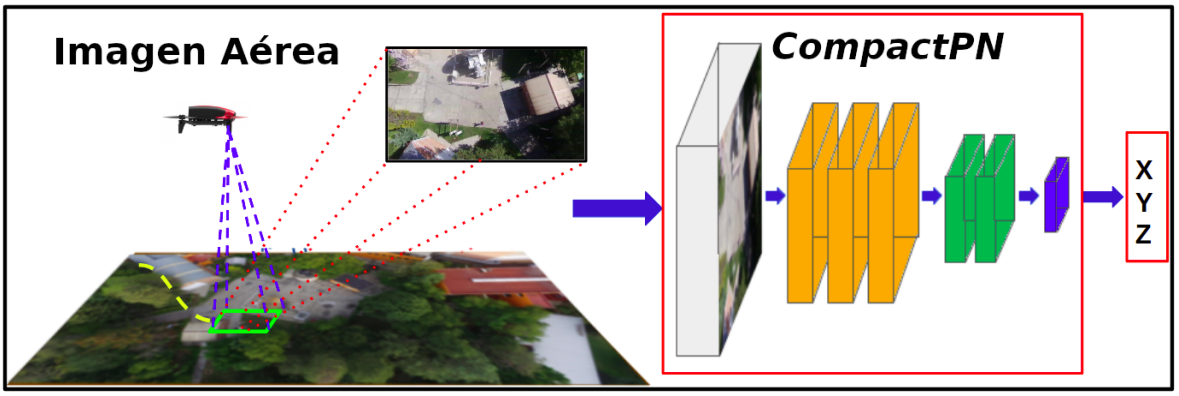

Figura 3.3: Diagrama de localización aérea utilizando la versión ligera *CompactPN* a partir de una imagen capturada con un dron.

neuronales recurrentes con odometría visual para permitir modelar la dinámica del escenario. Trabajos como (Mohanty et al., 2016) aprovechan la correlación entre imágenes consecutivas para estimar simultáneamente la trayectoria, la posición y la profundidad. Además, enfoques como (Li et al., 2018; Wang et al., 2017) extienden esta idea generando conjuntos de datos que integran información tanto de posición como de profundidad, mejorando su uso en tareas de localización y navegación.

Trabajos más recientes extienden la integración de la odometría visual con el aprendizaje profundo al incorporar correlaciones entre imágenes y características extraídas de mapas, permitiendo compartir información auxiliar a la red y mejorar la precisión de estimación (Lin et al., 2019; Valada et al., 2018). Este enfoque no solo optimiza la navegación autónoma, sino que también permite la fusión de información visual para estimar la posición en entornos complejos (Ott et al., 2020). Además, estas propuestas demuestran que aprovechar la información de mapas a múltiples escalas proporciona una interpretación más completa del entorno, construyendo modelos de estimación más robustos (Zhi et al., 2022). Finalmente, la capacidad de combinar mapas, características visuales y aprendizaje profundo permite que estos métodos se apliquen en escenarios reales con vehículos aéreos para tareas de localización y navegación (Bednář et al., 2022).

Además de los modelos basados en PoseNet y odometría visual, se han propuesto otros enfoques que utilizan metodologías híbridas y arquitecturas especializadas para mejorar la estimación de posición. Por ejemplo, algunos trabajos combinan procesos de regresión gaussiana con aprendizaje profundo para mejorar la precisión entre las imágenes y la posición de la cámara estimada (Cai et al., 2018). Por otra parte, redes siamesas han sido utilizadas

para obtener posiciones absolutas y relativas a partir de dos imágenes, aprovechando así las vistas consecutivas del entorno (Kim & Ko, 2022). Otros trabajos se centran en aprender representaciones jerárquicas del entorno y directamente de los píxeles de las imágenes para obtener una localización robusta (Sarlin et al., 2021; Yang et al., 2019).

En los últimos años, investigaciones recientes han mejorado los modelos de estimación de posición, reduciendo la parametrización de los pesos durante el proceso de entrenamiento (Seifi & Tuytelaars, 2019). De igual manera, se han incorporado arquitecturas basadas en transformadores para optimizar la complejidad del modelo sin sacrificar precisión, aprovechando la información temporal, espacial y visual para una regresión más robusta (Jantos et al., 2023; Qiao et al., 2023). En complemento, se han propuesto técnicas como la *"Salpicadura Gaussiana"* que combina correspondencias de rayos de los píxeles con modelos geométricos para estimar con precisión la posición de la cámara en escenas 3D (Matteo et al., 2024). En conjunto, estos avances ofrecen nuevas direcciones hacia una localización más precisa, aprovechando el uso entre redes profundas y la geometría del escenario.

Mientras que todas estas metodologías comparten el objetivo de estimar la posición de la cámara, la mayoría utiliza conjunto de datos de imágenes en escenarios controlados. No obstante, investigaciones recientes dirigen estos avances a aplicaciones del mundo real como la robótica donde el aprendizaje profundo se emplea como sistema de localización (Liu et al., 2022). En estos sistemas, la red agrega información como mapas de profundidad, flujo óptico y deformaciones de características, permitiendo estimar la distancia entre una cámara y su entorno, recuperando así su localización basándose únicamente en imágenes (Guo et al., 2021; Mandal & Jain, 2022; Taguchi & Hirose, 2022). Los resultados en robots móviles demuestran que estas técnicas ofrecen un rendimiento robusto en entornos reales, superando las limitaciones de los métodos visuales clásicos.

Uno de los escenarios más desafiantes para la localización es el entorno aéreo debido a las posibles pérdidas de señal GPS causadas por condiciones ambientales, interferencias o limitaciones de cobertura. En estas situaciones, la localización visual surge como alternativa para mantener la navegación y recuperar la posición del vehículo aéreo (Shen et al., 2025). Diversos trabajos enfrentan este problema mediante aprendizaje profundo y aprendizaje por refuerzo, combinando información visual de imágenes aéreas con datos de imágenes satelitales para estimar una posición cercana a un punto de referencia (He et al., 2023; Pirinen et al., 2022; Vallone et al., 2022). Estos enfoques permiten reconocer ubicaciones previamente visitadas así como relocalizar al vehículo aéreo, ofreciendo un sistema confiable para la navegación autónoma durante misiones de vuelo.

Además de las imágenes aéreas, diversos trabajos utilizan información complementaria para reforzar la estimación de posición y mejorar la localización visual. Entre ellas se

encuentran sensores IMU, imágenes georreferenciadas, imágenes térmicas (Helgesen et al., 2019) y puntos clave basadas en red de gráficos relacionales (Jin et al., 2019). Estos tipos de estrategias son esenciales en escenarios complejos, como misiones sobre océanos con drones de ala fija (Herrera et al., 2025; Wickramasuriya et al., 2025) o escenarios nocturnos (Xing et al., 2025) donde las condiciones dificultan la localización precisa. Sin embargo, el uso de imágenes georreferenciadas y satelitales presentan aún desafíos ya que carecen de información de posición, complicando su uso para el entrenamiento de modelos de aprendizaje profundo. Para enfrentar este problema surgen propuestas como AstroLoc (Berton et al., 2025), un método que empareja imágenes tomadas por astronautas con vistas aéreas para generar etiquetas de localización e integrarlas. La Figura 3.4 muestra la integración de información complementaria para obtener la localización visual de un dron de ala fija, utilizando mapas de referencia para estimar la posición de la cámara.

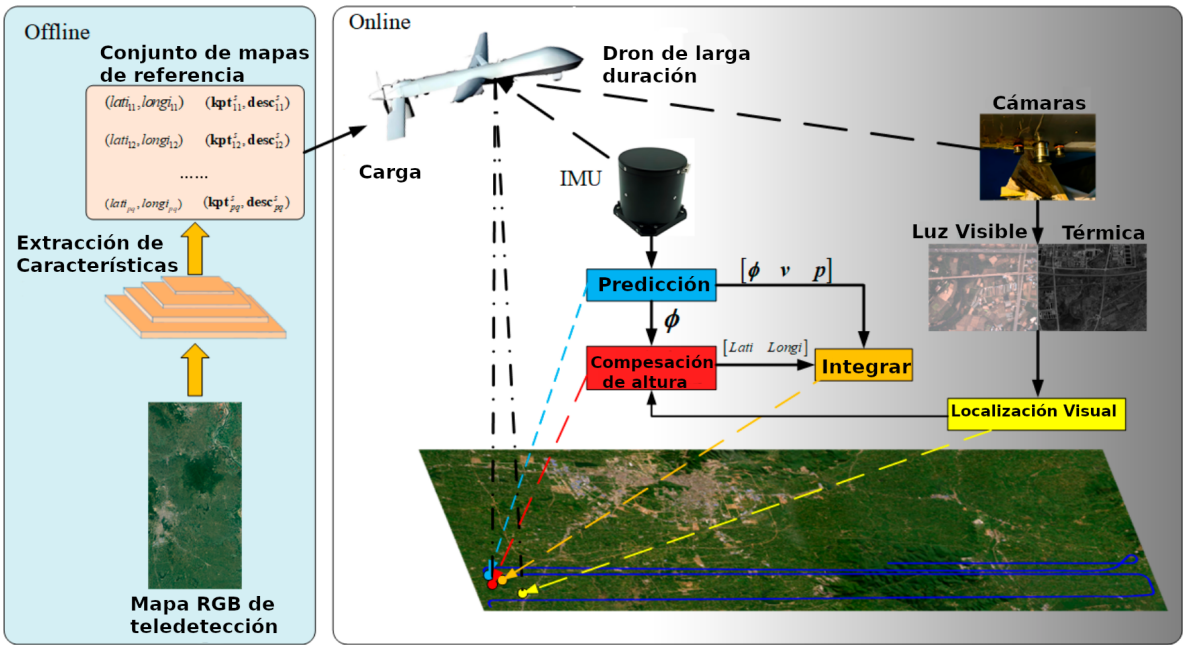

Figura 3.4: Localización visual utilizando imágenes térmicas, imágenes RGB e información IMU para estimar la posición de georreferencia a partir de un mapa de referencia.

Los trabajos revisados demuestran que la integración de información complementaria es útil para mejorar la precisión de la localización visual en escenarios aéreos complejos. Sin embargo, a pesar de los avances logrados mediante aprendizaje profundo, permanecen limitaciones tales como: conjunto de datos grandes, altos recursos computacionales y extensos tiempos de entrenamiento. Estas limitaciones dificultan su aplicabilidad directa

en escenarios reales con drones, donde se necesita una localización persistente. Por lo tanto, estas restricciones han motivado al desarrollo de nuevas metodologías que buscan reducir la dependencia de grandes conjuntos de datos y entrenamientos prolongados. Entre estas metodologías se encuentra el aprendizaje continuo como un enfoque que permite actualizar los modelos de forma incremental, adaptándose a nueva información sin necesidad de reentrenar desde cero lo que resulta adecuado para tareas de localización visual y estimación de posición en robótica aérea.

3.3. Aprendizaje Continuo para Localización

Como revisamos en la sección anterior, el aprendizaje profundo ha demostrado ser efectivo para la extracción de características visuales y asociarlas a modelos de estimación de posición. No obstante, presentan limitaciones tales como: el uso de grandes volúmenes de datos y prolongados tiempos de entrenamiento, dificultando su aplicación en escenarios reales como la robótica. Para enfrentar estas restricciones, en la última década surgió un método llamado aprendizaje continuo cuyas estrategias estan diseñadas para permitir que los modelos aprendan de manera incremental a partir de nuevos lotes de información, evitando reentrenar con conjuntos de datos extensos. Un trabajo popular en este campo fue CORe50 (Lomonaco & Maltoni, 2017), quien introdujo un conjunto de datos diseñado para el aprendizaje continuo utilizando estrategias basadas en memorias externas, donde se almacenan patrones previamente aprendidos y se mezcla con nueva información, manteniendo así una repetición latente del conocimiento. Este enfoque presentó la base para el desarrollo de nuevas metodologías capaces de integrar el aprendizaje continuo y la estimación de posición, ofreciendo un camino hacia sistemas eficientes y adaptativos para la localización visual en entornos cambiantes.

Dado que los escenarios pueden cambiar con frecuencia, el aprendizaje continuo permite adapatar los modelos de manera incremental sin olvidar el conocimiento previo. De esta manera, trabajos como (He et al., 2020; Li et al., 2020) proponen estrategías basadas en destilación cruzada y partición de lotes, haciendo que el entrenamiento se divida en subcojuntos para que la red aprenda sin necesidad de procesar todos los datos simultaneamente. Estas estrategías permiten que el aprendizaje mitigue el olvido catastrófico, permitiendo que el modelo mantenga su rendimiento en escenarios previamente visitados mientras incorpora nueva información. Por otro lado, trabajos como (Cui & Chen, 2023; Zaffar et al., 2023) utilizan estrategias combinando la repetición latente con interpolación y extrapolación de descriptores visuales obtenidos con LiDAR, para mejorar el reconocimiento de lugares. Este enfoque permite que el sistema sea capaz de

almacenar representaciones de entornos visitados y reutilizarlos para la identificacion de lugares conocidas, siendo adecuado para tareas de localización visual.

Además de los trabajos anteriores, el aprendizaje continuo también se ha integrado en sistemas de mapeo y odometría visual para mejorar la estimación de posición en escenarios dinámicos. De esta manera, trabajos como (Cai & Müller, 2023; Yan et al., 2021) emplean estrategias de repetición latente y destilación de conocimiento en combinación con representaciones implícitas del entorno como campos de radiación neuronal (NeRF, por sus siglas en inglés) que permiten reconstruir mapas 3D a partir de imágenes. Estas metodologías aprovechan el aprendizaje continuo para actualizar progresivamente los párametros del modelo, incorporando nueva información sin requerir reentrenar desde cero. De igual manera, otros trabajos (Cudrano et al., 2024; Pan et al., 2024; Vödisch et al., 2023b) combinan la repetición latente con odometría visual y redes duales, permitiendo que el sistema integre experiencias pasadas y mantenga correspondencias visuales consistentes en entornos cambiantes. Estos avances demuestran que combinar aprendizaje continuo con técnicas de mapeo y odometría visual permite desarrollar modelos capaces de adaptarse a nuevos cambios de escenarios mientras se minimiza el olvido catastrófico.

En la misma línea, el aprendizaje continuo se ha incorporado con sistemas SLAM para mejorar la estimación de posición, trayectoría y la construcción de mapas en escenarios dinámicos. Por ejemplo, (Chen et al., 2021) propone alinear características de profundidad con modelos 3D, integrando información visual para obtener localizaciones más precisas, mientras que (Sucar et al., 2021) actualiza el mapa de forma incremental, completando regiones no vistas a partir de nuevas observaciones. SLAM continuo (Vödisch et al., 2023a) combina memorias duales con repetición latente para preservar información de lugares visitados mientras se óptimiza la detección de cierre de bucles en mapas 3D. Por otra parte, BioSLAM (Yin et al., 2022), es un trabajo inspirado en el comportamiento de las palomas, utilizando una memoria a largo plazo para reconocer trayectorias y facilitar la navegación en entornos previamente explorados. Finalmente, (Li et al., 2024) introduce un enfoque basado en destilación y repetición de conocimiento, conservando características relevantes y olvidando las innecesarias, permitiendo adaptar la localización a escenarios dinámicos.

No obstante, el uso de sistemas basados en mapeo puede requerir emparejar características visuales con descriptores 3D para estimar la posición de la cámara, incrementando el costo computacional (Moreau et al., 2023). Para abordar esta limitación, otros enfoques emplean estrategias de aprendizaje continuo que reducen la dependencia de los sistemas de mapeo. Por ejemplo, (Truong et al., 2024) introduce una estrategía que restringue los cambios de distribución, mientras que (Cabrera-Ponce et al., 2021) aplica repetición latente en sistemas multicámara para adaptar el modelo a diferentes percepciones visuales y estimar la posición

en escenarios reales de navegación. En (Cabrera-Ponce et al., 2024) presenta una revisión de técnicas de aprendizaje continuo aplicadas a la estimación de posición y localización visual, donde la repetición latente es la estrategia más utilizada, ya que permite preservar el conocimiento previo mientras se incorpora nueva información (Figura 3.5) (Wang et al., 2021).

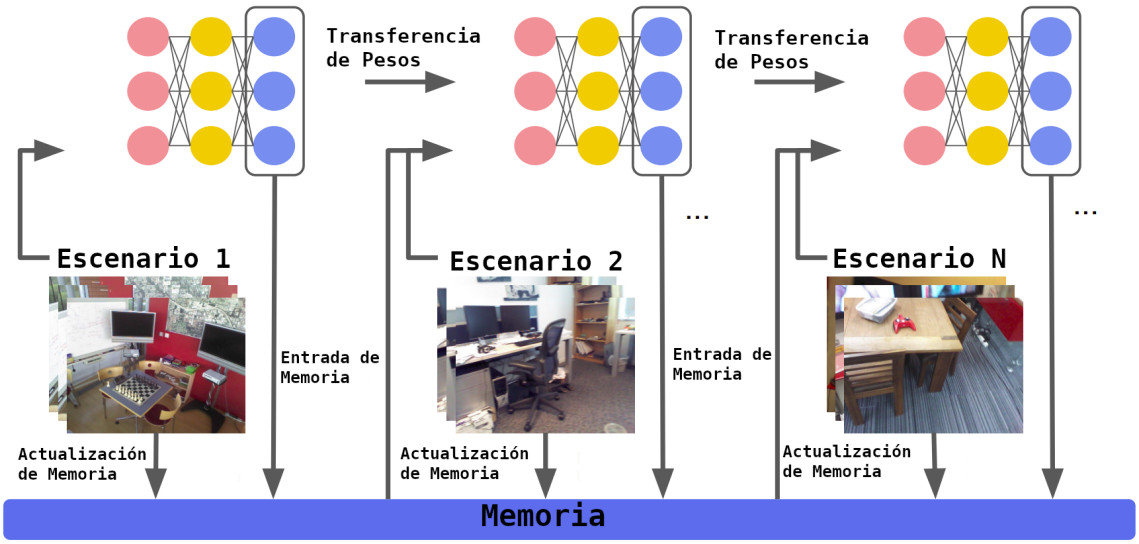

Figura 3.5: Repetición de información para localización visual. El modelo se actualiza utilizando muestras actuales y anteriores (Wang et al., 2021).

Por lo tanto, el aprendizaje continuo basado en repetición latente resulta adecuado para tareas de localización visual y estimación de posición, ya que permite actualizar el modelo al integrar conocimiento previo y nuevo. En robótica, esta estrategía facilita que los robots se adapten a entornos cambiantes aprovechando experiencias pasadas (Yoon et al., 2024), mientras que en robótica aérea es ideal para enfrentar escenarios desafiantes donde los drones deben explorar áreas desconocidas (Safa et al., 2022). Siguiendo esta idea, se han propuesto enfoques que combinan múltiples modelos con repetición latente (Cabrera-Ponce et al., 2022), almacenando información por región para mejorar el reconocimiento del lugar y la localización topológica del dron (Cabrera-Ponce et al., 2023a). Finalmente, metodologías de redes duales han demostrado ser efectivas para identificar regiones dentro de trayectorias y estimar la posición de un dron mientras navega (Cabrera-Ponce et al., 2023b).
En conclusión, el aprendizaje continuo ha demostrado ser un método ideal para mejorar la localización visual y la estimación de posición en entornos dinámicos, actualizando y adaptando modelos de manera incremental al aprovechar experiencias pasadas sin necesidad de reentrenar desde cero. Además, su aplicación en escenarios reales de robótica aérea,

presenta una solución a los diversos retos relacionados con la generalización y la persistencia de la localización. Esto permite el desarrollo de nuevas metodologías que integren el aprendizaje continuo con aprendizaje profundo, optimizando el proceso de entrenamiento y la estimación de posición en drones al adaptarse en entornos cambiantes.

3.4. Otros Métodos de Aprendizaje para Localización

Aunque los métodos basados en aprendizaje profundo demuestran un rendimiento alto en tareas de estimación de posición y localización visual, su elevado costo computacional y el manejo de grandes conjuntos de datos han motivado a explorar otros enfoques. Entre ellos, se encuentran las máquina de vectores de soporte (SVM, por sus siglas en inglés) los cuales comenzaron como métodos de clasificación para determinar la posición del cuerpo humano (Ardizzone et al., 2000; Chen et al., 2011). Estos métodos resultaron eficientes por el rápido entrenamiento y por no requerir un extenso conjunto de datos, reduciendo la complejidad computacional y acelerando el tiempo de inferencia. Además, ofrecen modelos más ligeros y eficientes, ideales para escenarios donde la velocidad de estimación y adaptabilidad en entornos dinámicos son clave.

En consecuencia, las SVM han sido utilizadas en tareas de estimación de posición humana y estimación de estados para la localización visual de objetos (Ionescu et al., 2009). También se han combinado en distintos escenarios para el seguimiento de objetivos mediantes cámaras, integrando algoritmos genéticos con SVM para mejorar la precisión (Madhukaro & Rao, 2024). Asimismo, se han aplicado en reconocimiento de lugares para seguimiento de trayectorias y localización visual, utilizando imágenes estéreo y modelos dinámicos (Qiao et al., 2015). Finalmente, variantes como las Máquinas de Soporte Vectorial para Regresión (SVR, por sus siglas en inglés), han sido empleadas en tareas de reconocimiento de lugares a partir de mediciones de rádio en vehículos aéreos (Lima et al., 2019), destacando su potencial más allá de la clasificación tradicional.

De esta manera, las SVR han demostrado ser eficientes en tareas más complejas tales como la localización visual, la estimación de posición, la navegación autónoma y la detección de objetivos (Jondhale et al., 2022; Wang et al., 2024). A diferencia de las SVM, las SVR están diseñadas para predecir valores continuos, ideales para estimar posiciones con mayor precisión, aprovechando la información contenida en los vectores de soporte que actúan como una memoria que guía al modelo durante la inferencia. Trabajos iniciales como (Ruping, 2001; Syed et al., 1999) exploraron el aprendizaje incremental al actualizar los vectores de soporte, obteniendo resultados prometedores aunque limitados en tareas de clasificación. Posteriormente, (Cabrera-Ponce et al., 2025) extendió este concepto al

proponer una metodología basada en múltiples modelos de SRV combinado con aprendizaje continuo para la localización visual a partir de imágenes capturadas por drones.
Este avance impulsó la exploración de enfoques más ligeros y eficientes, entre ellos están las redes neuronales binarias (BNNs, por sus siglas en inglés) como alternativa para tareas de localización visual. Inicialmente estas redes fueron exploradas para modelos de lenguaje con el objetivo de reducir la parametrización y el tamaño de los modelos al binarizar los pesos y las activaciones (Courbariaux et al., 2016). Posteriormente, su uso se extendió a visión por computadora mostrando resultados en aplicaciones de estimación de postura humana (Bulat & Tzimiropoulos, 2017) y el reconocimiento de imágenes en tiempo real en tarjetas FPGA (Jokic et al., 2018). Además de reducir la complejidad y el peso del modelo, la binarización acelera la inferencia en las estimaciones mostrando un amplio uso para tareas de regresión (Bulat et al., 2019).
Si bien el auge de las redes binarias sigue en crecimiento, este presenta detalles en cuanto a la precisión debido a la pérdida de información durante la binarización de los pesos y activaciones. Por lo tanto, trabajos como (Ma et al., 2024; Wang et al., 2023) abordan este problema utilizando una técnica de decuantización, lo cual permite recuperar la información despues de la binarización sin perder precisión en el resultado final. Esto garantiza a la red cambiar puntos flotantes a datos binarios, manteniendo un desempeño adecuado comparable a los modelos tradicionales. De esta manera, las redes binarias logran convertirse en modelos ligeros para sistemas embebidos, sin sacrificar precisión (Chee et al., 2023). Asimismo, estrategias como la destilación de conocimiento han demostrado mejorar el entrenamiento y el rendimiento en tareas de estimación de pose (Zhang et al., 2023).
En conclusión, las redes binarias representan un novedoso método para acelerar el proceso de entrenamiento y reducir su complejidad, siendo ideales para tareas de visión por computadora tales como el reconocimiento, estimación y localización. Además, la incorporación del aprendizaje continuo amplía aún más sus posibilidades, ofreciendo modelos ligeros, rápidos y adaptables para el reconocimiento de lugares (Wang et al., 2024) y la estimación de posición con drones (Ponce et al., 2024). Así, las redes binarias ofrecen resultados prometedores para diversos desafíos en robótica, representando un método novedoso para escenarios reales.

3.5. Análisis Sistemático y Retos Actuales

Las investigaciones expuestas en esta revisión abarcan desde enfoques tradicionales hasta métodos basados en aprendizaje profundo, mostrando las múltiples estrategias para abordar el problema de localización visual. Gran parte de estos estudios utilizan métodos de mapeo y aprendizaje con conjuntos de datos diseñados mayormente en diferentes escenarios terrestres.

Esto abre una brecha en la robótica para escenarios aéreos, donde los conjuntos de datos disponibles son limitados, llevando a crear conjuntos de imágenes aéreos propios. En la Figura 3.6, se muestran los conjuntos de datos más utilizados en los diferentes enfoques de localización presentados en este capítulo.

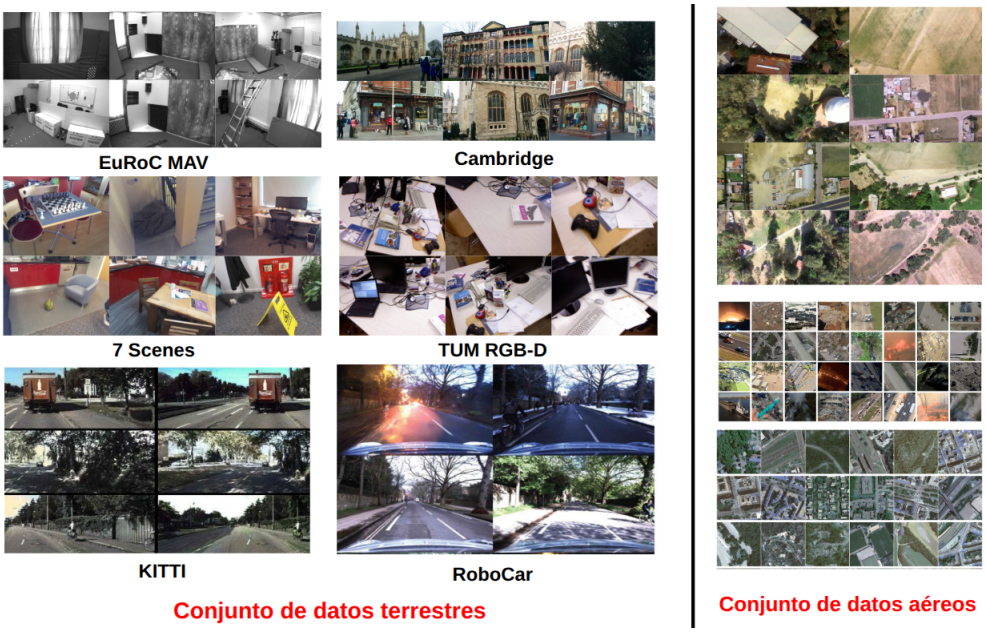

Figura 3.6: Conjuntos de datos utilizados en la literatura para estimación de posición.

A pesar de la amplia variedad de métodos y estrategias para la localización visual, la integración de redes neuronales convolucionales con aprendizaje continuo sigue siendo un campo poco explorado. Esta limitación abre la posibilidad de nuevas investigaciones hacia enfoques con modelos más ligeros y rápidos en aplicaciones con drones e imágenes aéreas. Si bien han comenzando a surgir alternativas como redes binarias, aún se encuentran en etapas tempranas de incorporación al aprendizaje continuo lo que ofrece un campo de investigación a explorar. Debido al limitado número de trabajos en esta área, se busca ampliar este método para distintas soluciones a los procesos clásicos de estimación de posición utilizando cámaras monoculares con drones.

La dirección para esta área de investigación está dirigida al desarrollo de sistemas de localización visual adaptativos, capaces de ajustarse dinámicamente a distintos escenarios. Como vimos en el capítulo, la combinación de aprendizaje continuo con otras técnicas ofrece un marco novedoso para diseños metodológicos que utilicen únicamente la cámara como fuente de datos. Tal como muestran trabajos representativos en la literatura (Cabrera-Ponce et al., 2021; Jin et al., 2019; Sucar et al., 2021; Vödisch et al., 2023b; Wang et al., 2021) los

métodos evolucionan incluyendo elementos complementarios como ubicaciones de sensores externos, imágenes satelitales, mapeo, entre otras anotaciones. Por lo tanto, la tendencia hacia modelos adaptativos crece con el aprendizaje continuo permitiendo que en el futuro los sistemas de localización visual tengan un papel central tanto en la robótica como en aplicaciones de interés científico.

A pesar de los avances recientes los retos actuales continúan con la integración del aprendizaje continuo en la robótica para lograr una localización visual precisa, navegación autónoma robusta, exploración y adaptación a entornos desconocidos. Precisamente, la estimación de posición precisa sigue siendo fundamental para tareas complejas las cuales enfrentan limitaciones frente a escenarios dinámicos, requiriendo un sistema de localización alternativo. De esta manera, en los siguientes años la aplicación del aprendizaje continuo en robótica aérea podría permitir que los drones mantengan su localización en ausencia de GPS. Además, este avance permitirá que los modelos de localización se adapten a nuevos entornos para relocalizar al dron durante misiones de vuelo.

3.6. Sumario

En resumen, los estudios revisados presentan múltiples enfoques para resolver la estimación de posición y la localización visual con imágenes monoculares, utilizando métodos clásicos de visión hasta técnicas avanzadas de aprendizaje profundo. Dentro de estos métodos, se explora el aprendizaje continuo como una alternativa para adaptar modelos de estimación sin entrenar desde cero, otorgando un resultado de manera más rápida. Sin embargo, este avance enfrenta el reto del olvido catastrófico, donde la incorporación de nueva información ocasiona pérdida del conocimiento previamente adquirido. Por lo tanto, un avance para el progreso de este método en estimación de posición y localización visual es requerido para alcanzar las limitaciones abiertas dentro de este campo.

Con este propósito, se presenta una línea de tiempo de las investigaciones en aprendizaje continuo, haciendo énfasis en aquellas enfocadas a tareas de estimación de posición y localización visual entre 2017 y 2025 (Figura 3.7). En los primeros años (2017-2018), el enfoque principal fue la actualización de modelos de clasificación preservando el conocimiento mientras se incorporan nuevos datos. Para los años (2019-2020), la investigación se expandió hacia el área de visión y robótica, donde el aprendizaje continuo ofreció alternativas de adaptación sin necesidad de largos entrenamientos y conjuntos de datos grandes. Esto representó un cambio en tareas de localización visual, más allá del marco tradicional de clasificación y reconocimiento de lugares, dando la base para nuevos avances en estimación de posición.

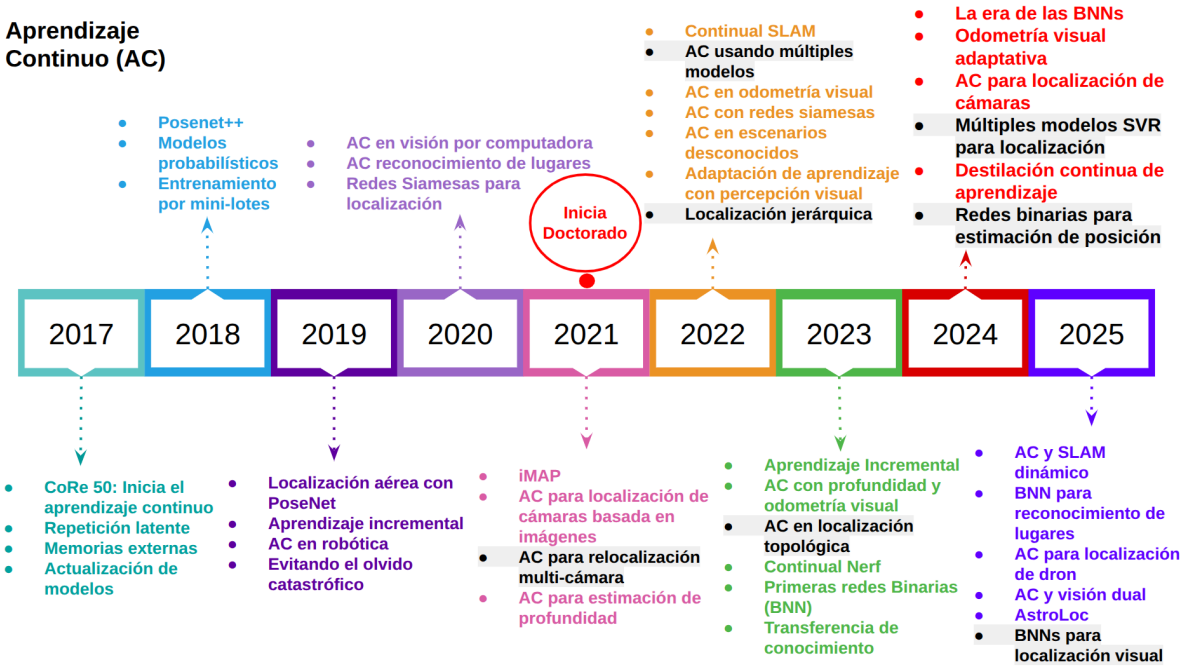

Figura 3.7: Cronología del aprendizaje continuo y sus avances en estimación de posición y localización visual, resaltando los avances en la localización aérea utilizando drones.

Para 2021 surgieron los primeros estudios que incorporaron estrategias de repetición latente para localización visual, representando un avance hacia nuevas metodologías. Ese mismo año se dio inicio esta investigación doctoral, enfocada en el aprendizaje continuo para la estimación de posición con cámaras monoculares en drones, un campo poco explorado hasta entonces. Durante los años (2022-2023) surgen contribuciones que demuestran cómo el aprendizaje continuo es aplicado en diferentes campos de la robótica, logrando su uso como método de localización visual. De esta manera, su uso se expande a sistemas de mapeo con métodos de destilación y repetición latente, estableciendo metodologías para la estimación de posición.

Entre 2024 y 2025, el aprendizaje continuo avanzó hacia enfoques de aprendizaje progresivo, incorporando múltiples modelos, submapas y estrategias ligeras como las máquinas de soporte vectorial y redes binarias. Estas últimas ofrecen modelos compactos para acelerar el entrenamiento y la inferencia, adecuadas para sistemas embebidos en drones. Finalmente, la presente investigación doctoral iniciada en 2021, propone una metodología de localización alternativa basada en aprendizaje continuo, diseñada como sistema de respaldo en escenarios donde la señal GPS se pierde. Así, buscamos obtener un esquema de localización visual capaz de adaptarse a misiones de vuelo reales utilizando modelos ligeros y entrenamiento continuo.

Capítulo 4

Metodología General

Este capítulo describe la metodología general de la investigación para la estimación de posición de una cámara a bordo de un dron y su localización visual empleando arquitecturas neuronales con aprendizaje continuo. Así mismo exponemos las variables de estudio que se definieron en esta investigación, la población, las métricas, así como la infraestructura y el impacto socioeconómico. Finalmente presentamos los procedimientos realizados para la adquisición del conjunto de datos, la estrategia metodológica de aprendizaje continuo y las arquitecturas base para el entrenamiento.

4.1. Variables de Estudio

Para este trabajo las variables que se utilizaron para estimar la posición y encontrar la localización visual con una cámara monocular fueron las imágenes capturadas por la misma. Las imágenes fueron empleadas por la observación y percepción del escenario visto desde la perspectiva de la cámara, cuya información dentro de las imágenes está determinada por unidades en valores de píxel. Así, esta información y el uso de redes neuronales convolucionales las transforman a conceptos más abstractos expresados en pesos para la construcción de un modelo de entrenamiento. Un ejemplo detallado de esta variable se expresa en la siguiente información:

- **Denominación:** Imágenes monoculares a color.
- **Tipo de Variable:** Variable independiente
- **Naturaleza:** Variable cuantitativa debido a que la captura de imágenes puede ser continua o discreta.

- **Medición:** Numérico
- **Indicador:** Píxeles
- **Unidad de medida:** Dimensión y color
- **Instrumento:** Cámara monocular para captura de video
- **Dimensión:** $(128 \times 128 \times 3)$, $(224 \times 224 \times 3)$ y $(320 \times 240 \times 3)$
- **Definición operacional:** Dimensión de la imagen y canales de color.
- **Definición conceptual:** La dimensión de la imagen está dada por la posición de los píxeles en ancho y alto, la profundidad corresponde al rango de los canales de color de la imagen donde un canal representa escala de grises y el uso de 3 canales representa los colores RGB (Rojo, Verde, Azul). Los píxeles son procesados para obtener características visuales robustas, utilizados para determinar la posición de la cámara por medio de redes neuronales convolucionales, permitiendo la localización dentro del escenario.

Otro tipo de variable utilizado en esta investigación fue la posición global obtenida del sensor GPS. Su medición está dada en coordenadas tridimensionales de latitud, longitud y altura representando la ubicación obtenida por medio de observaciones satelitales. Esta señal permite una triangulación con diferentes satélites obteniendo así la posición dentro de un escenario. Un ejemplo de esta variable se expresa en la siguiente información:

- **Denominación:** Posición global tridimensional por GPS
- **Tipo de Variable:** Variable dependiente
- **Naturaleza:** Variable cuantitativa. Las posiciones capturadas pueden interpretarse a datos tridimensionales de X, Y, Z.
- **Medición:** Numérico
- **Indicador:** Posición tridimensional a un punto de referencia.
- **Unidad de medida:** Principalmente en grados con el sistema de coordenadas WGS84 y posteriormente representada a metros pasando al sistema UTM.
- **Instrumento:** Sistema de posicionamiento global (GPS)

- **Dimensión:** La dimensión de la posición depende de la trayectoria de vuelo, teniendo distancias representadas en metros con valores entre 0 y 1000 metros.
- **Definición operacional:** La posición se obtiene con respecto a un punto de referencia y el punto estimado. La distancia entre ambos puntos puede ser en centímetros, metros y kilómetros en función del error entre ellas y la normalización utilizada.
- **Definición conceptual:** Las coordenadas GPS están representadas en un inicio en grados geográficos, donde por medio de una brújula interna se indican los diferentes puntos cardinales. Estos representan el Norte (0°), Este (180°), Sur (180°) y Oeste (270°) de la Tierra. De esta manera, los grados proporcionan una medida angular para determinar la orientación y dirección de un objeto en relación a un punto de referencia. Por otro lado, la distancia entre el punto estimado y el punto de referencia indica el espacio físico entre ellos, utilizado para medir la separación de un punto a un objeto o lugar. Así, la posición es determinada indicando la ubicación del lugar donde se encuentra un objeto con respecto a un punto de referencia.

4.1.1. Población, Muestra, Métricas e Instrumentos de Recolección

La población que utilizamos es el conjunto de imágenes capturadas con una cámara y las posiciones tridimensionales adjuntadas a cada imagen. A partir de esta población se requiere obtener conclusiones de localización y posición estimada por medio del entrenamiento de los datos. Las muestras que utilizaremos son las mismas imágenes dividida en subconjuntos de datos pequeños, ideales para crear modelos de estimación con métodos de aprendizaje para adquirir la posición estimada con respecto a la posición de referencia de nuestra población. En la siguiente Tabla 4.1 mostramos el tamaño del conjunto de datos de nuestra población así como las muestras que utilizaremos para llevar a cabo esta investigación. Tanto la población como las muestras están conformadas por imágenes monoculares y su posición tridimensional anexa a cada una de ellas.

Tabla 4.1: Población y Muestra definidas para esta investigación.

Tipo	Descripción de la información
Población	Conjunto de datos de imágenes (200 a 6000 imágenes)
Muestra	Para entrenamiento en lotes (20 a 600 imágenes)

Las métricas de evaluación utilizadas fueron divididas en dos partes: 1) Evaluar la localización midiendo el error de la distancia entre posiciones; 2) Evaluar la velocidad y tiempo de

procesamiento del sistema. Para la localización se utilizaron las metricas de Precisión, Raíz del Error Cuadrático Medio (RMSE, por sus siglas en inglés), Error Medio de Distancia Euclidiana, Porcentaje de Error, Error Medio de Localización y Error Mediano de Localización. Mientras que para el procesamiento del sistema se utilizaron el Tiempo Medio de Procesamiento y Cuadros por Segundo (FPS, por sus siglas en inglés). Todo calculado en metros y mili segundos para una mejor representación del resultado obtenido y compararlo con posiciones de referencia así como resultados del estado del arte.
Finalmente, los instrumentos de recolección de datos utilizados fueron cámaras monoculares y sistemas de posicionamiento global (GPS) abordo de un dron, permitiendo la recolección de datos para el procesamiento de la información. Así mismo, para esta investigación utilizamos las siguientes estrátegias de recolección de información:

- **Observación controlada:** Nuestras variables de estudio dependen de las imágenes capturadas con la cámara monocular del dron, cuya información de posición debe estar ubicada en entornos texturizados y con una amplia gama de información. La estrategia de observación controlada permite manejar las variables en base a las perspectivas del escenario, estableciendo un mejor control ante fenómenos observados en su forma natural.

- **Experimentación en campo:** Esta estrategia va de la mano con la observación controlada ya que permite recolectar datos a través de ejercicios en el campo por medio de grabaciones de vídeo y captura de imágenes. También es ideal para registrar información de posición de los escenarios creando así una población de un conjunto de datos cuyas muestras permitirán el entrenamiento y evaluación de los modelos de aprendizaje en esta investigación.

4.1.2. Infraestructura

Para el desarrollo de este trabajo de investigación utilizamos la infraestructura de la Benemérita Universidad Autónoma de Puebla (BUAP) así como del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). Las instalaciones cuentan con laboratorios de cómputo y supercómputo en la BUAP, así como el laboratorio de robótica del INAOE cuyos drones fueron utilizados para realizar los experimentos. El equipo que se empleó principalmente fue una cámara a bordo en una plataforma aérea llamada Matrice 100 de DJI. El laboratorio de supercómputo fue utilizado para el entrenamiento de redes neuronales, debido a la velocidad y capacidad computacional de sus computadoras. Finalmente, las instalaciones del INAOE cuenta con espacios abiertos donde la captura de la información

fue realizado en el centro de información así como el parque tecnológico del instituto. La estación de control utilizada fue una computadora de uso personal con tarjeta gráfica y conexión WIFI lo que permite una adquisición de los datos provenientes del dron mientras realiza una misión de vuelo. Finalmente, para el desarrollo de esta investigación se contó con un presupuesto establecido en la Tabla 4.2, mostrando un costo estimado del material empleado en la metodología y experimentación.

Tabla 4.2: Presupuesto estimado del material utilizado para esta investigación.

Equipo	Precio
Computadora personal	$ 32,000.00 MXN
Matrice 100 DJI	$ 24,694.00 MXN
Cámara Zenmuse X3	$ 3,260.00 MXN
	$ 59,954.00 MXN

4.1.3. Impacto Socioeconómico

El impacto socioeconómico esperado de esta investigación podría encontrarse dentro de los Programas Nacionales Estratégicos (PRONACES) de la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) en la categoría de seguridad humana. De esta manera, la presente investigación puede contribuir en la localización dentro de escenarios rurales y urbanos utilizando imágenes aéreas del entorno donde se encuentran grupos de personas, estableciendo movilidad y seguridad en la población en general. Asimismo, puede llevar a regiones donde nuestra investigación pueda dar información sobre los sectores de población, localizando y estableciendo ubicaciones en lugares alejados de las grandes ciudades lo que beneficiaría una integración a sistemas geográficos e instituciones como INEGI.

4.1.4. Alcances y Limitaciones

El desarrollo de esta investigación y su uso metodológico en diferentes campos de la robótica cuenta con los siguientes alcances:

1. Una metodología capaz de aprender continuamente las posiciones de una cámara monocular con un tiempo menor a los métodos tradicionales.

2. Actualización y adaptación de modelos de aprendizaje que incorporan nueva información de posición mientras conserva el conocimiento previamente adquirido.

3. Una metodología de localización que pueda llevarse a diferentes plataformas robóticas y computadoras a bordo de drones.

4. Relocalización de sistemas aéreos como drones capaces de aprender información de localización en trayectorias de vuelo en escenarios nuevos y previamente visitados.

Asimismo, esta investigación cuenta con algunas limitaciones realistas con respecto al hardware y el software empleado:

1. Capacidad de memoria limitada a las especificaciones de la computadora utilizada.

2. La adquisición de datos y el entrenamiento de los modelos de aprendizaje está limitado a la señal de transferencia con la plataforma aérea.

3. Exceder una distancia mayor a 2 km donde el vehículo esté fuera de vista del piloto puede comprometer al sistema y al escenario donde se realiza el vuelo.

4.2. Recursos Metodológicos

El desarrollo de esta investigación se apoya en un conjunto de datos compuesto por imágenes aéreas capturas con la cámara monocular de un dron. Dicho conjunto fue generado para propósitos experimentales y de validación de la metodología de localización visual propuesta en esta tesis. Asimismo, se presentan las arquitecturas de aprendizaje utilizadas en los procesos de entrenamiento y validación del enfoque de aprendizaje continuo, detallando brevemente su diseño para el manejo eficiente de la información. Finalmente, se presentan de forma general las dos estrategias de aprendizaje continuo implementadas en esta investigación, cuyo propósito es mostrar su aplicación en tareas de estimación de posición y localización visual.

4.2.1. Conjunto de Datos

La generación del conjunto de datos se realizó utilizando el Sistema Operativo Robótico (ROS), estableciendo la comunicación entre el dron y una estación de control en tierra. Esta configuración permite obtener el flujo de imágenes provenientes de la cámara monocular del vehículo aéreo y los datos de posición por el sensor GPS. A partir de esto, se diseñaron 4 trayectorias de vuelo para la recolección de imágenes aéreas en resolución HD. Las imágenes obtenidas fueron organizadas en dos conjuntos: 1) Imágenes asociadas a coordenadas GPS para fines de entrenamiento; 2) Fotogramas clave cada cierta distancia para fines

de búsqueda del lugar. Además, las coordenadas GPS se convirtieron en un sistema métrico para facilitar la interpretación y manejo de la posición.
En la Figura 4.1 se ilustran las cuatro trayectorias de vuelo diseñadas en Google Earth, donde los puntos marcados corresponden a las coordenadas GPS registradas durante los recorridos realizados por el dron. Estas trayectorias representan las rutas de entrenamiento utilizadas en los experimentos de esta investigación. De igual manera, se emplearon recorridos similares con el objetivo de recolectar imágenes de prueba para la evaluación del enfoque propuesto. En la figura, el marcador rojo indica el inicio de la trayectoria, mientras que el marcador verde señala el punto final del recorrido. Para la generación del conjunto de pruebas, las rutas de vuelo fueron ejecutadas en sentido inverso, simulando el retorno del dron al punto de despegue en situaciones de pérdida. Finalmente, la longitud de cada trayectoria fue de 0,53 km, 1.4 km, 2.4 km y 2.9 km respectivamente.

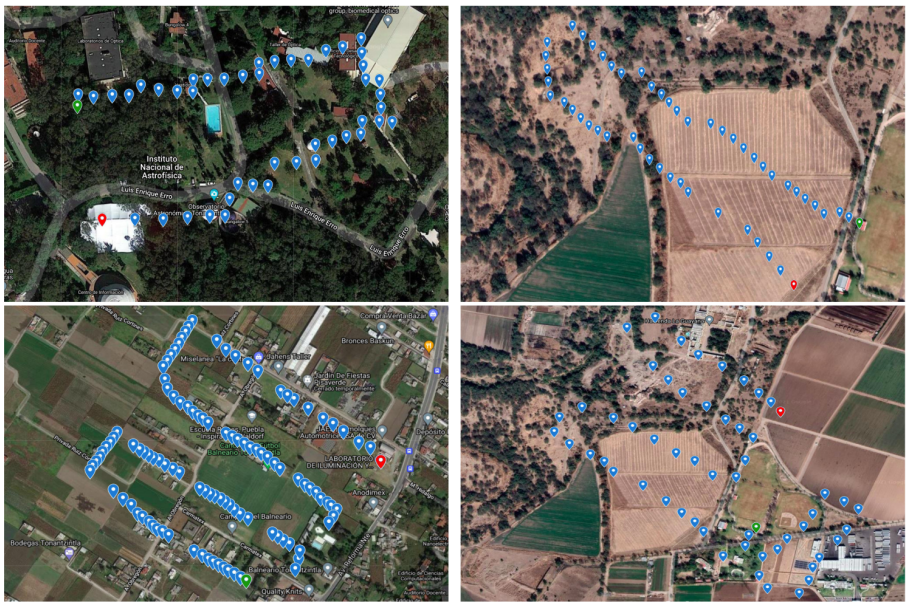

Figura 4.1: Trayectorias de vuelo para la generación del conjuto de datos representado en Google Earth.

Cabe destacar que las imágenes fueron capturadas por el dron a una altura aproximada de 70 a 100 metros, permitiendo obtener una visualización amplia de cada escenario. Posteriormente, las imágenes fueron redimensionadas a los tamaños requeridos para el entrenamiento de las diferentes arquitecturas utilizadas en esta investigación, con resoluciones entre 128×128 y 320×240 píxeles. Las coordenadas GPS, previamente convertidas a metros también fueron utilizadas en el entrenamiento representadas como posiciones medias para localizaciones generales del escenario y posiciones estimadas cercanas

a la ubicación capturada. Finalmente, la Figura 4.2 presenta un ejemplo del conjunto de datos generado en las diferentes trayectorias, donde se aprecian variaciones de perspectiva y altura en escenarios urbanos, rurales y dentro de las instalaciones del INAOE.

Figura 4.2: Imágenes aéreas capturadas para la generación del conjunto de datos.

4.2.2. Redes y Arquitecturas Base

En esta investigación se utilizaron cuatro arquitecturas principales como base para el desarrollo de la metodología de aprendizaje continuo aplicada a tareas de estimación de posición y localización visual. Adicionalmente, se utilizaron dos redes como extractores de características, con el objetivo de obtener descriptores más representativos de las imágenes y utilizarlos en el entrenamiento. Las arquitecturas consideradas incluyen dos redes neuronales convolucionales, una arquitectura basada en Máquinas de Soporte Vectorial para Regresión (SVR) y una arquitectura en Redes Binarias (BNN). Las redes profundas fueron adoptadas de trabajos del estado del arte, mientras que las arquitecturas de SVR y BNN fueron diseñadas en esta investigación con el fin de adaptarlas a tareas de regresión de posición. A continuación, se presenta un listado de las arquitecturas utilizadas, describiendo brevemente los elementos principales que las componen:

- **MobileNetV2:** Esta arquitectura se compone de una capa inicial de convolución con 32 filtros 2 pasos, seguida de una serie de bloques residuales invertidos. Dichos bloques permiten mantener un número reducido de párametros al expandir y comprimir las dimensiones de los tensores, mejorando la eficiencia computacional y la precisión. Además, la red incluye etapas de reducción de tamaño controlando la complejidad y ajustando el campo receptivo. Finalmente, incorpora una convolución de 1280 filtros con una capa de agrupación promedio global y una capa completamente conectada con activación softmax.

- **InceptionV4:** Esta arquitectura es utilizada como base en trabajos como PoseNet ya que ofrece una capacidad de estimación de alta precisión gracias a su diseño profundo. La red está compuesta por 23 capas convolucionales y 9 módulos Inception, los cuales combinan convoluciones de diferentes tamaños de filtro para extraer información en múltiples escalas. Al final cuenta con una capa de convolución de 2048 filtros, seguida de una capa completamente conectada con activación ReLu.

- **SVR:** Más que una arquitectura es un algoritmo de aprendizaje supervisado basado en máquinas de soporte vectorial, cuyo objetivo es aproximar una función que relacione los datos de entrada con valores de salida continuos. Este algoritmo permite crear modelos a partir de la transformación de los datos de entrada utilizando funciones de base radial (RBF, por su siglas en inglés) y un parámetro de regularización. Estas características controlan el equilibrio entre la complejidad del ajuste y la tolerancia al error, ajustando así valores estables y generalizables.

- **BitNet:** Las arquitecturas de redes binarias se caracterizan por sustituir las operaciones en punto flotante a representaciones binarias, permitiendo aplicar técnicas de cuantización y optimización para reducir el consumo de memoria y acelerar la inferencia. La red binaria utilizada en esta investigación está compuesta por 2 bloques convolucionales binarios (bitConv2d), seguidos de capas de activación ReLu y una capa de agrupación máxima. Los bloques binarios tienen 3 capas completamente conectadas las cuales realizan la regresión en un vector de valores continuos. Un componente clave dentro de su arquitectura es el uso del Estimador Directo (STE, por sus siglas en inglés) la cual permite aproximar el gradiente de la función de binarización durante la propagación.

Para las redes base como extractor de características utilizamos 2 ampliamente conocidas en la literatura, gracias a la calidad de las características extraidas. A continuación, se presenta un listado de estas arquitecturas, describiendo los elementos principales que la componen:

- **ResNet18:** Esta red se compone de 18 capas convolucionales organizadas en bloques residuales, cada uno integra dos capas de convolución y una conexión de acceso directo. Su diseño residual permite mitigar el desvanecimiento del gradiente, mejorando la capacidad de entrenamiento en redes más profundas. Además, su arquitectura ligera y eficiente es utilizada como extractor de características, obteniendo representaciones robustas de 512 características en la capa final.

- **DeepPilot4Pose:** Esta arquitectura es una versión destilada de PoseNet, diseñada específicamente para reducir la parametrización del modelo. Su diseño consiste de 4 capas convolucionales seguidas de 3 módulos Inception y una función de activación ReLu, permitiendo combinar filtros de diferentes tamaños para una extracción de representaciones visuales más robustas. También, su diseño compacto genera 1024 características eficientes para tareas de estimación de posición con imágenes monoculares.

4.2.3. Estrategias de Aprendizaje Continuo

En esta investigación se utilizaron dos estrategias de aprendizaje continuo: 1) El método de reproducción latente; 2) Aprendizaje progresivo. La primera consiste en el uso de memorias externas que almacenan representaciones esenciales de experiencias pasadas. Esto permite incorporar información de nuevas posiciones mientras se refuerza el conocimiento previamente adquirido, evitando el olvido de los datos ya aprendidos. Esta estrategia fue aplicada en las redes MobileNetV2 e InceptionV4.
Por otro lado, el aprendizaje progresivo surge dentro de las estrategias arquitectónicas del aprendizaje continuo, cuyo propósito consiste en expandir el conocimiento sin modificar lo previamente adquirido. En esta investigación, se implementó esta estrategia mediante el uso de múltiples modelos, en lugar de modificar directamente la arquitectura de las redes. De este modo, cada modelo conserva la información aprendida de lugares anteriores mientras se expande el conocimiento en nuevos modelos. El aprendizaje progresivo fue aplicado a los modelos SVR y BitNet para mostrar su eficacia y adaptación a nuevos entornos.

Capítulo 5

Localización Topológica

Este capítulo presenta la metodología propuesta y los resultados obtenidos para la localización de una cámara con el uso de aprendizaje continuo aplicado a imágenes aéreas capturadas con un dron en entornos exteriores. A diferencia de los métodos de entrenamiento clásicos, la estrategia utilizada permite incorporar nueva información sin olvidar el conocimiento previamente adquirido, asegurando al vehículo en una posición conocida. Para este propósito, se empleó un conjunto de datos dividido en minilotes junto con la técnica de reproducción latente, permitiendo generar una estructura de posiciones medias que sirve como referencia topológica en caso de pérdida de la señal GPS. En la Figura 5.1, se muestra el flujo metodológico de este proceso de localización, en el cual se utilizó la red MobileNetV2 como base para el entrenamiento continuo.

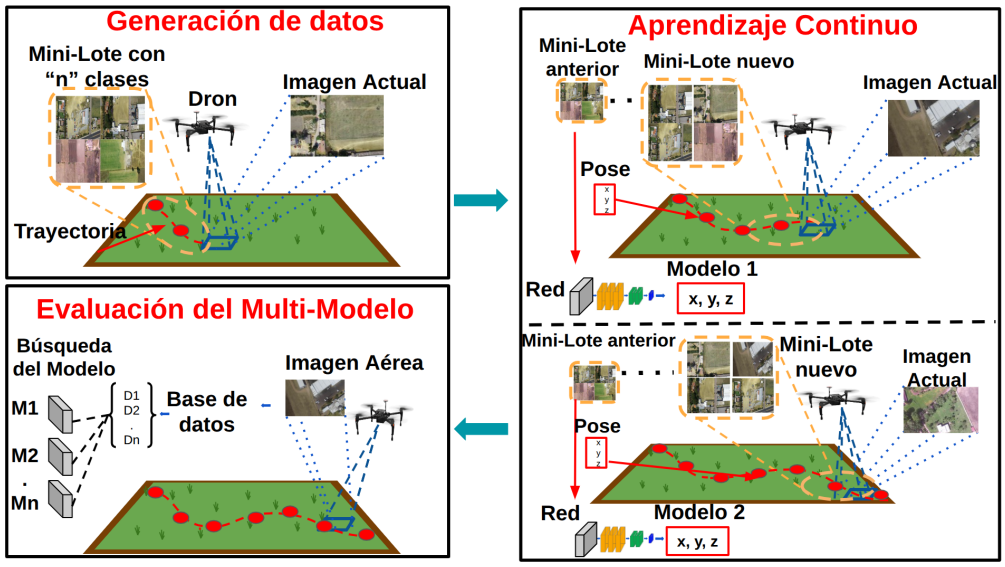

Figura 5.1: Metodología de localización topológica con aprendizaje continuo.

A este proceso lo denominamos localización topológica, ya que representa el entorno como un grafo compuesto por nodos y conexciones, donde cada uno corresponde a una posición media dentro de la trayectoria. Estas posiciones medias no indican coordenadas exactas pero sí proporcionan una aproximación estimada cercana a la localización en la que se encuentra el dron durante el vuelo. De esta forma, la metodología captura una relación espacial relativa entre diferentes segmentos de la trayectoria, estableciendo nodos conectados que representan posiciones cercanas. Así, este enfoque permite que el dron navegue en un escenario mientras construye una estructura topológica, actualizando las referencias de posición en caso de pérdida de señal GPS.

5.1. Preparación del Conjunto de Datos

A diferencia de otros enfoques de aprendizaje profundo, el aprendizaje continuo no requiere mantener un balance estricto del conjunto de datos, lo que permite entrenar la información a partir de minilotes. Para la preparación del conjunto de datos en esta metodología se adoptó la división presentada en (Cabrera-Ponce et al., 2022), adecuada para escenarios de aprendizaje continuo. Esta división permite organizar la información en grupos discretos correspondientes a segmentos en las trayectorias. Así esta interpretación facilita el proceso de entrenamiento continuo, ya que el modelo adquiere conocimiento de nuevos grupos de imágenes sin necesidad de un entrenamiento completo.

Las posiciones capturadas fueron agrupadas en referencias medias, calculadas a partir de las coordenadas reales obtenidas durante el vuelo. Así, la red puede interpretar esta información como clases de localización topológica, definidas en términos de índices que representan regiones aproximadas. La generación de las referencias se realizó volando el dron manualmente mientras recorría la trayectoria vuelo, estableciendo una nueva posición media correspondiente a esa región. En la Figura 5.2 se ilustra este proceso donde la trayectoria en rojo y sus puntos representan las coordenadas de entrenamiento, mientras que la verde corresponde al vuelo de regreso. Los puntos amarillos indican las coordenadas medias generadas de la agrupacion de posiciones, utilizadas como representaciones de las localizaciones de entrenamiento.

Por lo anterior, las trayectorias de vuelo son representadas por puntos intermedios que funcionan como nodos de referencia en las cuales el dron puede relocalizarse en caso de pérdida temporal de la posición. Cada punto de referencia se entrena junto con un minilote de imágenes que corresponden a esa región específica de la trayectoria. Para una mejor comprensión de las etiquetas que representan estas posiciones medias, en la Tabla 5.1 se presenta un ejemplo de los nodos de posición obtenidos del escenario 1. Esta representación

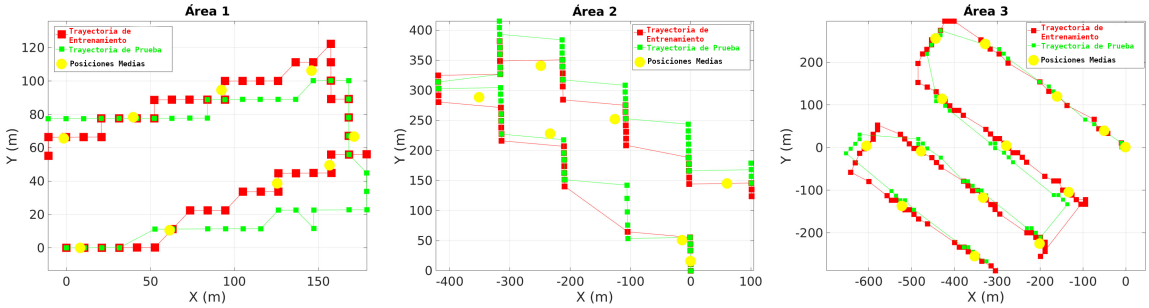

Figura 5.2: Referencias de posiciones medias en cada escenario de vuelo. Las trayectorias de entrenamiento y prueba fueron realizadas en la misma misión de vuelo.

es utilizada para el entrenamiento manejado como clases indexadas a posiciones aproximadas que permiten localizar al dron en un región cercana a la trayectoria. Por lo tanto, la trayectoria 1 cuenta con 9 posiciones de referencia, la trayectoria 2 con 7 referencias, la trayectoria 3 con 14 y la última con 26 referencias. La información detallada de las posiciones medias en cada uno de los escenarios se presenta en (Cabrera-Ponce et al., 2023a).

Tabla 5.1: Posiciones medias de referencia asociadas a minilotes del escenario 1.

índices	Posiciones Medias
0	x: 8.14760, y: 0.03170, z: 30.81
1	x: 61.4042, y: 10.5910, z: 30.94
2	x: 125.455, y: 38.6834, z: 30.71
3	x: 156.644, y: 49.5136, z: 30.34
4	x: 171.174, y: 66.7043, z: 30.38
5	x: 145.899, y: 106.227, z: 30.52
6	x: 92.3082, y: 94.5960, z: 30.52
7	x: 39.7532, y: 78.3282, z: 30.50
8	x: -1.6134, y: 65.6737, z: 30.41

5.2. Entrenamiento con Aprendizaje Continuo

La estrategia de aprendizaje continuo empleada en este estudio fue la repetición latente, la cual introduce una memoria externa en una capa de la red con el propósito de almacenar patrones esenciales de los datos aprendidos mientras se incorporan nuevos. Para la metodología el proceso se implementó en la red MobileNetV2 poniendo la memoria externa en la capa 6 de agrupación, permitiendo almacenar un número de patrones intermedios lo

que ralentiza el aprendizaje en capas inferiores y deja libre las capas superiores. Así, cuando llega un nuevo minilote se extraen sus características y se combinan con las previamente almacenadas en la memoria, rejuveneciendo el conocimiento previo con el reciente. De este modo, el modelo se actualiza de manera incremental integrando información por minilotes y adaptándose a las nuevas posiciones.

Se definió un número de patrones almacenados en la memoria de 1500 a partir del primer minilote. Conforme se incorporan nuevos datos se descartan los poco relevantes para mantener aquellos considerados esenciales, preservando el conocimiento previo junto al nuevo. La red MobileNetV2 está diseñada para aprender hasta 1000 clases diferentes, la cual para este estudio fue adaptada para aprender las posiciones de referencia media descritas en la sección anterior. Para un mejor conocimiento del volumen de datos y su distribución en los distintos minilotes, se presenta la Tabla 5.2 donde se resume la cantidad de imágenes utilizadas en el proceso de entrenamiento.

Tabla 5.2: Imágenes usadas para entrenamiento y validación, así como el número de posiciones medias de referencia y minilotes.

Área	Entrenamiento	Prueba	Posiciones Medias	MiniLote
1	1116	574	9	4
2	644	566	7	3
3	560	340	14	6
4	7826	654	26	12

Finalmente, las especificaciones del entrenamiento fueron las siguientes: 1) Optimizador SGD; 2) Tamaño de minilote de 128 imágenes; 3) Épocas de entrenamiento 4; 4) Función de pérdida basada en entropía cruzada; 5) Tasa de aprendizaje de 0.001; 6) Almacenamiento de 1,500 patrones en la memoria externa. El entrenamiento se llevó a cabo en una computadora personal con Ubuntu 20.04, PyTorch 1.4 y una tarjeta gráfica NVIDIA GeForce GTX 960M. Adicionalmente, se realizaron entrenamientos bajo las mismas condiciones pero incrementando a 3,500 patrones y 10 épocas, con el propósito de comparar el desempeño de los modelos en la localización topológica. Cabe destacar que, aunque el aprendizaje continuo actualiza el modelo al incorporar nueva información, también se generaron modelos independientes por cada minilote, los cuales fueron entrenados con las posiciones de referencia media asociadas.

5.3. Estrategia de Búsqueda

En este estudio se incorporó una estrategia de búsqueda motivada por la necesidad de gestionar las múltiples posiciones medias de referencia a lo largo de las trayectorias y los modelos independientes generados en cada minilote. Con este fin, se adoptó una idea simple basada en el histograma de color de la imagen como descriptor visual. El concepto consiste en identificar distribuciones similares entre una imagen de consulta y una base de datos asociada a las posiciones de referencia media. Por lo tanto, se transformó el espacio de color de BGR a HSV permitiendo una interpretación más robusta de la información visual con un vector de características de 288 dimensiones para su uso en el proceso de búsqueda.

En la Figura 5.3 se presenta el diagrama del proceso de búsqueda con el histograma de color permitiendo una mayor claridad de la estrategia. Para ello, cada imagen se dividió en cinco cuadrantes calculando un histograma de color en cada uno de ellos generando un descriptor de 288 valores en punto flotante. Después, se compara el descriptor de la imagen de consulta con las imágenes representativas asociadas a las posiciones medias de referencia, encontrando la similitud entre descriptores utilizando la distancia chi cuadrada. De esta manera, al identificar el emparejamiento más cercano se recupera la posición media correspondiente a esa imagen. El proceso de búsqueda es aplicado tanto al modelo entrenado con todas las posiciones medias para encontrar la correspondiente, y a los modelos independientes asociados a regiones de la trayectoria permitiendo obtener la posición de esa región.

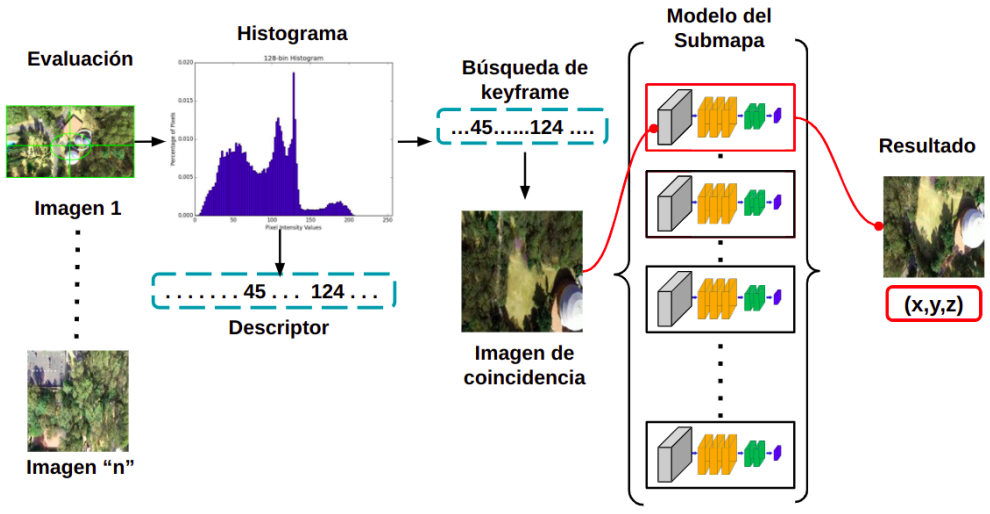

Figura 5.3: Estrategia de búsqueda por medio de histogramas de color para encontrar la posición media correspondiente.

Esta estrategia de búsqueda resulta necesaria para el proceso de localización topológica, ya que permite identificar el nodo de referencia cuya información de posición media está asociada. Además, permite obtener una localización aproximada adecuada para ubicar a un dron durante una misión de vuelo. Si bien, el uso de histograma de color como descriptor señala ser un técnica de busqueda simple y puede ser sustituida más adelante por métodos más robustos de búsqueda. No obstante, su incorporación en este estudio demostró encontrar de manera efectiva una posición aproximada de la cámara dentro de un escenario. En este sentido, la presente tesis muestra un avance inicial hacia la localización visual de un dron, identificando regiones del escenario y cargando el modelo entrenado de esa zona para obtener una localización aproximada.

5.4. Evaluación Experimental

El rendimiento de la metodología de localización topológica fue evaluado utilizando un conjunto de datos de prueba obtenido durante el vuelo de regreso del dron. En nuestros primeros pasos, comparamos la eficiencia de los 3 diferentes modelos, todos entrenados mediante el método de repetición latente en el marco del aprendizaje continuo. El primer modelo consideró 1,500 patrones almacenados en la memoria externa, el segundo 3,500 patrones, y el tercero mantuvo 1,500 patrones pero con un entrenamiento a 10 épocas. Cada modelo se actualizó a medida que nuevos minilotes de datos llegaban, incorporando la información reciente sin olvidar lo previamente aprendido. Los resultados de precisión en las cuatro áreas de prueba se presentan en la Tabla 5.3, donde se observa que el modelo 2 obtiene el mejor desempeño en términos de precisión.

Tabla 5.3: Resultados de precisión con el conjunto de evaluación utilizando tres modelos. El mejor resultado es resaltado en negritas.

Modelo	Área 1	Área 2	Área 3	Área 4
1	0.61	0.72	0.46	**0.67**
2	**0.81**	**0.75**	0.85	0.62
3	0.79	0.68	**0.89**	0.64

Con base en los resultados anteriores, se seleccionó el modelo 2 para la localización topológica debido a los resultados de precisión obtenidos. Así este modelo es utilizado para identificar las posiciones de referencia media en las cuatro trayectorias de vuelo del conjunto de datos de prueba. Esta evaluación se implementó en ROS donde el flujo de imágenes

se procesan a través del modelo de aprendizaje, generando como salida el índice asociado al nodo de la posición media. Los resultados de esta evaluación se presentan en la Tabla 5.4, reportando el número de imágenes correctamente asociadas a su posición, la precisión alcanzada por el modelo y la velocidad de operación expresada en cuadros por segundo (FPS).

Tabla 5.4: Resultados de precisión con el conjunto de datos de evaluación utilizando el segundo modelo.

Trayectoría	Imágenes	Correcto	Precisión	Fps
1	144	89	0.6180	152.397
2	117	73	0.6239	153.044
3	84	70	0.8333	154.096
4	606	378	0.6237	151.564

Esta primera evaluación demostró que el aprendizaje continuo presenta una precisión media de 0.62, siendo aceptable para tareas de relocalización. Si bien este resultado se ve limitado al número de lotes de información que llega, el modelo puede adaptarse evitando el olvido del conocimiento previo utilizando la repetición latente.

5.4.1. Relocalización Topológica

Una de las pruebas más valiosas de esta investigación consiste en determinar la posición de la cámara de un dron durante una misión de vuelo. Los resultados de la sección anterior demostraron que es posible recuperar una posición media de referencia, hacia la cual el vehículo puede dirigirse mientras restablece la señal GPS. No obstante, con el propósito de incrementar la precisión de la localización topológica, se planteó la relocalización del dron en escenarios de pérdida de posición. Para esta evaluación se realizaron pruebas tanto con los tres modelos previamente presentados como con los múltiples modelos entrenados de manera independiente para cada minilote. El objetivo fue integrar la metodología de relocalización, comenzando por la búsqueda de la imagen representativa para identificar el modelo correspondiente a la región de la trayectoria en la que se encuentra el dron.
De esta manera, se evaluó el desempeño de los modelos actualizados y los múltiples modelos entrenados de forma independiente para cada trayectoria. El procedimiento consistió en comparar el descriptor de color a partir del histograma de la imagen con la imagen de referencia para cargar el modelo correspondiente a la posición media de esa región. Además, se incorporó los resultados obtenidos con ORB-SLAM2, aprovechando su módulo

de relocalización para obtener posiciones aproximadas a la referencia. Así, en la Tabla 5.5 se presentan la comparación de los resultados de precisión obtenidos con los tres modelos de repetición latente, los múltiples modelos y el sistema ORB-SLAM2.

Tabla 5.5: Resultados de precisión con el conjunto de datos de prueba utilizando los tres modelos aprendidos, los múltiples modelos y ORB-SLAM2. El mejor resultado es resaltado en negritas.

Enfoque	Tray. 1	Tray. 2	Tray. 3	Tray. 4
Modelo 1	0.618	0.726	0.464	0.671
Modelo 2	**0.815**	0.752	0.857	0.623
Modelo 3	0.791	0.683	**0.892**	0.643
ORB-SLAM2	0.034	**0.913**	0.869	0.584
Multi-Model	0.735	0.817	0.856	**0.714**

La comparación con el sistema ORB-SLAM2 se llevó a cabo generando un mapa a partir del conjunto de entrenamiento, cuya nube de puntos se utilizó como referencia para el proceso de relocalización. Después, el sistema se evaluó con el conjunto de pruebas, intentando relocalizar cada imagen sobre el mapa previamente construido. De esta manera, la relocalización se determinó mediante un umbral de distancia entre la posición media de la imagen representativa y la de la consulta, cuya distancia se encuentra dentro del umbral. Por otro lado, la Tabla 5.6 muestra los resultados de velocidad de procesamiento expresado en FPS.

Tabla 5.6: Resultados de velocidad de procesamiento utilizando los tres modelos aprendidos, los múltiples modelos y ORB-SLAM2. El mejor resultado es resaltado en negritas.

Enfoque	Tray. 1	Tray. 2	Tray. 3	Tray. 4
Modelo 1	101.20	121.16	102.01	105.56
Modelo 2	107.73	127.33	106.16	106.27
Modelo 3	100.42	123.97	96.770	109.21
ORB-SLAM2	85.47	83.33	92.59	89.28
Multi-Model	**150.19**	**151.65**	**153.27**	**153.55**

Este resultado permite evaluar la eficiencia computacional de cada enfoque, proporcionando un panorama más completo entre la precisión y velocidad de procesamiento. Así, se observa que el sistema ORB-SLAM2 presenta dificultades al relocalizar las imágenes de prueba

en el Área 1, debido a que el conjunto de pruebas contiene imágenes no continuas, mientras que el SLAM requiere secuencias consecutivas para construir mapas consistentes. Como consecuencia, el módulo de relocalización no encuentra información suficiente para emparejar correctamente las imágenes. No obstante, en las demás áreas el desempeño mejora mostrando una relocalización más estable al encontrar posiciones aproximadas.

De igual manera, los modelos basados en aprendizaje continuo en combinación con la estrategia de búsqueda y la carga de múltiples modelos, demostraron un rendimiento más preciso en las áreas desafiantes. También, alcanza una velocidad de procesamiento mayor a los enfoques evaluados, lo que representa un balance adecuado entre precisión de localización y eficiencia. Para un mejor entendimiento de los resultados presentados, la Tabla 5.7 muestra una resultado de comparación utilizando la métrica RMSE con los enfoques de ORB-SLAM2, el modelo 2 de aprendizaje, y los múltiples modelos. La métrica indica que tan lejos están los valores obtenidos de los reales, donde un valor menor representa un mejor ajuste a la precisión destacando los resultados obtenidos utilizando múltiples modelos.

Tabla 5.7: Resultados usando la métrica RMSE. El mejor resultado es resaltado en negritas.

Enfoque	Tray. 1	Tray. 2	Tray. 3	Tray. 4
ORB-SLAM2	5.9389	**0.4090**	**1.8563**	8.2484
Modelo 2	2.1730	2.3056	4.4507	7.8642
Multi-Model	**1.8671**	1.6641	2.7080	**7.0858**

Finalmente, se realizó una comparación cualitativa de los resultados obtenidos en las cuatro trayectorias de vuelo utilizando la localización topológica basada en aprendizaje continuo. En la Figura 5.4 se muestran tanto las trayectorias de entrenamiento como las de prueba, donde cada coincidencia correcta entre una imagen de consulta y su posición media de referencia es representada mediante un nodo enlazado con una línea azul. En contraste, cuando la estrategia de búsqueda o el modelo entrenado devuelve una posición incorrecta, esta es marcada en negro y conectada con una línea negra, representando una relocalización fallida. Este resultado integra el proceso metodológico completo, desde la búsqueda de la imagen representativa hasta la asignación final del índice asociado a una posición aproximada de referencia.

De manera complementaria, la Figura 5.5 presenta los resultados obtenidos con el módulo de relocalización de ORB-SLAM2, mostrando las posiciones correctamente relocalizadas sobre la trayectoria de prueba. Además, se observa que las trayectorias 1 y 4 presentan dificultades de relocalización debido a la discontinuidad de las imágenes en ciertos tramos

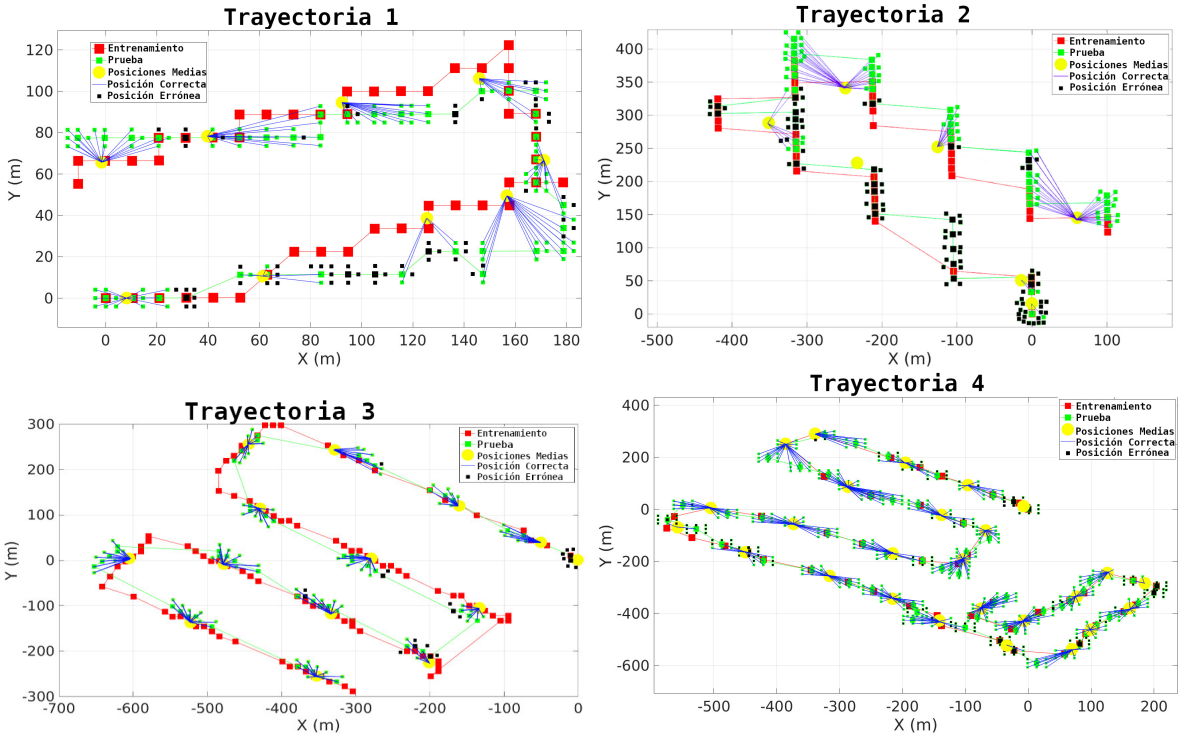

Figura 5.4: Localización topológica utilizando múltiples modelos y aprendizaje continuo con repeticion latente.

del recorrido. No obstante, los resultados demuestran que el sistema SLAM es capaz de recuperar una posición aproximada a la real, siempre que logre detectar un número suficiente de características visuales entre las imágenes de prueba y el mapa construido.

5.5. Sumario

En este capítulo se presentó la metodología de localización topológica basada en la estrategia de reproducción latente de datos, donde se entrenaron modelos asociados a índices de posiciones medias y se evaluaron en trayectorias de prueba. La principal aportación fue la integración del aprendizaje continuo con una estrategia de búsqueda de imágenes representativas, permitiendo recuperar la posición media correspondiente a una imagen de consulta. Los resultados obtenidos mostraron un desempeño adecuado tanto en precisión como en velocidad de procesamiento, siendo adecuado para su uso como un sistema de localización de respaldo en situaciones de pérdida de señal GPS. Asimismo, las comparaciones experimentales y los resultados visuales confirmaron la viabilidad de aplicar

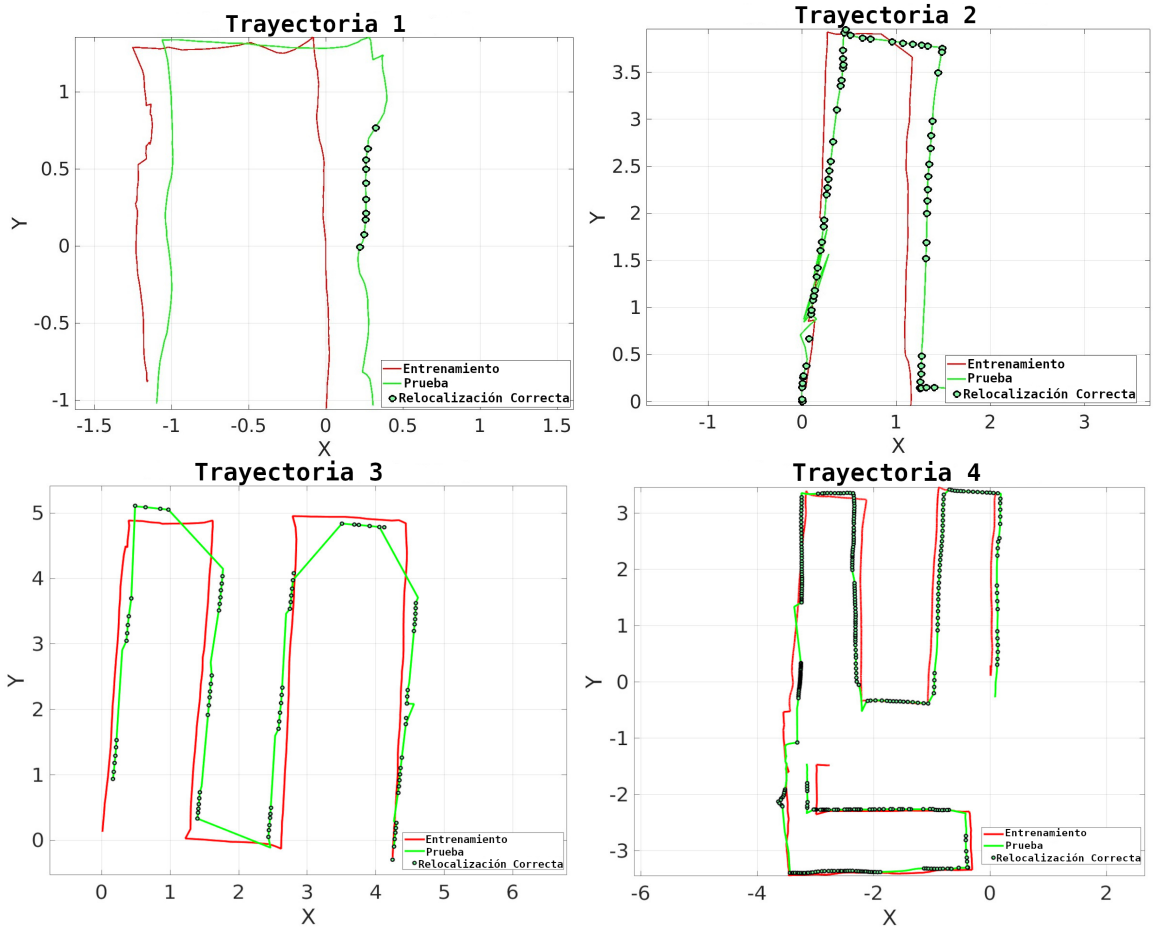

Figura 5.5: Relocalización de las imágenes utilizando ORB-SLAM2.

este enfoque en trayectorias reales utilizando la cámara de un dron. Finalmente, como trabajo futuro se plantea extender la metodología hacia escenarios reales con drones, así como explorar su aplicación en entornos donde la estimación de la posición es ideal.

Capítulo 6

Localización Jerárquica

Este capítulo introduce la metodología de localización jerárquica y los resultados obtenidos a partir de su aplicación en imágenes aéreas capturadas con un dron. Al igual que en el capítulo anterior, se utilizó la estrategia de reproducción latente como técnica de aprendizaje continuo. No obstante, para esta metodología se incorporó la creación de submapas a lo largo de la trayectoria, permitiendo organizar y gestionar múltiples modelos aprendidos por minilotes. La idea de los submapas surge de los sistemas SLAM, los cuales construyen mapas del entorno en diferentes zonas para relocalizar la cámara cuando llega a un punto conocido. Por lo anterior, denominamos a esta metodología localización jerárquica ya que el proceso implementado se realizó en varios niveles desde el entrenamiento continuo de la red, la generación de modelos, la búsqueda del submapa, y la asignación de la posición.

De esta manera, la localización jerárquica se implementó como una estructura modular que integra dos bloques principales: la búsqueda de submapas y la asignación de posición, proceso que permite relacionar la imagen de entrada con el lugar específico donde se encuentra el dron. Igualmente que el capítulo anterior se empleó la red MobileNetV2 para la creación de los múltiples modelos como arquitectura base del entrenamiento. Debido a su resultado en la localización topológica, la arquitectura presenta un equilibrio entre la eficiencia y capacidad de generalización con datos anteriores y nuevos.

Por otro lado, se propone utilizar la red InceptionV4 como extractor de características empleado como el sistema de búsqueda. Esto representa una ventaja al momento de obtener descriptores robustos para la localización de la imagen dentro de la trayectoria realizada. Por lo tanto, una estructura de dos redes con un método de aprendizaje continuo es aplicado en este capítulo, permitiendo la identificación de la región para después obtener la posición de la cámara. Finalmente, la Figura 6.1 presenta la metodología aplicada para obtener la localización jerárquica con imágenes capturadas por un dron durante una misión de vuelo.

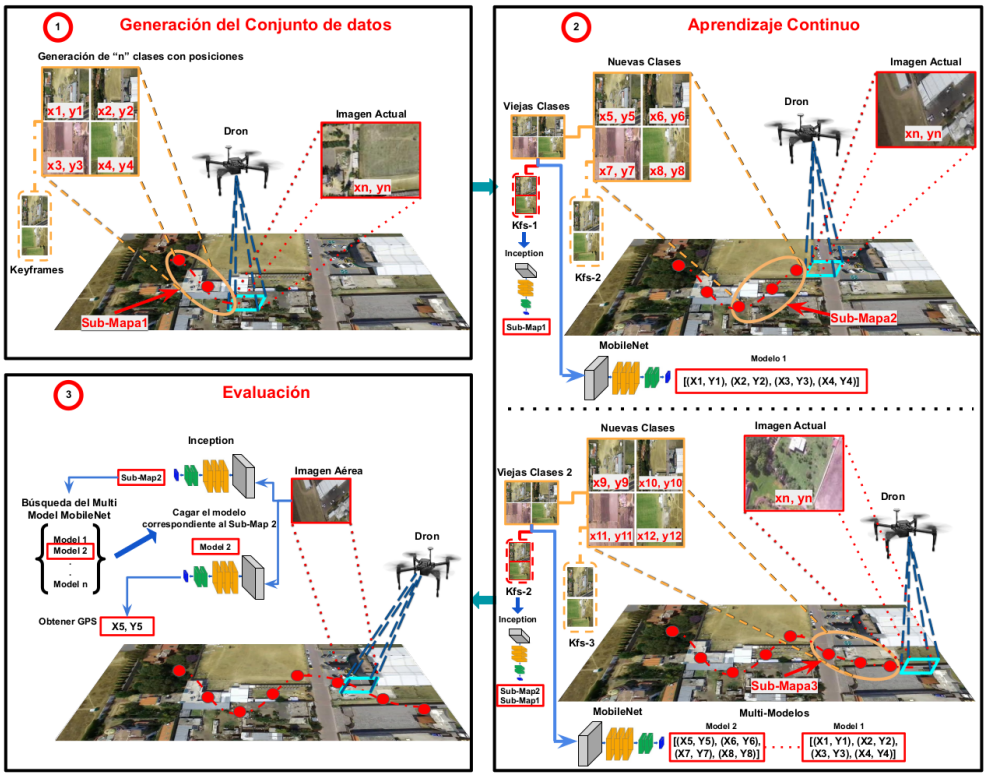

Figura 6.1: Metodología de localización jerárquica con aprendizaje continuo.

6.1. Preparación del Conjunto de Datos

Para la preparación del conjunto de datos en esta metodología se siguió el mismo procedimiento descrito en el capítulo anterior, dividiendo las imágenes en minilotes de entrenamiento. El propósito fue utilizar las imágenes aéreas para asociarlas a posiciones obtenidas mediante la conversión de las coordenadas GPS. No obstante, a diferencia de la preparación del conjunto anterior donde cada minilote integraba más de cinco posiciones medias, en esta ocasión se generaron 50 posiciones distribuidas en 10 submapas a lo largo de toda la trayectoria de vuelo. De esta manera, se cuenta con más información de posiciones intermedias cercanas a la ruta de vuelo realizada por el dron.

Adicionalmente a la generación de minilotes con las imágenes aéreas, se crearon submapas a lo largo de la trayectoria, dividiendo la ruta en varias regiones con agrupaciones de posiciones. De esta manera, cada minilote se asocia a cinco posiciones consecutivas lo que permite obtener localizaciones más cercanas a la ubicación real. Para ilustrar este procedimiento, la Figura 6.2 muestra el proceso de creación y preparación del conjunto de datos utilizado para el entrenamiento. Se observa que por cada cinco posiciones un nuevo submapa es

generado con una distancia de separación de 50 a 100 metros respecto de la última posición registrada.

Figura 6.2: Adquisición del conjunto de datos, generación de submapas e imágenes representativas por cada región a lo largo de la trayectoria.

La creación de los submapas se complementó con la selección de imágenes representativas utilizadas como referencia para guiar el proceso de búsqueda y cargar el modelo correspondiente. Con este fin, se almacenaron hasta tres imágenes representativas por cada submapa, lo que proporciona mayor diversidad en el proceso de emparejamiento cuando se utiliza una imagen de consulta. Este proceso es representado en la Figura 6.2 donde la trayectoria de entrenamiento está representada en color rojo junto con las posiciones capturadas por el GPS del dron. Los cuadros amarillos indican los submapas generados y los cuadros rojos señalan las imágenes representativas almacenadas para el proceso de búsqueda. Por lo tanto, se obtuvieron 47 posiciones a lo largo de la trayectoria 1, 50 posiciones en las trayectorias 2 y 3, y 52 posiciones en la última trayectoria.

6.2. Entrenamiento con Aprendizaje Continuo

La estrategia de aprendizaje continúo empleado en esta etapa siguió el mismo proceso presentado en el capítulo anterior utilizando el método de repetición latente. Sin embargo, en esta propuesta se desarollaron dos entrenamientos paralelos, uno con la red MobileNetV2 y otro con la red InceptionV4. El entrenamiento con MobileNetV2 permitió combinar la

información previa con la nueva de las posiciones a lo largo de la trayectoria, almacenando la información en un memoria externa donde se actualiza el conocimiento adquirido. Para el proceso jerárquico la memoria externa incorporó hasta 50 posiciones distintas distrubuidas en toda la trayectoria, en lugar de utilizar posiciones medias como en la localización topológica. Las 50 posiciones aprendidas se organizaron en minilotes de hasta 200 imágenes, cada uno asociado a un conjunto de cinco posiciones representativas dentro del submapa correspondiente a esa región. En paralelo, se generaron imágenes representativas de los submapas utilizando etiquetas como índices que determinan la región de la trayectoria. La Figura 6.3, ilustra el flujo de entrenamiento utilizando las imágenes para ambas redes, donde MobileNetV2 se entrenó con posiciones agrupadas mientras que InceptionV4 utilizó los índices asociados a las imágenes de los submapas. Como resultado, se obtuvieron 27 imágenes representativas para la trayectoria 1, 30 para las trayectorias 2 y 3, y 52 para la trayectoria 4.

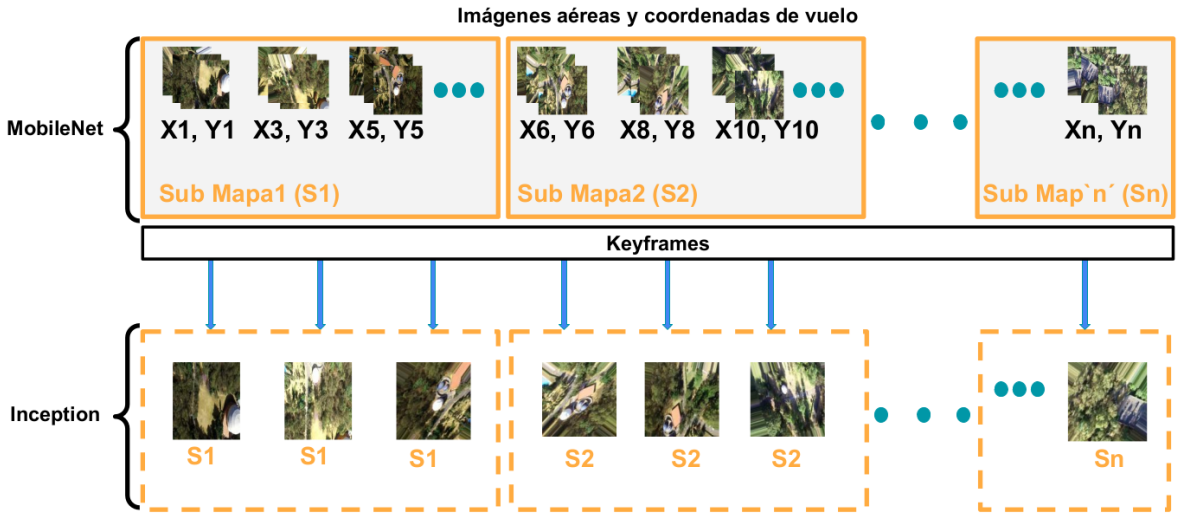

Figura 6.3: Flujo de entrenamiento con MobileNetV2 e InceptionV4 utilizando imágenes con posiciones e imágenes representativas con etiquetas asociadas a submapas.

El entrenamiento continuo de InceptionV4 se llevó a cabo utilizando las imágenes representativas y sus respectivas etiquetas, donde cada etiqueta corresponde a un índice de submapa identificado desde S1 hasta S10. Este proceso fue posible a una modificación en la arquitectura la cual permitió entrenar la red imagen por imagen, encapsulando sus características en un vector de representación. Conforme nuevas imágenes eran incorporadas el vector de patrones se actualizaba con la información reciente, generando a un vector global de características que integra la información de todos los submapas. Este proceso por

niveles lo denominamos entrenamiento jerárquico, permitiendo el aprendizaje de dos redes con información distinta pero complementaria para la adquisición de la posición. Finalmente, la Tabla 6.1 muestra los parámetros utilizados en el entrenamiento de ambas redes.

Tabla 6.1: Parámetros utilizados para el entrenamiento continuo de las redes MobileNetV2 e InceptionV4.

Parámetros	MobileNetV2	InceptionV4
Optimizador	Gradiente Descendiente	Gradiente Descendiente
Tamaño del Lote	128	1
Épocas	20	1
Función de Perdida	Entropía Cruzada	Entropía Cruzada
Tasa de Aprendizaje	0.001	0.001
Capa Latente	Agrupación 6	Vector Temporal
Patrones	3500	1536

6.3. Evaluación Experimental

En este capítulo se realizaron dos experimentos con el propósito de evaluar la metodología de localización jerárquica utilizando las mismas condiciones descritas en la metodología anterior ejecutados en una computadora portátil con Ubuntu 20.04, Python 3.8, PyTorch 1.4 y ROS Noetic. Los resultados se organizaron en dos categorías: 1) La búsqueda correcta del submapa perteneciente a cada imagen de prueba; 2) La adquisición de la posición una vez identificado el submapa. En los primeros pasos, se evaluó la eficiencia de la red InceptionV4 como método de búsqueda, identificando la región correspondiente para las imágenes de prueba. Una vez encontrado el submapa, cada imagen fue asociada a una de las cinco posiciones aproximadas a lo largo de la trayectoria del dron. La evaluación se aplicó a las cuatro trayectorias de vuelo, emparejando las imágenes de prueba con las imágenes representativas de cada submapa.

Para fines de comparación, se utilizó la estrategia de búsqueda mediante histogramas de color previamente empleada en la localización topológica, permitiendo observar la eficiencia de la búsqueda frente a la red InceptionV4. En la Tabla 6.2 se presentan los resultados obtenidos incluyendo el número de imágenes de prueba encontradas por cada estrategia y la precisión alcanzada. Donde la búsqueda basada en la red profunda mostró una mayor capacidad para identificar similitudes de características al comparar las imágenes con las representativas de

los submapas. El proceso de búsqueda genera 1536 patrones de los cuales si una imagen tiene un valor cercano implica una correspondencia con la imagen representativa.

Tabla 6.2: Resultados de precisión con el conjunto de prueba utilizando la búsqueda basada en histogramas de color y la red InceptionV4 para encontrar las imágenes asociadas a sus submapa.

Trayectoria	SubMapa	Histograma		InceptionV4	
		Imágenes	Precisión	Imágenes	Precisión
1	9	87	0.6041	**117**	**0.8125**
2	10	65	0.5555	**84**	**0.7179**
3	10	56	0.6666	**60**	**0.7142**
4	10	310	0.5107	**467**	**0.7693**

De esta manera, los resultados demuestran que el uso de un método basado en aprendizaje profundo proporciona una mayor calidad en la extracción de características visuales, superando las del histograma de color. Así, su uso permite un emparejamiento más robusto entre la imagen de prueba y la imagen representativa del submapa. No obstante, cuando una imagen es asociada a un submapa incorrecto las posiciones quedan descartadas, lo que genera un error en la localización. A pesar de ello, el resultado de búsqueda demuestra ser adecuado para la adquisición de la localización jerárquica, ofreciendo un paso inicial para asignar la posición dentro de una región identificada.

6.3.1. Relocalización Jerárquica

La siguiente evaluación se llevó a cabo utilizando el proceso completo de localización jerárquica, el cual integra tres niveles principales: 1) La búsqueda del submapa a partir de imágenes de prueba: 2) La carga del modelo correspondiente; 3) La asignación de la posición a la imagen. Este proceso tiene un flujo jerárquico donde la imagen aérea capturada por el dron entra en la red InceptionV4, extrayendo las características para compararlas con las imágenes representativas de cada submapa. El resultado de esta comparación genera una etiqueta de índice entre 0 a 9 representando los diez submapas creados en la trayectoria. Después de obtener el índice se carga el modelo asociado al submapa con la información de cinco posiciones entrenadas específicamente en esa región. Posteriormente, el modelo se utiliza nuevamente con la imagen de prueba para asignar una de las cinco posiciones. El proceso de evaluación jerárquica se ilustra en la Figura 6.4, mostrando el flujo de una imagen de prueba al pasar por ambas redes con el objetivo de obtener su localización.

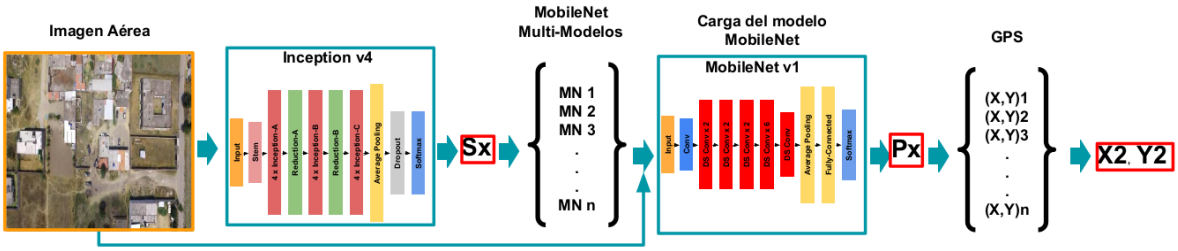

Figura 6.4: Evaluación de imágenes de prueba para la relocalización jerárquica.

De esta manera, la metodología jerárquica se muestra como un proceso de dos niveles, donde el primero identifica el submapa correspondiente y el segundo asigna una posición aproximada con respecto a la trayectoria del dron. Para fines de comparación, se entrenó un modelo por cada trayectoria incorporando todas las posiciones a lo largo de la ruta utilizando el método de repetición latente. Este entrenamiento se realizó con un único modelo mientras se actualizaba el conocimiento con los nuevos datos de posiciones adquiridas en el vuelo. De esta manera, se evaluó si un modelo único, entrenado de forma continua es capaz de preservar las posiciones aprendidas previamente o presenta problemas de olvido catastrófico. Por otro lado, se utilizó nuevamente el módulo de relocalización de ORB-SLAM2, generando un mapa a partir del conjunto de entrenamiento y registrando las posiciones obtenidas en un archivo de texto. Posteriormente, se desactivó el proceso de mapeo y se utilizó el conjunto de prueba, encontrando las coincidencias de características visuales utilizando el módulo. De esta manera, si una imagen de prueba comparte suficientes características con el mapa creado, está se relocaliza estimando una posición cercana tomada de la distancia entre la imagen de prueba y la imagen representativa. También, se implementó la misma metodología de localización jerárquica sustituyendo la búsqueda con InceptionV4 por la búsqueda con histogramas de color, comparando los resultados de todos los enfoques en la Tabla 6.3.

Tabla 6.3: Resultados de precisión para la relocalización utilizando un único modelo, ORB-SLAM2, la metodología jerárquica con histogramas de color y con InceptionV4.

Tray.	Posiciones	Imágenes	Modelo Único	ORB-SLAM2	Histrograma	InceptionV4
1	47	144	0.2500	0.1041	0.3055	**0.7083**
2	50	117	0.2564	0.5897	0.3076	**0.7777**
3	50	84	0.2976	**0.9166**	0.6071	0.8928
4	52	607	0.4382	**0.7166**	0.6690	0.7001

Como se observó en la tabla anterior, los resultados de comparación muestran las posiciones

a lo largo de cada trayectoria, las imágenes del conjunto de prueba y la precisión alcanzada por cada enfoque. Para ello, el resultado corresponde al número de imágenes correctamente relocalizadas a partir de la identificación correcta del submapa y la asignación a su posición más cercana. Además, los resultados muestran que la metodología jerárquica con InceptionV4 mantiene un desempeño consistente, logrando un promedio de precisión de 0.74. Por otra parte, ORB-SLAM2 obtuvo un desempeño bajo en la primera trayectoria, pero obtuvo un valor alto en la trayectoria 3. En contraste, el modelo entrenado de manera continua con todas las posiciones presentó olvido catastrófico, obteniendo un resultado bajo de precisión a lo largo de todas las trayectorias.

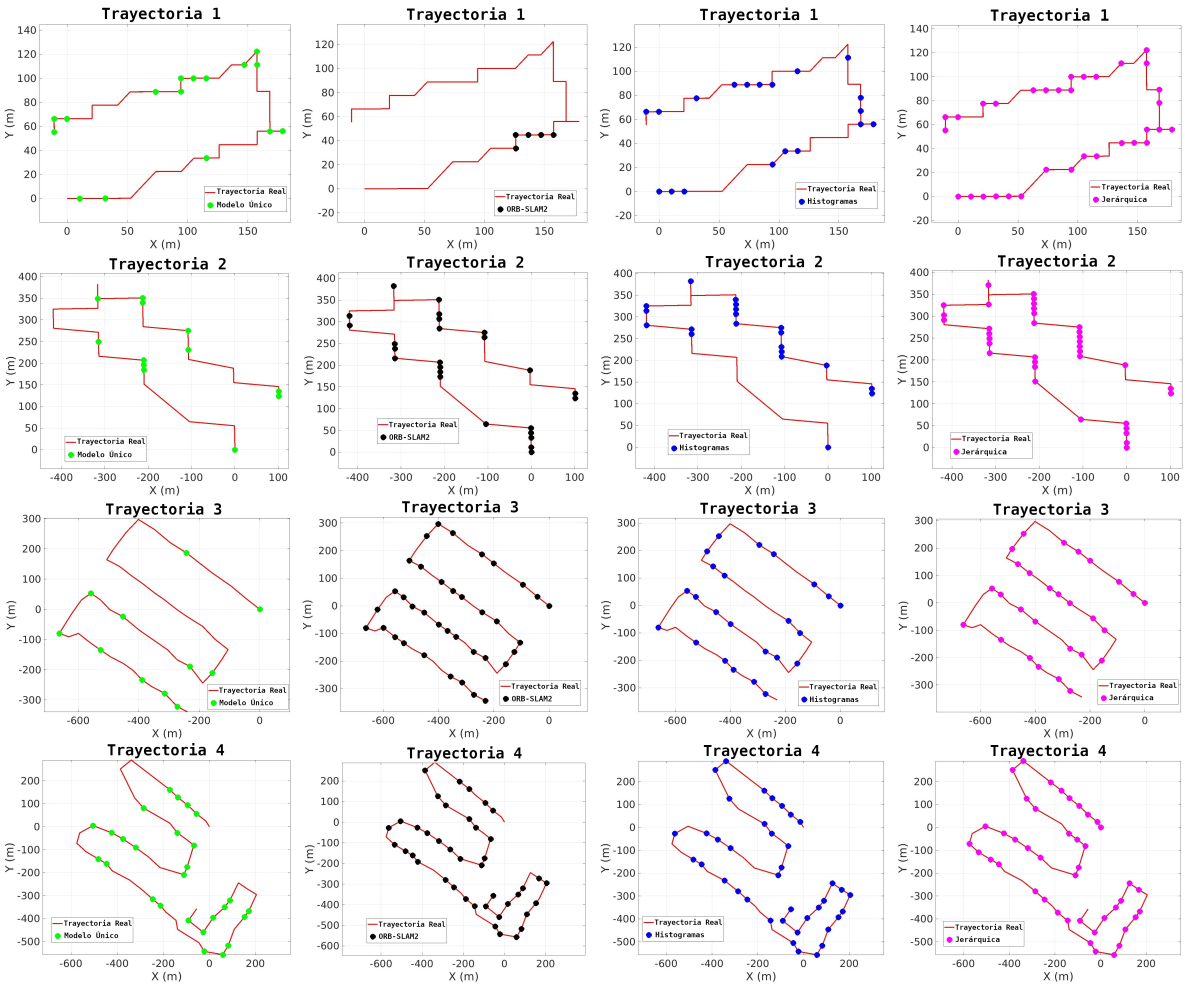

Figura 6.5: Resultado visual de la relocalización con los cuatro enfoques utilizados.

Esto presenta una limitación en el proceso de aprendizaje continuo aplicado a un único modelo que aprende y se actualiza con información de 50 posiciones diferentes. Es decir, que

entre más datos se agreguen al modelo, el olvido catastrófico se presenta en el conocimiento previamente aprendido, llevando a la necesidad de adoptar procesos de entrenamiento compactos y dividir los datos para preservar el conocimiento previo. Por otro lado, la metodología jerárquica propuesta utilizando la búsqueda de submapas con InceptionV4 mostró un mejor desempeño en comparación a la búsqueda con histogramas de color. Una visualización cualitativa de estos resultados se presenta en la Figura 6.5, comparando visualmente las posiciones correctamente relocalizadas con cada uno de los 4 enfoques.

Los resultados visuales obtenidos con los cuatro enfoques muestran el número de posiciones correctamente relocalizadas representadas como puntos de color, donde al no mostrarse un punto en la trayectoria, significa que la posición no fue asignada correctamente. El número total de imágenes correctamente relocalizadas, abarca desde la búsqueda del submapa hasta la asignación de la posición correspondiente. La Tabla 6.4, resume este resultado donde ORB-SLAM2 obtuvo un desempeño alto en la trayectoria 4 pero presenta dificultades en las dos primeras trayectorias debido a la complejidad del escenario para asociar correctamente las imágenes. Además, este resultado se debió a la carencia de información visual limitando la adquisición de la localización. Sin embargo, los enfoques basados en modelo de aprendizaje y búsqueda mediante histograma de color, presentó un rendimiento menor con una cantidad reducida de imágenes relocalizadas.

Tabla 6.4: Imágenes correctamente relocalizadas utilizando cada uno de los cuatro enfoques. El mejor resultado se resalta en negritas.

Método	Tray. 1	Tray. 2	Tray. 3	Tray. 4
ORB-SLAM2	15	69	**77**	**435**
Modelo Único	36	30	25	225
Histograma	44	36	51	266
InceptionV4	**102**	**91**	75	425

Finalmente, la Tabla 6.5 presenta los resultados de la velocidad de procesamiento para cada enfoque expresando en fps. Los resultados indican que ORB-SLAM2 alcanzó un rendimiento de velocidad mayor que la de los otros enfoques. En contraste, el uso de un único modelo y la estrategia de búsqueda con histogramas de color presentan un desempeño similar, manteniendo una velocidad de procesamiento arriba de los 50 fps. Por último, la velocidad alcanzada con la propuesta jerárquica de este estudio, obtuvó una velocidad arriba de los otros pero menor a la del SLAM, esto representa que la metodología puede utilizarse como sistema de localización de respaldo para determinar la posición aproximada de la cámara.

Tabla 6.5: Velocidad de procesamiento en fps con cada método de comparación. La información en negrita muestra el mejor resultado.

Método	Tray. 1	Tray. 2	Tray. 3	Tray. 4
ORB-SLAM2	**85.47**	**83.33**	**92.57**	**89.28**
Modelo Único	55.30	48.28	55.64	62.76
Histograma	59.63	62.40	64.80	61.87
InceptionV4	77.44	68.52	67.63	63.37

6.4. Sumario

En este capítulo se presentó la metodología de localización jerárquica utilizando el método de reproducción latente como estrategia de aprendizaje continuo y en un esquema de búsqueda de submapas. La principal aportación de esta propuesta fue la incorporación de un proceso por niveles que permite determinar la posición de una imagen capturada por la cámara a bordo de un dron. Para ello, se utilizó un entrenamiento dual con dos redes profundas: una para la identificación del submapa o región donde fue captura la imagen y otra enfocada en la asignación de una posición aproximada a la ubicación real. Los resultados obtenidos mostraron un desempeño adecuado para la relocalización de imágenes, recuperando la mayoría de las posiciones a lo largo de las trayectorias de vuelo evaluadas.

La comparación experimental con otros enfoques demostró que el proceso jerárquico es apropiado para emplearse como un sistema de localización de respaldo en situaciones de pérdida de señal GPS. Además, esta metodología puede extenderse a escenarios más complejos y dinámicos, donde el uso de múltiples mapas permita integrar nueva información sin incurrir en el olvido catastrófico. Como trabajo futuro, se plantea aplicar esta estrategia en entornos reales con drones de modo que el sistema pueda localizar al vehículo a partir de las imágenes capturadas durante el vuelo. De igual manera, la integración de un entrenamiento dual permitirá su incorporación en diferentes campos de la robótica enfocados en la localización visual.

Capítulo 7

Localización Progresiva

En este capítulo se presenta una metodología de localización progresiva basada en estrategia de aprendizaje continuo y aplicada a la estimación de posición con imágenes aéreas y terrestres. A diferencia de las metodologías de localización topológica y jerárquica a partir de nodos y submapas, la propuesta progresiva aborda la regresión directa de las posiciones utilizando máquinas de soporte vectorial para regresión (SVR) y redes neuronales binarias (BNNs). El concepto de aprendizaje progresivo surge de los enfoques de aprendizaje continuo basados en arquitectura, donde el conocimiento se expande dinámicamente para integrar nuevas tareas sin olvidar lo previamente aprendido. En contraste de estos métodos que reutilizan o adaptan una red, la estrategia en este capítulo emplea la incorporación progresiva de múltiples modelos donde cada uno amplía el conocimiento hacia nuevas regiones de la trayectoria.

La estructura progresiva presenta algunas ventajas en comparación de las otras metodologías: permite preservar la información previa, facilita la adaptación incremental durante el entrenamiento y permite el uso de múltiples modelos en lugar de un único modelo. En este contexto, la metodología propuesta en este capítulo integra dos enfoques: 1) Estimación de posición con SVR; 2) Estimación de posición con redes binarias (BNN). Estos enfoques se exploraron para reducir la complejidad computacional y aprender las posiciones de manera más directa a partir de características visuales a posiciones espaciales. Finalmente, este capítulo presenta los resultados de la localización progresiva utilizando múltiples modelos y comparando su eficiencia con otros enfoques mostrando su uso en aplicaciones de estimación de posición tanto en entornos aéreos como terrestres.

7.1. Preparación del Conjunto de datos

Para la preparación del conjunto de datos se emplearon los mismos procedimientos descritos en capítulos anteriores, utilizando ROS para la adquisición de imágenes aéreas y posiciones GPS, las cuales fueron convertidas a coordenadas métricas. Al igual que en la metodología jerárquica, se realizaron cuatro trayectorias de vuelo, cada una dividida en diez submapas con longitudes entre 1.0 y 6.0 km. La generación de estos submapas se realizó cuando el dron avanzaba entre 50 y 10 metros respecto a la última posición registrada, obteniendo en promedio entre 5 y 20 posiciones por submapa. Asimismo, para cada submapa se seleccionaron tres imágenes representativas como referencia para la localización de la región donde fue capturada la imagen aérea, teniendo un total de 30 por submapa.

Las imágenes capturadas para los conjuntos de datos se redimensionaron a 224×224 píxeles y asociadas a coordenadas tridimensionales en metros. La Tabla 7.1, presenta el detalle del conjunto de datos utilizado, incluyendo el total de imágenes de entrenamiento, los submapas creado y las imágenes representativas guardadas para cada trayectoria.

Tabla 7.1: Imágenes capturadas para entrenamiento, prueba e imágenes representativas generadas por cada submapa.

Tray.	Imágenes Aéreas		Imágenes Representativas	
	Entrenamiento	Prueba	1 Submapa	10 Submapas
1	282	144	3	30
2	188	117	3	30
3	236	84	3	30
4	7826	607	3	30

Además, el conjunto de datos de entrenamiento y prueba no es continúo debido al propósito metodológico de evaluar el aprendizaje continuo en escenarios no secuenciales. De esta manera permite analizar el desempeño de los modelos cuando se enfrentan a regiones nuevas que no mantienen continuidad con las anteriores. Así, las primeras trayectorias cuentan con número no mayor a 300 imágenes de entrenamiento, mientras que la trayectoria más extensa obtuvimos 7800 con el objetivo de evaluar el método en entornos prolongados. Este conjunto de datos integra todas las imágenes con sus posiciones asociadas, permitiendo el entrenamiento para modelos SVR y redes binarias en tareas de estimación directa de posición en lugar de obtener una posición media o discreta como en metodologías previas.

7.2. Entrenamiento Progresivo

El entrenamiento progresivo surge de la estrategia de expansión del conocimiento dentro del marco del aprendizaje continuo. Este enfoque consiste en modificar la arquitectura de una red neuronal para incorporar capas, integrar subredes auxiliares o el uso de arquitecturas maestro–alumno, permitiendo la destilación del conocimiento hacia bloques internos dedicados a nuevas tareas. De esta manera, la red puede extender gradualmente su entrenamiento sin comprometer lo previamente aprendido. En esta investigación, el aprendizaje progresivo se interpreta como una expansión donde el conocimiento no solo se amplía dentro de una única red, sino que también puede distribuirse a través de modelos complementarios. La Figura 7.1 ilustra la metodología de entrenamiento progresivo utilizando modelos SVR para la localización visual de un dron.

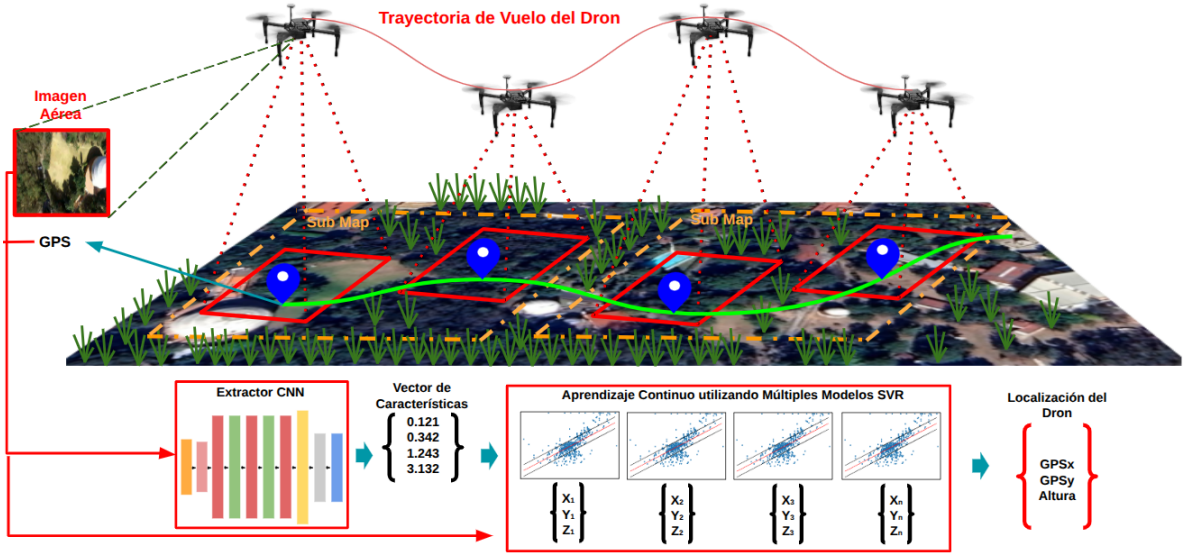

Figura 7.1: Entrenamiento progresivo, donde se extraen características visuales de las imágenes para entrenar tres modelos SVR.

7.2.1. Entrenamiento con Modelos SVR

Para esta fase se llevaron a cabo dos procesos: 1) Entrenamiento de múltiples modelos SVR con información de posición; 2) Entrenamiento de InceptionV4 con imágenes representativas. En el primer proceso, se entrenaron tres modelos SVR independientes, uno por cada coordenada dado su naturaleza unidimensional que permite predecir un único valor numérico

continuo. Para el entrenamiento se utilizó un filtro de función de base radial (RBF), el cual permite modelar relacionales no lineales entre las características y las posiciones. También, se fijó un valor de C=100, donde actúa como parámetro de regularización que controla el equilibro de precisión y la generalización. Se utilizaron como extractores de características las redes DeepPilot4Pose y ResNet18, cuyas salidas son usadas como vectores de entrada a los modelos SVR.

De este modo, las etiquetas de posición fueron asociadas a cada vector de características lo que permite al modelo aprender representaciones más directas entre la información visual y las coordenadas espaciales. El segundo proceso empleó el método de reproducción latente permitiendo agregar una memoria externa para retener información de los datos anteriores e integrar la nueva información. Así, se definió un vector de 10 dimensiones donde se almacenaron las características de las imágenes correspondientes a los submapas. El proceso extrae las características con InceptionV4 para después compararlas con la información almacenada en el vector de información. De esta manera se identifica el submapa correspondiente a la imagen de consulta. La Figura 7.2 ilustra el procedimiento del entrenamiento progresivo con los modelos SVR y el entrenamiento con InceptionV4 para la búsqueda de submapas en función de las imágenes generadas en cada región.

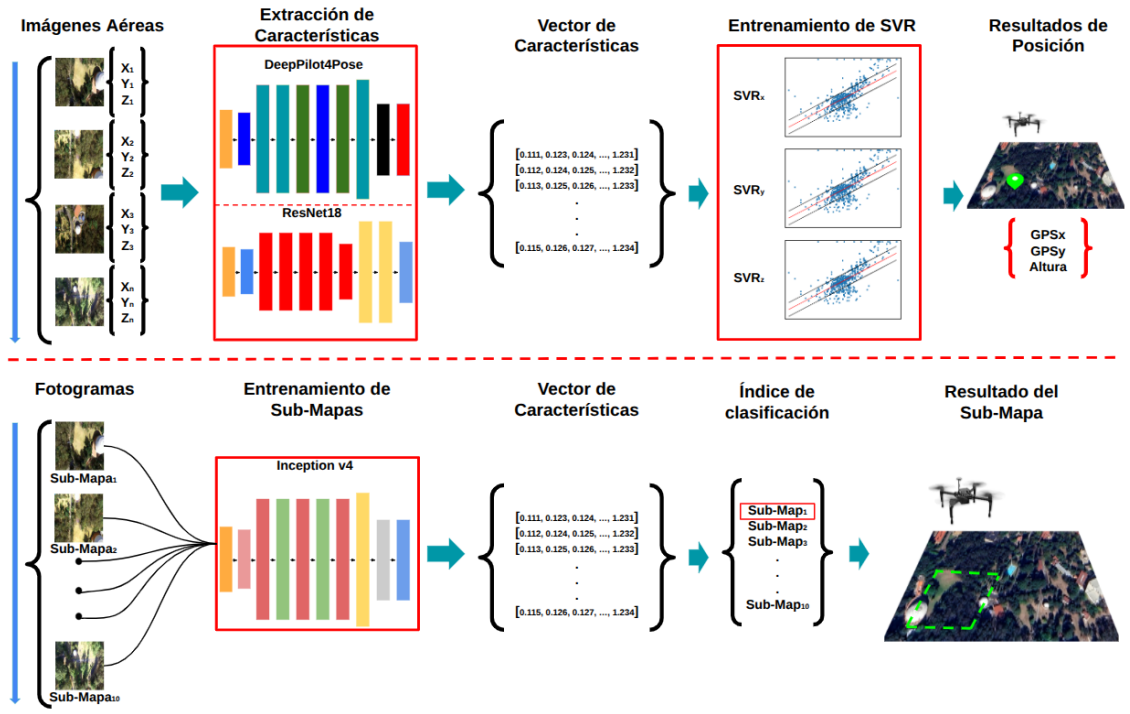

Figura 7.2: Entrenamiento progresivo: 1) Entrenamiento de múltiples modelos SVR con información de posición; 2) Entrenamiento de InceptionV4 con información de los submapas.

7.2.2. Entrenamiento con Redes Binarias

El entrenamiento progresivo con redes binarias se realizó bajo el mismo principio aplicado en los modelos SVR con dos procesos principales: 1) Entrenamiento basado en estrategia de expansión de conocimiento con el método de aprendizaje progresivo; 2) Entrenamiento basado en la reproducción latente de los datos. En el primer proceso cada modelo binario fue entrenado con imágenes aéreas y sus respectivas posiciones tridimensionales como etiquetas de salida, lo que permitió que la red aprendiera a estimar coordenadas continuas a partir de características visuales. El segundo proceso, se utilizaron las imágenes representativas generadas en cada submapa para entrenar la red InceptionV4, cuya función fue identificar el índice del submapa correspondiente a una imagen de consulta y así determinar la región aproximada en la que se encuentra el dron. Como se explicó en capítulos anteriores, este entrenamiento se llevó a cabo imagen por imagen, extrayendo tanto características máximas como mínimas almacenadas en un vector de 10 dimensiones representado los distintos submapas. La Figura 7.3 ilustra la metodología general adoptada para el entrenamiento progresivo con redes binarias, integrando ambos procesos en un esquema general.

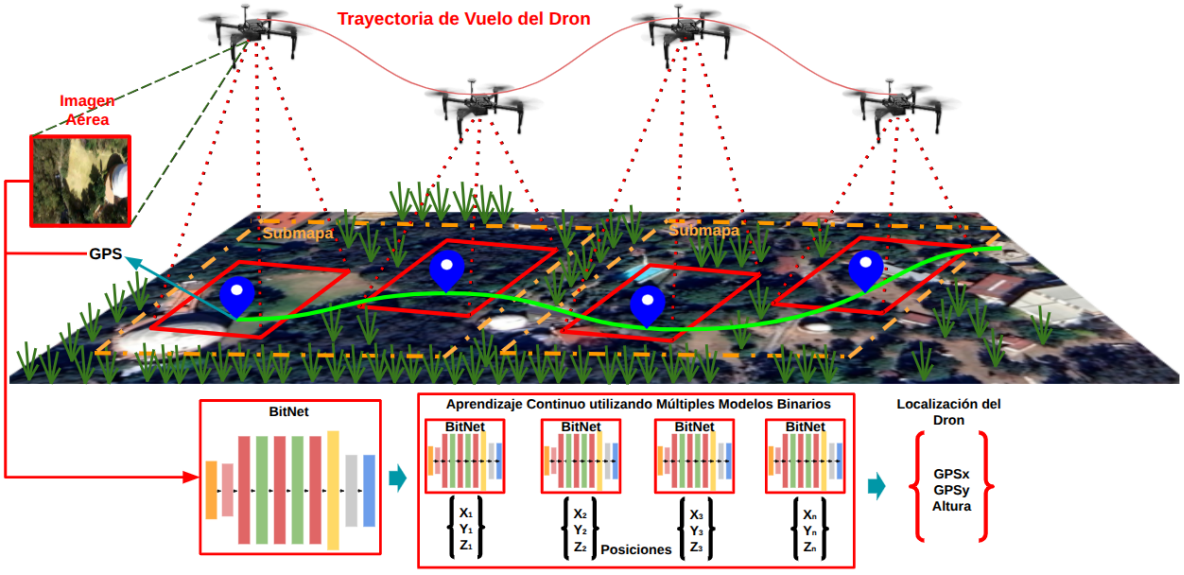

Figura 7.3: Entrenamiento progresivo, donde se extraen características visuales de las imágenes para entrenar múltiples modelos binarios (BNN).

La red binaria base utilizada fue presentada por Wang et al. (2023) pero realizando modificaciones en las capas profundas y de salida con el objetivo de adaptarla a tareas de regresión. En particular, se sustituyeron las capas originales por capas convolucionales

diseñadas para tareas de regresión y estimación. La arquitectura final consiste de dos capas convolucionales binarias personalizadas llamadas "bitConv2d", seguidas de una función de activación ReLU y una capa de agrupamiento máximo. Después, se incorporaron dos capas totalmente conectadas que generan la salida del modelo en un vector tridimensional que representa la posición estimada de la cámara y complementado con información de orientación. En la Figura 7.4 se ilustra el flujo de esta arquitectura, desde la entrada hasta obtención del vector final de predicción. Así la red constituye una arquitectura ligera de cuatro capas convolucionales binarizadas adecuada para tareas de estimación de posicion y localizacion visual.

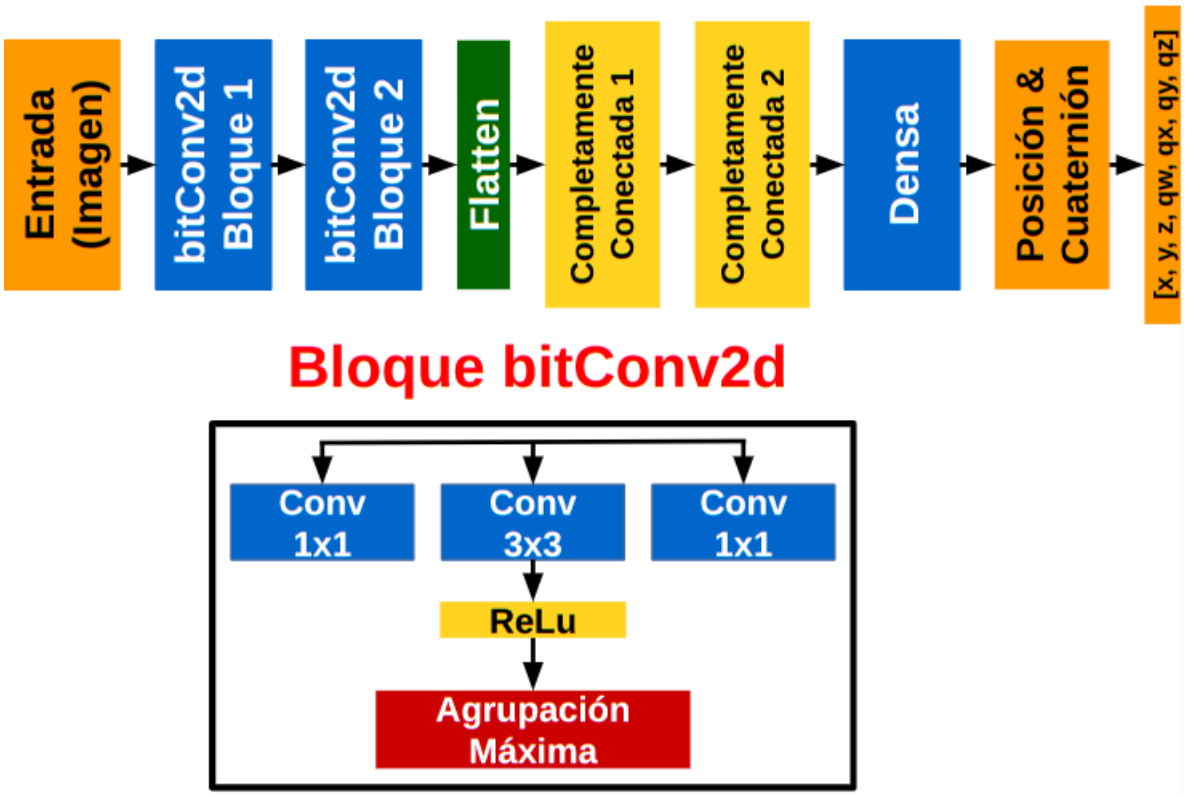

Figura 7.4: Arquitectura de BitNet: Consiste de cuatro capas, dos bloques de bitConv2d y dos capas completamente conectada.

Así, el entrenamiento progresivo consistió en expandir el conocimiento en múltiples modelos entrenadas por cada submapa. Sin embargo, para fines de evaluación se entrenó un modelo único el cual se actualizó progresivamente con toda la información de imágenes y posiciones. En ambos procesos, se utilizó 100 épocas de entrenamiento con el optimizador Adam y una tasa de aprendizaje de 0.001, tanto para los múltiples modelos como el modelo único.

Dado que las capas binarias representan un avance reciente en la literatura, se utilizaron las últimas versiones de CUDA a 12.2 y PyTorch 2.3, ejecutados en la misma computadora empleada en las metodologías anteriores. Finalmente, en la Figura 7.5 se presenta el flujo del entrenamiento propuesto donde cada imagen es procesada para identificar el submapa correspondiente y posteriormente estimar la posición de la cámara abordo del dron. Este enfoque progresivo con redes binarias ofrece un esquema ligero y adaptable para la relocalización en escenarios aéreos.

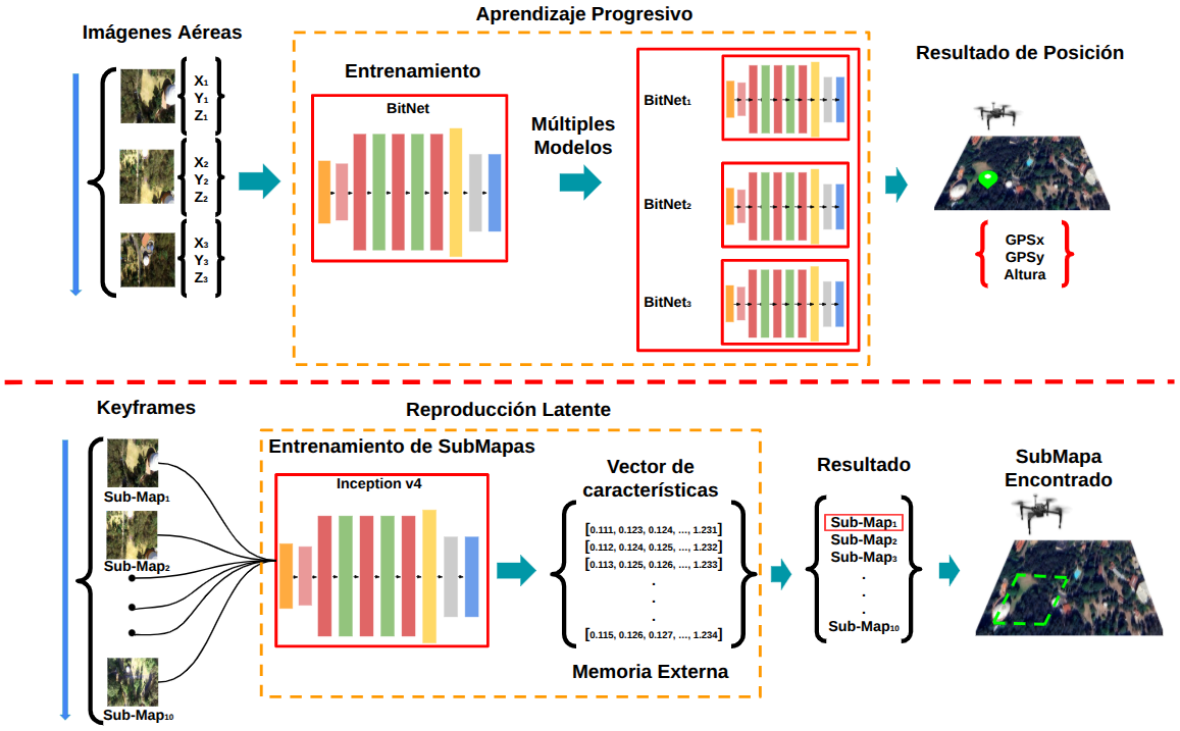

Figura 7.5: Metodología de entrenamiento para la localización progresiva a partir de la identificación submapas y la estimación de posición utilizando la red binaria correspondiente.

7.3. Evaluación Experimental

En esta sección se presentan los experimentos y resultados obtenidos con modelos SVR y redes binarias para la estimación de posición de una cámara aplicando aprendizaje progresivo. El primer experimento consistió en la búsqueda de submapas, retomando la idea de la metodología jerárquica donde previamente se evaluó el la búsqueda utilizando el enfoque de histogramas y la red Inceptionv4. En esta ocasión, además de los previos enfoques, se utilizó

las redes MobileNetV2 y ResNet18 como extractores de características complementarios, con el fin de fortalecer la identificación de las imágenes de prueba. Los resultados de este proceso se presentan en la Tabla 7.2 donde se muestra el desempeño de cada red en la búsqueda de submapas a partir del conjunto de imágenes de prueba.

Tabla 7.2: Resultados de precisión con el conjunto de prueba utilizando la búsqueda con las redes MobileNetV2 y ResNet18 para encontrar las imágenes asociadas a los submapa.

Trayectoria	Imágenes	MobileNetV2		ResNet18	
		Encontradas	Precisión	Encontradas	Precisión
1	144	96	0.66	**104**	**0.75**
2	117	71	0.60	**77**	**0.65**
3	84	55	0.65	**62**	**0.73**
4	607	381	0.62	**427**	**0.70**

El uso de redes profundas como esquema de búsqueda demostró ser efectivo tanto en la calidad de la información extraída como en el rendimiento del proceso de emparejamiento de imágenes. Los resultados anteriores muestran que, aunque ResNet18 superó en precisión a MobileNetV2, la red InceptionV4 sigue alcanzado un desempeño superior con una precisión media de 0.75. Este resultado se explica por la capacidad de capturar tanto características máximas como mínimas a través de sus módulos Incepción lo que permite representar de manera más robusta las imágenes de prueba y asociarlas con el submapa correcto. En constraste, MobileNetV2 y ResNet18 al ser arquitecturas más ligeras, ofrecen descriptores menores lo cual limita su rendimiento en tareas de búsqueda.

7.3.1. Estimación de Posición con SVR

Para la estimación de posición con SVR a partir de las imágenes de prueba, se ejecutó el proceso completo de localización progresiva. Este procedimiento inicia con la búsqueda del submapa correspondiente, lo que permite identificar los modelos previamente entrenados en esa región específica. Así, se realiza la extracción de características visuales de la imagen de prueba y se compara con las imágenes representativas almacenadas. Una vez localizada la imagen más similar, se asigna el índice del submapa asociado indicando qué modelos SVR deben ser cargados para la estimación.

Esta estimación se realiza con cada imagen de prueba donde se procesan inicialmente con las redes DeepPilot4Pose o ResNet18, obteniendo vectores de características de 1024 y 512 dimensiones. Estos vectores se utilizan como entrada para los tres modelos SVR,

estimando la posición asociada a la imagen y por lo tanto la ubicación del dron. El procedimiento se repite para todas las imágenes de prueba, produciendo una serie de estimaciones que determina la localización aproximada del vehículo. La Figura 7.6, ilustra el flujo de procesamiento desde la extracción de características hasta la estimación final.

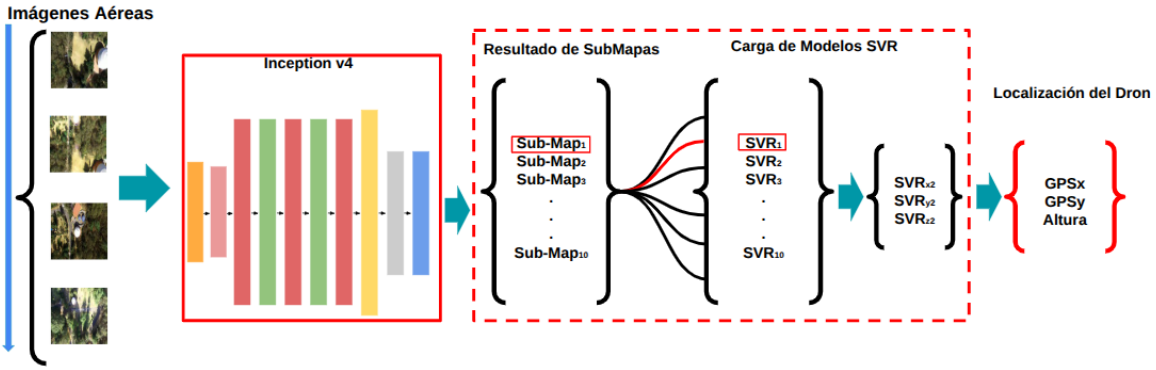

Figura 7.6: Estimación de posición utilizando modelos SVR a partir de la búsqueda del submapa correspondiente.

Para fines de comparación, se evaluó la estimación de posición mediante diferentes enfoques: 1) Red neuronal (NN) de 4 capas utilizando un modelo único que concentra todo el aprendizaje y múltiples modelos entrenados por submapas; 2) PoseNet; 3) ORB-SLAM2 para localizar las imágenes de prueba y estimar la posición correspondiente; 4) Modelos SVR tanto con tres modelos únicos (x,y,z) como la de múltiples modelos con la búsqueda de submapas. Los resultados obtenidos se presentan en la Tabla 7.3, reportando el error medio de distancia euclidiana en metros utilizando DeepPilot4Pose como extractor de características.

Tabla 7.3: Error medio de distancia euclidiana en metros utilizando DeepPilot4Pose como extractor de características. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1			Trayectoria 2			Trayectoria 3			Trayectoria 4		
	x	y	z	x	y	z	x	y	z	x	y	z
PoseNet	47.4	15.3	5.07	94.1	82.3	12.7	59.9	38.3	19.0	146.1	130.8	2.09
ORB-SLAM2	-	-	-	**13.04**	**10.68**	6.90	-	-	-	-	-	-
NN - Único	48.0	18.8	11.3	104.0	93.7	18.0	54.2	45.5	11.2	158.5	136.6	16.8
NN - Múltiple	46.4	26.7	23.5	82.51	100.8	30.9	77.2	32.3	25.8	94.86	87.51	36.1
SVR - Único	43.0	15.9	0.16	101.4	85.8	0.16	57.9	52.1	0.21	143.2	130.3	0.27
SVR - Múltiple	**5.85**	**3.41**	**0.11**	16.77	19.1	**0.050**	**14.6**	**12.1**	**0.05**	**27.1**	**29.6**	**0.22**

De manera similar, en la Tabla 7.4 se reportan los resultados expresados como porcentaje de error donde un valor más alto indica una mayor error en la estimación de posición, o por

lo tanto en asignación correcta del submapa. El resultado muestra el porcentaje de error por coordenada donde el eje z presenta los valores más bajos debido a que todas las posiciones fueron constantes a la misma altura de vuelo. En contraste, los errores para los x,y son más elevados atribuidos a un entrenamiento limitado en términos de cantidad o quizá al olvido catastrófico aún presente. No obstante, el enfoque propuesto con múltiples SVR mantiene un menor porcentaje de error en comparación con los enfoques alternativaos, demostrando la capacidad para adquirir posiciones cercanas a los reales.

Tabla 7.4: Porcentaje de error por coordenada utilizando DeepPilot4Pose como extractor de características. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1			Trayectoria 2			Trayectoria 3			Trayectoria 4		
	x	y	z	x	y	z	x	y	z	x	y	z
PoseNet	48.7	30.8	9.91	60.4	43.0	12.7	19.0	33.6	18.9	76.9	53.2	2.11
ORB-SLAM2	-	-	-	16.6	**5.74**	7.56	-	-	-	-	-	-
NN - Único	49.3	37.8	22.4	66.8	49.0	17.9	17.1	39.9	11.1	83.4	55.5	16.7
NN - Múltiple	47.7	53.7	46.6	52.9	52.7	30.8	24.4	28.4	25.7	49.9	35.5	35.9
SVR - Único	44.2	31.9	0.32	65.1	44.9	0.16	18.3	45.9	0.21	75.3	53.0	0.27
SVR - Múltiple	**6.02**	**6.85**	**0.22**	**10.7**	9.99	**0.05**	**4.64**	**10.6**	**0.05**	**14.2**	**12.0**	**0.22**

Finalmente, en la Tabla 7.5 se reportó el error porcentual total, calculado como la suma del error de traslación acumulado a lo largo de cada trayectoria. La metodología propuesta obtuvo valores entre 4,7 % y 10,5 %, lo cual resulta adecuado considerando que el entrenamiento se realizó con menos de 300 imágenes y que los datos fueron divididos. En contraste, PoseNet presentó un error superior previsto por el número reducido de ejemplos, así como la discontinuida de las posiciones entrenadas. En esta ocasión ORB-SLAM2 no logro completar el mapeo en 3 trayectorias ocasionando que no se pudiera adquirir la posición estimada para cada imagen, pero si alcanzando un desempeño favorable en la trayectoria 2.

Tabla 7.5: Porcentaje total de error utilizando DeepPilot4Pose como extractor de características. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1	Trayectoria 2	Trayectoria 3	Trayectoria 4
PoseNet	34.3	42.2	22.1	52.0
ORB-SLAM2	-	11.7	-	-
NN - Único	39.5	48.2	20.9	58.1
NN - Múltiple	49.0	47.8	25.5	40.7
SVR - Único	29.9	41.9	20.8	51.0
SVR - Múltiple	**4.75**	**8.03**	**5.06**	**10.5**

En la Figura 7.7 se presentan los resultados visuales obtenidos con tres métodos de

estimación de posición. En la primera columna se muestra el desempeño de PoseNet, una red utilizada en la literatura para localización presentando saltos abruptos en la estimación de posición con imágenes de prueba. La segunda columna se muestran los resultados de nuestro enfoque utilizando la red neuronal de 4 capas, logrando estimaciones estables en un inicio, pero con fallos hacia el final de las trayectorias. Finalmente, la tercera columna se observa nuestro enfoque multimodelo con SVR con posiciones estimadas más alineadas con las reales, demostrando un desempeño superior frente a los otros enfoques.

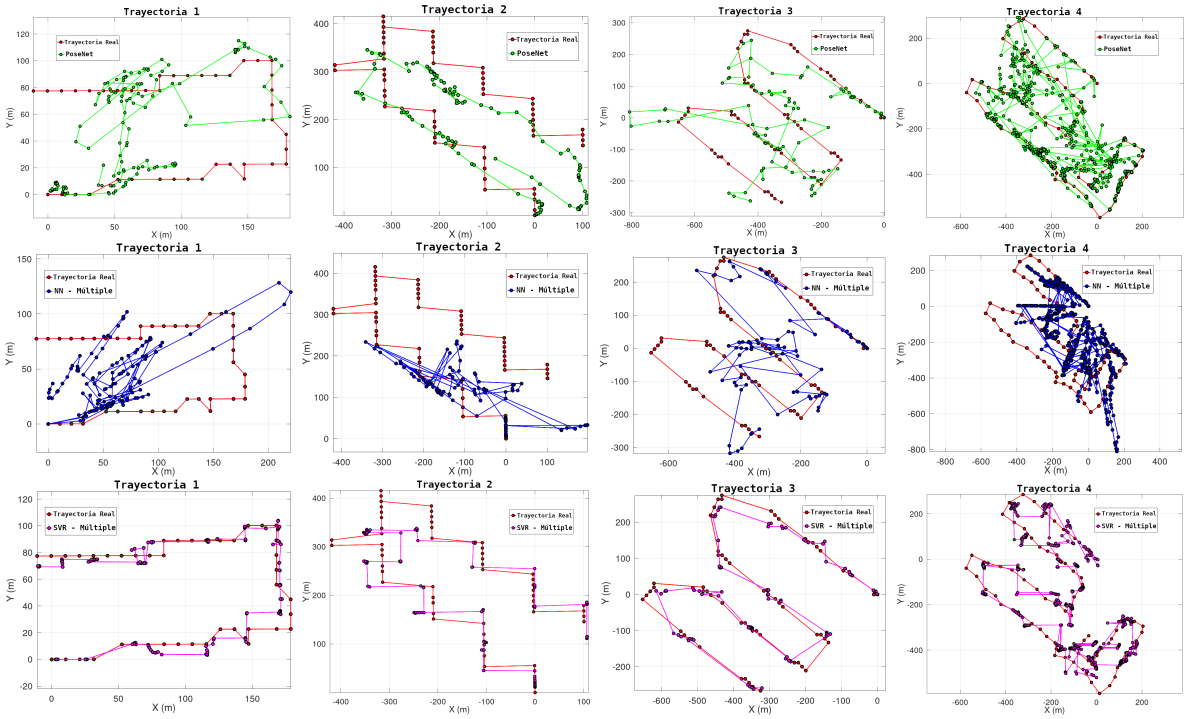

Figura 7.7: Estimación de posición utilizando DeepPilot4Pose como extractor de características, presentando las estimaciones con: PoseNet, NN-Múltiple y SVR-Múltiple.

En la Tabla 7.6 se reportan los errores medios de distancia euclidiana utilizando la red ResNet18 como extractor de características, donde nuestro enfoque se mantiene consistente en rendimiento que los alternativos. La Tabla 7.7 presenta el porcentaje de error por coordenada en la cual el enfoque SVR-múltiple mantiene un porcentaje menor en todas las trayectorias evaluadas. En comparación con los resultados anteriores, ResNet18 ofrece una representación de características más robusta y de mayor rendimiento tanto para los modelos SVR como para las redes neuronales, reduciendo la dispersión de las predicciones. Esto se evidencia en la Tabla 7.8 donde se presenta una ligera reducción en el porcentaje de error total con valores entre 4.65 % y 10.1 %.

Tabla 7.6: Error medio de distancia euclidiana en metros utilizando ResNet18 como extractor de características. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1			Trayectoria 2			Trayectoria 3			Trayectoria 4		
	x	y	z	x	y	z	x	y	z	x	y	z
PoseNet	47.4	15.3	5.07	94.1	82.3	12.7	59.9	38.3	19.0	146.1	130.8	2.09
ORB-SLAM2	-	-	-	**13.04**	**10.68**	6.90	-	-	-	-	-	-
NN - Único	39.4	9.72	0.10	82.09	75.7	5.80	34.4	37.9	6.06	140.7	120.6	7.30
NN - Múltiple	12.8	10.9	4.07	45.8	48.47	13.5	27.8	21.5	7.05	52.64	52.75	23.2
SVR - Único	35.3	9.59	0.18	87.9	74.0	0.20	28.1	39.7	0.22	141.2	119.2	0.27
SVR - Múltiple	**6.12**	**3.85**	**0.10**	19.61	17.2	**0.04**	**13.6**	**10.9**	**0.04**	**25.22**	**29.35**	**0.23**

Tabla 7.7: Porcentaje de error por coordenada utilizando ResNet18 como extractor de características. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1			Trayectoria 2			Trayectoria 3			Trayectoria 4		
	x	y	z	x	y	z	x	y	z	x	y	z
PoseNet	48.7	30.8	9.91	60.4	43.0	12.7	19.0	33.6	18.9	76.9	53.2	2.11
ORB-SLAM2	-	-	-	16.6	**5.74**	7.56	-	-	-	-	-	-
NN - Único	40.5	19.5	5.54	52.6	39.6	5.70	10.9	33.3	6.03	74.0	49.0	7.26
NN - Múltiple	13.1	22.0	8.07	29.4	25.3	13.4	8.82	18.8	7.02	27.7	21.4	12.1
SVR - Único	36.3	19.2	0.36	56.4	38.9	0.20	8.91	34.9	0.22	74.3	48.5	0.24
SVR - Múltiple	**6.29**	**7.75**	**0.21**	**12.5**	17.2	**0.04**	**4.34**	**9.58**	**0.04**	**13.2**	**11.9**	**0.04**

Los resultados obtenidos confirman que ResNet18 es capaz de extraer características visuales de alta calidad más adecuadas para el entrenamiento que las obtenidas con DeepPilot4Pose. Al utilizar estas representaciones el enfoque de NN-Múltiple logra estimaciones más cercanas a las reales, reduciendo el error en las estimaciones incorrectas con respecto a la trayectoria de prueba. De manera similar, nuestro enfoque SVR-Múltiple presenta una ligera mejora en la precisión, permitiendo una localización más estable alrededor de los submapas encontrados. Sin embargo, para alcanzar resultados más fluidos y evitar los saltos aún observados es necesario entrenar con un conjunto de datos constante para incrementar la diversidad del

Tabla 7.8: Porcentaje total de error utilizando ResNet18 como extractor de características. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1	Trayectoria 2	Trayectoria 3	Trayectoria 4
PoseNet	34.3	42.2	22.1	52.0
ORB-SLAM2	-	11.7	-	-
NN - Único	26.3	46.5	14.8	50.0
NN - Múltiple	14.1	24.0	10.6	21.9
SVR - Único	22.8	36.3	12.8	48.6
SVR - Multiple	**5.10**	**8.24**	**4.65**	**10.1**

escenario en relación a sus posiciones. Finalmente, la Figura 7.8 presenta las posiciones estimadas utilizando ResNet18 como extractor de características y comparando los resultados con PoseNet, NN-Múltiples y nuestro enfoque de SVR-Múltiple.

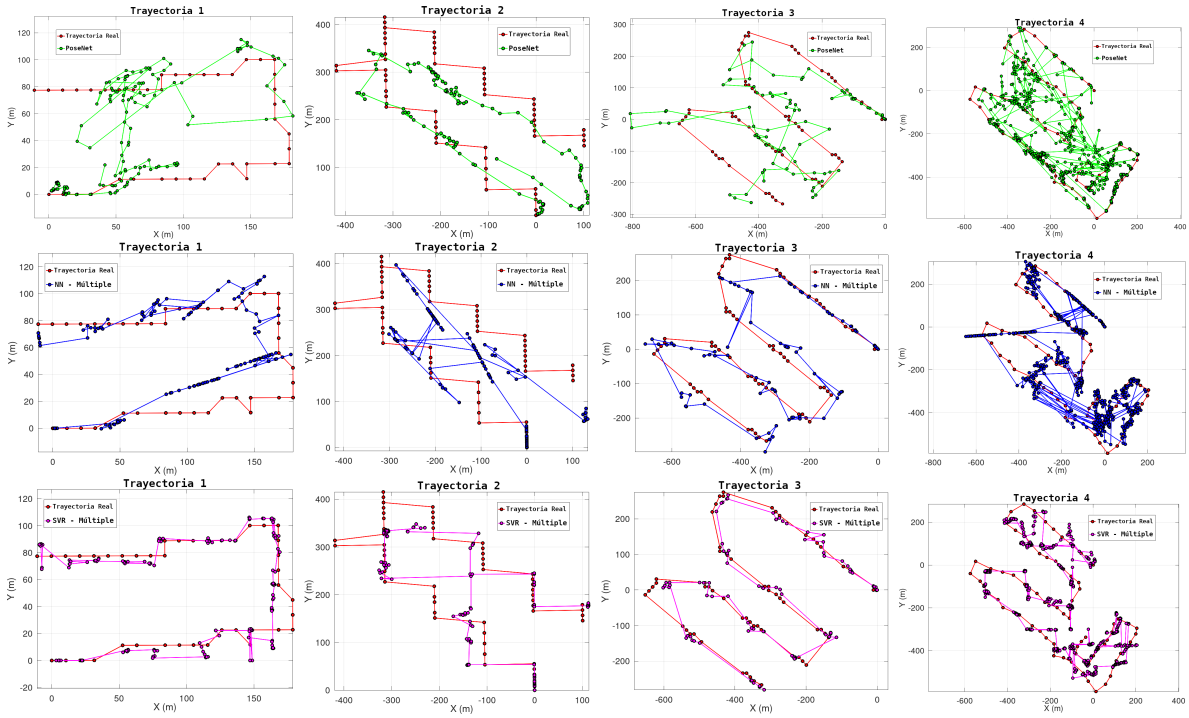

Figura 7.8: Estimación de posición utilizando ResNet18 como extractor de características, presentando las estimaciones con: PoseNet, NN-Múltiple y SVR-Múltiple.

Para evaluar la eficiencia de la metodología propuesta se realizaron dos pruebas de tiempo de procesamiento expresadas en milisegundos (ms). Así, se comparó el tiempo de extracción de características utilizando ORB-SLAM2, DeepPilot4Pose y ResNet18 con el fin de determinar la rapidez en la generación de descriptores visuales. Del mismo modo, se tomó el tiempo de localización total la cual consiste desde la identificación del submapa hasta la estimación final de la posición mediante localización progresiva, proporcionando información sobre la eficiencia general de la metodología.

La Tabla 7.9 presenta los resultados del tiempo de extracción de características bajo condiciones controladas utilizando ROS y evaluando imagen por imagen a través de un nodo de procesamiento. La medición de las redes profundas se realizó desde el momento en que la red genera el vector de características. En contraste, para ORB-SLAM2 el tiempo corresponde al proceso de extracción de descriptores visuales obtenidos directamente de las imágenes. A pesar de la diferencia de procesos, incluimos el sistema SLAM para ofrecer un

punto de vista base en relación al tiempo de extracción. Este resultado permite observar cuánto tiempo consume una red compacta en la extracción de características y compararlo con un método tradicional basado en geometría.

Tabla 7.9: Tiempo medio de extracción de características utilizando ORB-SLAM2, DeepPilot4Pose y ResNet18 como extractores.

Extractor	Trayectoria 1	Trayectoria 2	Trayectoria 3	Trayectoria 4
ORB-SLAM2	77.17	70.94	72.97	74.90
DeepPilot4Pose	54.55	55.10	55.65	58.07
ResNet18	**43.35**	**39.91**	**41.65**	**40.09**

Finalmente, para cerrar la evaluación de la localización progresiva con modelos SVR, en las Tablas 7.10 y 7.11 se muestra el tiempo total de procesamiento considerando todos los enfoques de esta sección. Los resultados muestran que el tiempo requerido para localizar una imagen es menor cuando se utiliza un SVR único, es decir utilizando tres modelos (x,y,z) entrenados continuamente con todas las posiciones. En contraste, el enfoque con múltiples modelos presenta un tiempo ligeramente mayor, ya que una vez identificado el submapa es necesario cargar el modelo correspondiente antes de estimar la posición. Resultados similares se observaron con el uso de redes neuronales con modelo único y múltiples modelos, demostrando ser igual de rápidas que los modelos SVR lo que permite obtener una posición para localizar al dron a partir de la imagen capturada.

Tabla 7.10: Tiempo total de procesamiento (ms) para loa localización utilizando DeepPilot4Pose como extractor. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1	Trayectoria 2	Trayectoria 3	Trayectoria 4
ORBSLAM2	-	79.12	-	-
PoseNet	71.94	73.74	73.52	72.15
NN - Único	58.59	58.41	53.74	54.01
NN - Múltiple	60.75	62.08	61.06	60.90
SVR - Único	**49.52**	**51.13**	**49.90**	**51.10**
SVR - Múltiple	64.20	63.33	69.21	71.53

En conclusión, el uso de aprendizaje continuo con un método de expansión de conocimiento con múltiples modelos demostró ser eficiente tanto en precisión de localización como en velocidad de procesamiento. No obstante, aún con estos avances queda la pregunta si se puede aplicar a computadoras a bordo de un dron real. Para ello, es necesario explorar

Tabla 7.11: Tiempo total de procesamiento (ms) para la localización utilizando ResNet18 como extractor. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1	Trayectoria 2	Trayectoria 3	Trayectoria 4
ORBSLAM2	-	79.12	-	-
PoseNet	71.94	73.74	73.52	72.15
NN - Único	56.01	50.25	45.22	49.82
NN - Múltiple	61.45	61.43	64.60	64.18
SVR - Único	**52.92**	**40.65**	**41.51**	**53.91**
SVR - Múltiple	70.53	64.37	64.63	75.15

arquitecturas más ligeras como las redes binarias, que reducen la capacidad de cómputo, pero manteniendo una precisión adecuada en los resultados.

7.3.2. Estimación de Posición con BitNet

Para la estimación de posición se utilizó la arquitectura BitNet, modificada específicamente para tareas de regresión en las 4 trayectorias de vuelo. La evaluación consistió en el proceso completo de localización: desde la búsqueda del submapa hasta la carga del modelo correspondiente de BitNet y estimación de la posición. Para la etapa de búsqueda utilizamos InceptionV4 como extractor de características, permitiendo identificar el submapa correspondiente y obtener el índice que determina qué modelo debía cargarse. Después, la misma imagen de entrada fue procesada por el modelo seleccionado para estimar la posición final y obtener la localización de la cámara. La Figura 7.9 ilustra este flujo de procesamiento, desde la identificación inicial del submapa hasta la estimación final de la posición.

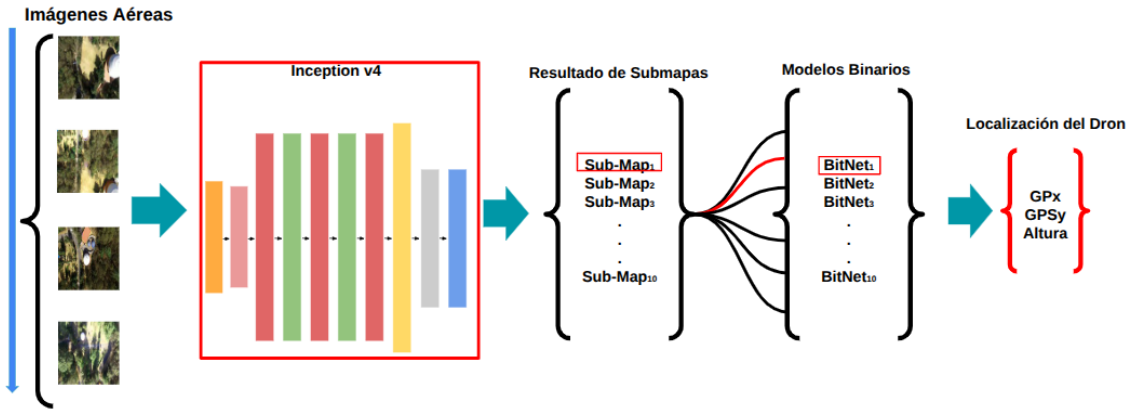

Figura 7.9: Estimación de posición utilizando modelos BitNet a partir del submapa.

Para fines de comparación, la estimación de posición se evaluó con distintos enfoques: 1) Red neuronal (NN) de 4 capas, tanto en un modelo único como en múltiples modelos con la metodología de submapas; 2) PoseNet; 3) ORB-SLAM2 para relocalizar las imágenes de prueba; 4) BitNet, evaluada en un modelo único y múltiples modelos con búsqueda de submapas. Los resultados se presentan en la Tabla 7.12, donde se muestra el error medio de distancia euclidiana en metros para cada enfoque. Este resultado permite medir la eficiencia de los enfoques donde un valor menor refleja una mejor estimación de la posición. Además, los resultados muestran la capacidad de arquitecturas compactas como NN y BitNet en términos de precisión frente a PoseNet y sistemas basados en SLAM.

Tabla 7.12: Error medio de distancia euclidiana en metros. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1			Trayectoria 2			Trayectoria 3			Trayectoria 4		
	x	y	z	x	y	z	x	y	z	x	y	z
PoseNet	47.4	15.3	5.07	94.1	82.3	12.7	59.9	38.3	19.0	146.1	130.8	**2.09**
ORB-SLAM2	-	-	-	**13.04**	**10.68**	6.90	-	-	-	-	-	-
NN - Único	39.4	9.72	**0.10**	82.09	75.7	**5.80**	34.4	37.9	**6.06**	140.7	120.6	7.30
NN - Múltiple	12.8	10.9	4.07	45.8	48.47	13.5	27.8	**21.5**	7.05	**52.64**	**52.75**	23.2
BitNet - Único	12.3	11.6	3.40	77.2	61.2	12.5	35.7	29.8	8.58	59.1	64.6.2	17.7
BitNet - Múltiple	**9.63**	**5.38**	2.97	33.8	28.8	7.32	**26.3**	21.9	7.39	57.3	57.3	11.9

De manera similar a la sección anterior, presentamos los resultados del porcentaje de error por eje de coordenadas (x, y, z), donde un valor alto indica un mayor error en esa coordenada de la posición. Es importante señalar que este error no depende únicamente de la estimación realizada por el modelo, sino que también puede deberse a una asignación incorrecta del submapa durante la búsqueda con la imagen de prueba. También, estos resultados permiten analizar con mayor detalle el desempeño por coordenada, demostrando que algunos ejes presentan resultados más estables mientras que otros concentran el error más alto. La Tabla 7.13 ilustra los resultados obtenidos del porcentaje de error por cada coordenada utilizando los 4 enfoques.

Tabla 7.13: Porcentaje de error por coordenada. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1			Trayectoria 2			Trayectoria 3			Trayectoria 4		
	x	y	z	x	y	z	x	y	z	x	y	z
PoseNet	48.7	30.8	9.91	60.4	43.0	12.7	19.0	33.6	18.9	76.9	53.2	**2.11**
ORB-SLAM2	-	-	-	**16.6**	**5.74**	7.56	-	-	-	-	-	-
NN - Único	40.5	19.5	**5.54**	52.6	39.6	**5.70**	10.9	33.3	**6.03**	74.0	49.0	7.26
NN - Múltiple	13.1	22.0	8.07	29.4	25.3	13.4	8.82	**18.8**	7.02	**27.7**	**21.4**	12.1
BitNet - Único	12.7	23.3	6.74	49.5	32.0	12.5	11.3	26.1	8.54	31.1	26.2	17.6
BitNet - Múltiple	**9.90**	**10.8**	5.89	21.7	15.0	7.29	**8.34**	19.2	7.35	30.1	23.3	11.9

Finalmente, la Tabla 7.14 presenta el porcentaje total de error obtenido con cada uno de los enfoques comparados. Este resultado corresponde a la suma del error de traslación a lo largo de toda la trayectoria real, ofreciendo una representación del desempeño en la localización de las imágenes. Sin embargo, el resulta presenta el error total juntando el resultado por una mala asignación del submapa así como por una estimación incorrecta de la posición. Los conjuntos de prueba con imágenes discontinuas incrementaron el error en el caso de PoseNet, mientras que en los sistemas SLAM esta discontinuidad dificultó completar el mapeo de manera adecuada. En contraste, los enfoques basados en redes compactas, como NN y BitNet, mostraron resultados más consistentes cuando se utilizó un único modelo, mientras que se redujo el error al aplicar el esquema de múltiples modelos. Este hallazgo refuerza el uso del aprendizaje continuo con entrenamiento progresivo para expandir el conocimiento mediante múltiples modelos, logrando reducir el error del resultado sin recurrir al olvido de la información.

Tabla 7.14: Porcentaje total de error con los enfoques comparados. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1	Trayectoria 2	Trayectoria 3	Trayectoria 4
PoseNet	34.3	42.2	22.1	52.0
ORB-SLAM2	-	**11.7**	-	-
NN - Único	26.3	46.5	14.8	50.0
NN - Múltiple	14.1	24.0	10.6	**21.9**
BitNet - Único	13.8	33.7	13.9	25.3
BitNet - Multiple	**9.11**	15.6	**10.5**	24.6

Por último, para evaluar la eficiencia de nuestro enfoque con redes binarias, se presentó en la Tabla 7.15 los resultados del tiempo total de procesamiento. Este tiempo consiste en todo el flujo de la metodología: desde la entrada de una imagen, la extracción de sus características, la identificación del submapa correspondiente, la carga del modelo para esa región y la estimación de la posición. El proceso fue evaluado en ROS procesando las imágenes una por una con el fin de medir el desempeño de la metodología. El resultado con la estrategia de búsqueda de submapas incrementa ligeramente el tiempo en comparación con un modelo único, pero manteniendo un resultado adecuado para aplicaciones de localización en escenarios reales.

Como resultado cualitativo y visual, en la Figura 7.10 se muestran las estimaciones de posición obtenidas en las cuatro trayectorias de vuelo con tres enfoques: PoseNet, la red neuronal (NN) de 4 capas con múltiples modelos, y nuestro enfoque con BitNet utilizando

Tabla 7.15: Tiempo total de procesamiento (ms) para la localización. Los mejores resultados están resaltados en negrita.

Enfoque	Trayectoria 1	Trayectoria 2	Trayectoria 3	Trayectoria 4
ORBSLAM2	-	79.12	-	-
PoseNet	71.94	73.74	73.52	72.15
NN - Único	56.01	50.25	45.22	49.82
NN - Múltiple	61.45	61.43	64.60	64.18
BitNet - Único	**18.47**	**19.94**	**18.73**	**18.69**
BitNet - Múltiple	51.80	52.22	49.91	51.85

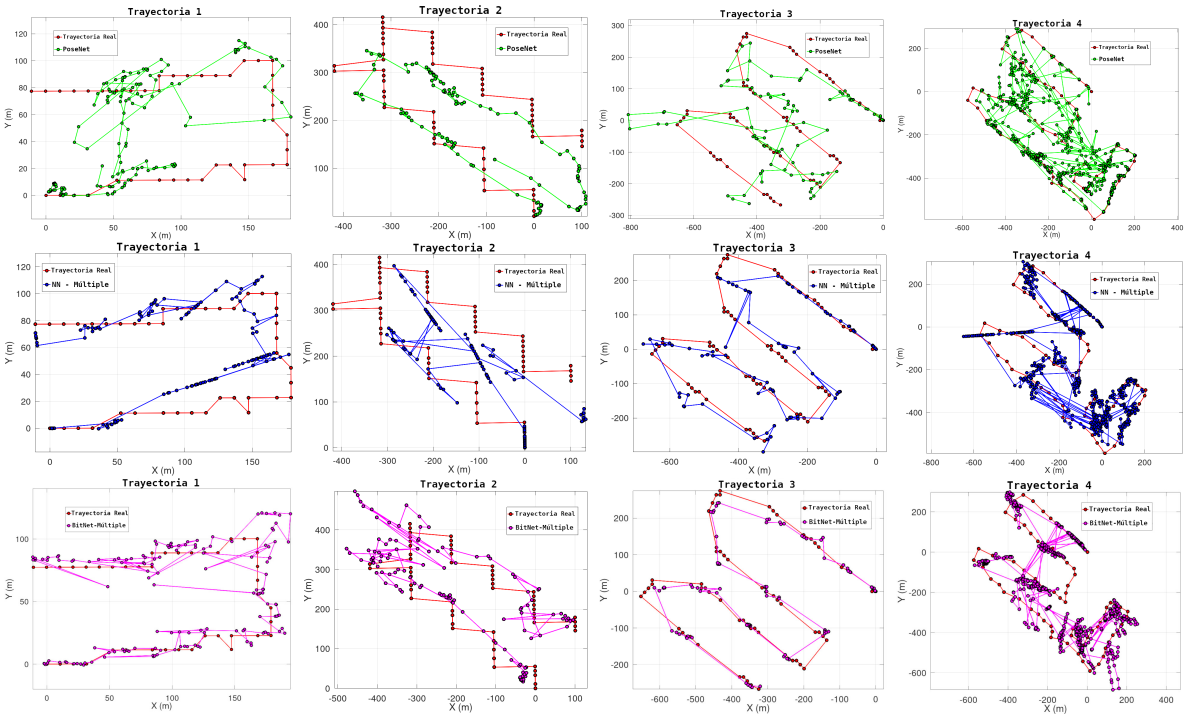

Figura 7.10: Estimación de posición utilizando PoseNet, NN-Múltiple y BitNet-Múltiple.

múltiples modelos. Para facilitar la interpretación, la trayectoria real de las imágenes de prueba se representa en color rojo, mientras que las trayectorias estimadas se presentan en tres columnas comparativas: la primera corresponde a PoseNet, la segunda a la NN de 4 capas con múltiples modelos, y la tercera a BitNet con múltiples modelos. Así, los resultados muestran que PoseNet presenta múltiples saltos en las estimaciones, generando trayectorias con errores y localizaciones inestables. En contraste, tanto la NN como BitNet

con múltiples modelos muestran un desempeño más robusto, reduciendo los saltos y manteniendo trayectorias más cercanas a la real.

7.3.3. Estimación de Posición 6D con BitNet

En esta sección se evaluó el desempeño de la red BitNet en distintos escenarios, aplicando un aprendizaje continuo en un único modelo sin utilizar la expansión del conocimiento mediante múltiples modelos. El objetivo fue valorar la efectividad de diversos enfoques utilizando seis conjuntos de datos: cuatro terrestres y dos aéreos. Para fines de comparación, se incluyeron además dos redes binarias de la literatura, originalmente diseñadas para la adquisición de la posición humana y el reconocimiento de lugares. Con el fin de adaptar estos métodos al problema de localización visual y estimación de posición, se modificaron las capas de salida para obtener tanto la posición como la orientación en los casos requeridos.
Así, los conjuntos de datos terrestres utilizados fueron los siguientes:

- **EuRoC MAV:** Consiste de 11 secuencias grabadas con un dron en escala de grises dentro de un escenario cerrado. Este conjunto es ampliamente utilizado para la validación de sistemas SLAM debido a su complejidad en entornos interiores.
- **TUM RGB-D:** Incluye diversas secuencias en interiores capturadas con cámaras RGB-D y es frecuentemente utilizado en tareas de reconstrucción tridimensional y estimación de posición.
- **7-Scenes:** Es un conjunto de datos de referencia para la estimación de posición 6D que incluye múltiples entornos interiores con diferentes condiciones de iluminación y textura, convirtiéndola en un conjunto de datos desafiante para la localización visual.
- **Cambridge Landmarks:** Este conjunto fue presentado originalmente en el trabajo de PoseNet, contiene secuencias en exteriores y está orientado a la estimación de la posición de cámara mediante el entrenamiento con la red PoseNet.

Por otro lado, se utilizaron dos conjuntos de datos aéreos: 1) Utilizando imágenes aéreas continuas con posiciones consecutivas capturadas dentro de las instalaciones del INAOE; 2) El segundo conjunto corresponde al utilizado previamente en los enfoques de localización con aprendizaje continuo donde las imágenes fueron capturadas de manera no secuencial con el propósito de evaluar el rendimiento del aprendizaje continuo. Así, el uso de estos conjuntos demuestran la efectividad de la estimación de posición en varios escenarios aplicando diversas métricas de evaluación sobre los seis conjuntos de datos y comparando cuantitativamente el desempeño de las redes binarias frente a enfoques tradicionales.

La primera evaluación se desarrolló midiendo la Raíz del Error Cuadrático Medio (RMSE) en los seis enfoques comparados. Los resultados obtenidos con el conjunto de datos EuRoC MAV se presentan en la Tabla 7.16, donde se resalta en negritas el mejor desempeño. Este conjunto de datos es ampliamente utilizado en sistemas SLAM debido a su elevado nivel de textura lo que facilita el proceso de mapeo. De esta manera, ORB-SLAM2 alcanza un rendimiento superior en este escenario dado que el conjunto de datos diseñado para este tipo de sistemas. Sin embargo, las redes binarias mantienen un margen de error bajo y competitivo, lo que permite su potencial para escenarios cerrados con condiciones visuales de imágenes a escala de grises.

Tabla 7.16: Resultados de la Raíz del Error Cuadrático Medio (RMSE) en metros utilizando EuRoC MAV. Los mejores resultados están resaltados en negrita.

Enfoque	V1Easy	V1Med.	V2Easy	MH2Easy	MH4Diff.
ORBSLAM2	**0.035**	**0.020**	**0.037**	**0.018**	**0.119**
PoseNet	-	-	-	0.925	0.941
Red Neuronal	0.084	0.088	0.065	0.138	0.184
BIHRNet	1.853	0.524	0.209	0.390	0.534
BinVPR	1.939	0.529	0.206	0.362	0.508
BitNet	0.356	0.500	0.218	0.331	0.479

Tabla 7.17: Resultados de la Raíz del Error Cuadrático Medio (RMSE) en metros utilizando TUM RGB-D. Los mejores resultados están resaltados en negrita.

Enfoque	desk	desk2	room	2desk	xyz	office	nst
ORBSLAM2	**0.016**	0.022	0.047	**0.009**	**0.004**	**0.010**	**0.019**
PoseNet	0.067	0.096	0.133	0.097	0.050	0.084	0.082
Red Neuronal	0.022	**0.021**	**0.039**	0.043	0.027	-	0.033
BIHRNet	0.198	0.382	0.377	1.746	0.037	0.248	3.220
BinVPR	0.170	0.354	0.304	1.716	0.020	0.186	3.218
BitNet	0.189	0.387	0.336	0.167	0.052	0.276	0.197

En la Tabla 7.17 se presentan los resultados de la métrica RMSE obtenidos con el conjunto de datos TUM RGB-D, utilizando los seis enfoques de evaluación. En este caso, ORB-SLAM2 alcanza nuevamente el mejor desempeño en cinco de los siete escenarios evaluados, mientras que la red neuronal de cuatro capas obtiene un resultado superior en los dos escenarios restantes. Por su parte, las redes binarias muestran un comportamiento

intermedio, mostrando un desempeño inferior en la mayoría de ellos pero manteniendose constantes. Este resultado puede atribuirse a la pérdida de precisión generada por la binarización de los pesos durante el entrenamiento, limitando la capacidad para estimar de manera precisa las posiciones en este conjunto de datos.

Las Tablas 7.18 y 7.19 muestran los resultados obtenidos con los conjuntos de datos 7Scenes y Cambridge utilizados en tareas de localización visual mediante aprendizaje profundo. En este caso, la comparación se centró en enfoques basados en entrenamiento con redes neuronales profundas. En los resultados con Cambridge, PoseNet alcanzó el mejor desempeño reportando un error mediano de localización inferior al de los demás. En contraste, en el conjunto 7Scenes la red neuronal de cuatro capas obtuvo un mejor rendimiento en la estimación de posición, mientras que PoseNet mostró un desempeño superior en la estimación de orientación. Por otro parte, las redes binarias mostraron resultados de errores más elevados, aunque mantuvieron un comportamiento más constante en 7Scenes. En cambio, en Cambridge los errores se incrementaron de manera notable, particularmente en el escenario de gran escala (Street), donde las largas trayectorias y la variabilidad visual complican la estimación precisa de la posición.

Tabla 7.18: Resultados del Error Mediano de Localización utilizando 7Scenes. Los mejores resultados están resaltados en negrita.

Entorno	Red Neuronal	PoseNet	BIHRNet	BinVPR	BitNet
Chess	**0.18m,** 17.8°	0.32m, **8.12°**	0.81m, 29.2°	0.77m, 25.9°	0.58m, 23.0°
Fire	**0.33m,** 33.9°	0.47m, **14.4°**	0.75m, 40.9°	0.73m, 35.7°	0.57m, 30.6°
Heads	**0.16m,** 25.7°	0.29m, **12.0°**	0.45m, 27.2°	0.34m, 24.7°	0.35m, 25.1°
Office	**0.33m,** 28.6°	0.48m, **7.68°**	0.96m, 51.9°	0.93m, 48.9°	0.76m, 31.1°
Pumpkin	**0.30m,** 17.0°	0.47m, **8.42°**	0.77m, 29.2°	0.75m, 27.4°	0.77m, 25.4°
Kitchen	**0.37m,** 28.8°	0.59m, **8.64°**	1.20m, 50.2°	1.09m, 37.9°	1.04m, 40.5°
Stairs	**0.36m,** 26.6°	0.47m, **13.8°**	1.36m, 50.7°	1.36m, 37.9°	0.56m, 25.7°

Tabla 7.19: Resultados del Error Mediano de Localización utilizando Cambridge Landmark. Los mejores resultados están resaltados en negrita.

Entorno	Red Neuronal	PoseNet	BIHRNet	BinVPR	BitNet
King's Colleg	4.04m, 15.1°	**1.92m, 5.40°**	10.1m, 33.5°	7.02m, 31.3°	8.32m, 110.°
Street	54.9m, 124.°	**3.67m, 6.50°**	115.m, 88.2°	99.5m, 83.6°	98.2m, 111.°
Old Hospital	4.91m, 28.4°	**2.31m, 5.38°**	13.5m, 80.2°	11.0m, 78.2°	7.12m, 44.1°
Shop Facade	3.36m, 27.5°	**1.46m, 8.08°**	5.49m, 87.5°	4.39m, 75.8°	5.16m, 66.5°
St Mary's Ch	6.79m, 47.7°	**2.65m, 8.48°**	19.1m, 55.1°	16.4m, 44.3°	15.9m, 90.5°

Por otro lado, al utilizar el conjunto de datos aéreo continuo capturado en el INAOE, se

observaron resultados diferentes al aplicar las redes binarias para la estimación de posición. La Tabla 7.20 presentó el error medio de localización obtenido con cinco enfoques de aprendizaje, donde PoseNet alcanzó el mejor desempeño debido a la información continua de las trayectorias, mejorando su entrenamiento en el modelo de estimación. En contraste, las redes binarias presentaron mayores dificultades para lograr localizaciones precisas cuyos errores son superiores a los 10.0 metros. No obstante, mantuvieron un comportamiento consistente a lo largo de las cinco trayectorias evaluadas, aunque su precisión es menor en escenarios aéreos amplios, ofrecen una estabilidad adecuada.

Tabla 7.20: Resultados de los Errores de localización medios en metros utilizando el conjunto aéreo de INAOE. El mejor resultado se resalta en negrita.

Enfoque	Tray. 1	Tray. 2	Tray. 3	Tray. 4	Tray. 5	Media
PoseNet	3.33	3.23	2.60	2.63	5.43	**3.44**
Red Neuronal	5.68	5.04	3.85	21.8	8.18	8.91
BIHRNet	21.2	22.6	19.6	16.5	20.4	20.0
BinVPR	20.0	19.4	17.3	14.6	18.5	17.9
BitNet	12.1	12.0	9.63	13.7	16.2	12.7

Por último, la Tabla 7.21 reporta el porcentaje total de error utilizando el conjunto de datos diseñado para evaluar el aprendizaje continuo. Debido a la naturaleza discontinua de las posiciones y de las imágenes, PoseNet presentó complicaciones para lograr una localización precisa en los cuatro escenarios. De manera similar, la red neuronal presentó limitaciones en trayectorias largas, donde la falta de continuidad afectó la precisión de las estimaciones. En contraste, los resultados obtenidos con las redes binarias muestran un desempeño más favorable en el caso de BitNet, mientras que otras arquitecturas tienen mayores dificultades para alcanzar localizaciones confiables. Este comportamiento puede atribuirse a un conjunto de factores: la discontinuidad del conjunto de datos, las arquitecturas binarias evaluadas y el proceso de binarización aplicada en cada modelo. No obstante, el rendimiento podría mejorarse optimizando alguno de estos elementos, resultando en estimaciones más precisas en escenarios desafiantes.

Finalmente, para concluir la evaluación de las redes binarias en comparación con los diferentes enfoques aplicados a seis conjuntos de datos, la Tabla 7.22 reporta el tiempo de inferencia expresado en milisegundos (ms). Este resultado muestra la velocidad con la que cada método procesa una imagen de entrada y devuelve una posición estimada. De esta manera, los resultados con las redes binarias presentan tiempos de inferencia menores a los otros enfoques aún con las limitaciones en términos de precisión de localización en

Tabla 7.21: Resultados del porcentaje total de error utilizando el conjunto aéreo para aprendizaje continuo. El mejor resultado se resalta en negrita.

Enfoque	Tray. 1	Tray. 2	Tray. 3	Tray. 4	Media
PoseNet	34.3	42.2	22.1	52.0	37.6
Red Neuronal	26.3	46.5	14.8	50.0	34.4
BIHRNet	55.3	55.3	42.7	50.1	50.8
BinVPR	50.2	52.1	36.7	40.3	44.8
BitNet	13.8	33.7	13.9	25.3	**21.6**

diferentes escenarios. Además, se muestra que las redes binarias destacan por su capacidad de acelerar el procesamiento sin requerir hardware de alto rendimiento, lo que las convierte en candidatas para implementaciones en plataformas a bordo de drones o robots móviles.

Tabla 7.22: Tiempos de inferencia en milisegundos (ms) utilizando cada uno de los enfoques con cada conjunto de datos evaluado. Los mejores resultados se resaltan en negritas.

Enfoque	EuRoC	TUM	7Scenes	Cambridge	Aéreo	Aéreo AC
PoseNet	-	15.8	14.2	**15.18**	14.6	72.8
Red Neuronal	8.56	14.3	14.6	27.5	8.13	50.3
BIHRNet	9.39	11.5	15.9	43.1	8.81	**16.8**
BinVPR	11.8	14.7	15.8	44.9	11.4	20.0
BitNet	**7.47**	**11.4**	**12.5**	37.5	**5.49**	18.9

7.4. Sumario

En este capítulo se presentó la metodología de localización progresiva basada en aprendizaje continuo, cuyo objetivo es expandir el conocimiento utilizando múltiples modelos entrenados de forma incremental. De esta manera, se abordaron entrenaron dos métodos principales: máquinas de soporte vectorial para regresión (SVR) y redes neuronales binarias (BNNs). Cada una de estos procesos de entrenamiento se integro a un esquema de búsqueda de submapas y extracción de características a través de redes profundas. Asimismo, se evaluaron redes binarias en escenarios terrestres y aéreos con imágenes asociadas a posiciones tridimensionales, permitiendo evaluar el desempeño de las redes binarias, así como la propuesta de la localización progresiva.

Los primeros resultados experimentales demostraron que el uso de múltiples modelos de

SVR son adecuados para la estimación de posición, alcanzando errores menores frente a enfoques convencionales especialmente en trayectorias no continuas. Aunque los modelos únicos tienden a sufrir de olvido catastrófico, la metodología progresiva basada en submapas permitió mantener la precisión en la localización sin perder información previa. Por otro lado, las redes binarias mostraron un rendimiento alto en términos de tiempo de inferencia permitiendo reducir los recursos computacionales necesarios y para estimar la posición a partir de una imagen, aunque aún presenta limitaciones en cuestión de precisión.

En resumen, este capítulo mostró la combinación de aprendizaje progresivo con modelos ligeros como SVR y arquitecturas optimizadas como las redes binarias, ofreciendo una alternativa para la localización visual en drones. La propuesta presentó un análisis entre precisión y eficiencia del aprendizaje continuo, permitiendo la posibilidad de ser aplicado en escenarios reales como sistemas de respaldo para la localización en caso de pérdida de señal GPS. Finalmente, se plantea aumentar la precisión de las redes binarias y evaluar el desempeño en tiempo real a bordo de vehículos aéreos, donde la metodología propuesta entre como un sistema de localización progresiva adaptable ante escenarios cambiantes.

Capítulo 8

Conclusiones

En esta tesis se presentó el desarrollo de una metodología para la estimación de posición de una cámara monocular a bordo de un dron, con el propósito de obtener la localización del vehículo mediante información visual. La investigación se centró en diseñar un sistema de localización de respaldo utilizando redes neuronales profundas y conjuntos de imágenes aéreas capturadas durante misiones de vuelo. Las propuestas realizadas en este trabajo presentan un avance en los sistemas de localización visual, ofreciendo una alternativa para escenarios donde la disponibilidad del GPS no está garantizada. Además, los métodos desarrollados permiten relocalizar una posición aproximada de la trayectoria de vuelo, proporcionando al dron una referencia mientras la localización GPS se restablece.

Las preguntas principales de esta investigación fueron: 1) ¿Qué estrategia de aprendizaje continuo, en conjunto con redes neuronales permite que una cámara monocular se localice mientras el modelo resultante aprende de manera dinámica a una velocidad cercana a la frecuencia de captura de la cámara? 2) ¿Qué tipo de esquema de localización puede implementarse a bordo de un dron para que, al recorrer una trayectoria, el sistema conserve el conocimiento previamente adquirido mientras incorpora información nueva sin incurrir en el olvido? Para responder a estas preguntas, se diseñó y evaluó una metodología de localización visual que integra estrategias de aprendizaje continuo, métodos de búsqueda y el uso de redes neuronales ligeras.

El principal hallazgo de esta investigación fue que los avances en aprendizaje continuo aplicados a la localización aérea o visual mediante cámaras monoculares aún no han sido explorados de manera extensiva en la robótica. La mayoría de los trabajos se concentran en tareas de clasificación, mientras que la localización visual en robótica ha recibido menor atención. Asimismo, se encontró que las redes convolucionales presentan limitaciones para trabajos de localización, especialmente por el uso de conjuntos de datos que requieren

largos tiempos de entrenamiento. Como consecuencia, los modelos de aprendizaje no están diseñados para actualizarse dinámicamente, impidiendo mantener una localización en tiempo real. Aunque existen métodos tradicionales como los sistemas de mapeo que ofrecen soluciones validas, su desempeño puede verse comprometido por los cambios constantes dentro del entorno.

Ante estos desafios, esta investigación demuestra que redes ligeras ofrecen una alternativa para acelerar el entrenamiento con conjuntos de datos reducidos. No obstante, sus arquitecturas requieren tiempos de procesamiento para obtener la localización de manera rápida. Bajo este contexto, un componente central de la metodología propuesta fue la incorporación del aprendizaje continuo basado en la repetición latente de los datos, permitiendo entrenar minilotes de información sin perder el conocimiento previamente adquirido. De esta manera, se reduce el tiempo de entrenamiento para crear modelos capaces de actualizarse casi al tiempo en que se desarrolla el vuelo.

Los resultados obtenidos muestran que la metodología propuesta logra una relocalización cercana a la trayectoria real, con niveles de precisión suficientes para operar como un sistema de respaldo ante la pérdida total de GPS. En la localización topológica se alcanzó una precisión promedio de 0.78, mientras que la localización jerárquica obtuvo 0.76, identificando la región o submapa correspondiente en trayectorias discontinuas. En la localización progresiva, el error porcentual total fue de 7.08 % utilizando múltiples modelos SVR y 14.95 % con BitNet. Aunque los errores en distancia se sitúan entre 5.8 m y 29.6 m según el escenario, son superiores al ideal para estimación precisa pero permanecen dentro del orden de magnitud del error típico de sistemas GPS comerciales (5–12 m). Esto demuestra que, aun con variabilidad, la metodología es funcional como mecanismo de localización de respaldo, pues permite recuperar una posición aproximada que mantiene la operación del dron mientras se restablece la señal GPS.

Además de la precisión alcanzada, la metodología demostró ser altamente eficiente en términos de tiempo de procesamiento. La velocidad de inferencia supera los 100 cuadros por segundo, con un tiempo promedio inferior a 20 milisegundos por estimación, lo que permite procesar cada imagen al mismo ritmo de captura de la cámara (30 Hz). Esta capacidad de operación es crucial en escenarios con pérdida de GPS, ya que proporciona una estimación inmediata que estabiliza la navegación y evita comportamientos erráticos del dron. En consecuencia, aunque la precisión obtenida no iguala la del GPS, la rápida disponibilidad de una localización aproximada ofrece un mecanismo de seguridad esencial, garantizando continuidad operativa y reduciendo el riesgo de pérdida del vehículo.

En relación con las hipótesis planteadas en esta tesis, la primera proponía integrar estrategias de aprendizaje continuo con redes neuronales permitiendo a un modelo aprender de manera

dinámica a partir de imágenes y manteniendo el conocimiento previamente adquirido. Esta hipótesis fue validada a través de las metodologías jerárquica y progresiva, en las que se aplicaron dos enfoques: el uso de múltiples modelos que preservan información previa y actual, y el método de reproducción latente con un único modelo. Así, el uso de múltiples modelos demostró ser eficiente para conservar la información de posiciones anteriores mientras se incorporaba nuevas. En el segundo, la estrategia de reproducción latente permitió almacenar patrones representativos del conocimiento previo, reduciendo el olvido catastrófico durante el entrenamiento. No obstante, a medida que aumentan los datos el desempeño de la memoria externa se ve limitado por la capacidad de patrones almacenados, lo que conduce a un olvido lento del conocimiento adquirido.

La segunda hipótesis planteaba que un esquema de múltiples modelos entrenados progresivamente permite conservar la información adquirida durante una misión de vuelo, evitando el olvido de conocimientos previos. Esta hipótesis fue validada mediante la metodología de búsqueda de submapas y la integración de modelos entrenados de forma progresiva en distintas secciones de la trayectoria. Así, la implementación del entrenamiento progresivo a partir de imágenes por cada submapa fue entrenado como un modelo independiente, expandiendo el conocimiento acumulado sin sobrescribir el previamente adquirido. Por lo tanto, la localización progresiva logra aprovechar la información almacenada en modelos anteriores para estimar la posición de la cámara con mayor precisión a partir de una imagen capturada con un dron. Este resultado confirma que el aprendizaje progresivo con múltiples modelos no solo conserva las posiciones pasadas, sino que también incorpora las nuevas de manera continua. Adicionalmente, esta estrategia ofrece un sistema adaptable hacia el diseño de sistemas de localización visual, garantizando un aprendizaje que permite actualizar un modelo al entorno dinámico capturado por la cámara del dron.

8.1. Limitaciones

La principal limitación de esta investigación fue la restricción de memoria y recursos computacionales disponibles en la plataforma utilizada. Si bien el aprendizaje continuo no requiere entrenamientos prolongados, el poder de cómputo sigue determinante al entrenar múltiples conjuntos de datos y generar modelos independientes de manera simultánea. Asimismo, las trayectorias largas pueden incurrir en el olvido parcial de la información previa a medida que se incorporan más datos, aunque la división en submapas ayudó a mitigar este problema no se elimina completamente. Finalmente, las pruebas se realizaron en entornos controlados con trayectorias capturadas en escenarios externos reales pero limitados. Esto presenta un nuevo reto para la localización en tiempo real durante misiones reales de vuelo,

lo cual representa un reto adicional no contemplado en esta investigación.

8.2. Trabajo a Futuro

Como trabajo futuro, esta investigación se enfocará en aumentar la precisión de la localización obtenida por las metodologías propuestas, reduciendo el error de posición para alcanzar la localización más cercana a las reales. De igual manera, se plantea continuar con la exploración de redes binarias y arquitecturas ligeras con el propósito de implementarlas en plataformas de bajo costo y a bordo de drones. Asimismo, uno de los retos más importantes será la validación en escenarios reales y dinámicos incluyendo vuelos en entornos cambiantes e incluso en escenarios desconocidos y nuevos. El objetivo será que el sistema pueda aprender y adaptarse en tiempo real, garantizando la localización del vehículo durante misiones de vuelo autónomo. Finalmente, la investigación de nuevos métodos de aprendizaje continuo y su aplicación a imágenes de profundidad o térmicas pueden ofrecer una localización más robusta que funcione en tiempo real, convirtiéndose en una alternativa para operaciones autónomas en escenarios del mundo real.

Referencias

Abdi, G., Samadzadegan, F., & Kurz, F. (2016). Pose estimation of unmanned aerial vehicles based on a vision-aided multi-sensor fusion. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, *41*.

Acharya, D., Khoshelham, K., & Winter, S. (2019). Bim-posenet: Indoor camera localisation using a 3d indoor model and deep learning from synthetic images. *ISPRS journal of photogrammetry and remote sensing*, *150*, 245–258.

Agency, U. S. D. M. (1987). *Department of Defense World Geodetic System 1984: its definition and relationships with local geodetic systems*, volume 8350. Defense Mapping Agency.

Ardizzone, E., Chella, A., & Pirrone, R. (2000). Pose classification using support vector machines. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 6, (pp. 317–322). IEEE.

Barfoot, T. D. (2017). *Primer on Three-Dimensional Geometry*, (pp. 165–204). Cambridge University Press.

Bednář, J., Petrlík, M., Vivaldini, K. C. T., & Saska, M. (2022). Deployment of reliable visual inertial odometry approaches for unmanned aerial vehicles in real-world environment. In *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*, (pp. 167–176). IEEE.

Benjumea, D., Alcántara, A., Ramos, A., Torres-Gonzalez, A., Sánchez-Cuevas, P., Capitan, J., Heredia, G., & Ollero, A. (2021). Localization system for lightweight unmanned aerial vehicles in inspection tasks. *Sensors*, *21*(17), 5937.

Berton, G., Stoken, A., & Masone, C. (2025). Astroloc: Robust space to ground image localizer. *arXiv preprint arXiv:2502.07003*.

Blanton, H., Greenwell, C., Workman, S., & Jacobs, N. (2020). Extending absolute pose regression to multiple scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (pp. 38–39).

Blanton, H., Workman, S., & Jacobs, N. (2022). A structure-aware method for direct pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (pp. 2019–2028).

Bradski, G. & Kaehler, A. (2000). Opencv. *Dr. Dobb's journal of software tools*, *3*.

Bulat, A. & Tzimiropoulos, G. (2017). Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the IEEE international conference on computer vision*, (pp. 3706–3714).

Bulat, A., Tzimiropoulos, G., Kossaifi, J., & Pantic, M. (2019). Improved training of binary networks for human pose estimation and image recognition. *arXiv preprint arXiv:1904.05868*.

Cabrera-Ponce, A. A., Manuel, M.-O., & Martinez-Carranza, J. (2024). Continual learning for camera localization. *Machine Learning for Complex and Unmanned Systems*, 14–33.

Cabrera-Ponce, A. A., Martin-Ortiz, M., & Martinez-Carranza, J. (2021). Continual learning for multi-camera relocalisation. In *Mexican International Conference on Artificial Intelligence*, (pp. 289–302). Springer.

Cabrera-Ponce, A. A., Martin-Ortiz, M., & Martinez-Carranza, J. (2023a). Continual learning for topological geo-localisation. *Journal of Intelligent & Fuzzy Systems*, *44*(6), 10369–10381.

Cabrera-Ponce, A. A., Martin-Ortiz, M. I., & Martinez-Carranza, J. (2022). Multi-model continual learning for camera localisation from aerial images. In *13th International Micro Air Vehicle Conference*, (pp. 103–109).

Cabrera-Ponce, A. A., Martin-Ortiz, M. I., & Martinez-Carranza, J. (2023b). Hierarchical continual learning for single image aerial localisation. In Moormann, D. (Ed.), *14^{th} annual International Micro Air Vehicle Conference and Competition*, (pp. 40–48)., Aachen, Germany. Paper no. IMAV2023-5.

Cabrera-Ponce, A. A., Martin-Ortiz, M. I., & Martinez-Carranza, J. (2025). Continual learning via multiple support vector models for localization with a single aerial image. *Unmanned Systems*, 1–14.

Cabrera-Ponce, A. A. & Martinez-Carranza, J. (2019). Aerial geo-localisation for mavs using posenet. In *2019 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED UAS)*, (pp. 192–198). IEEE.

Cabrera-Ponce, A. A., Martinez-Carranza, J., & Martinez-Carranza, J. (2022). Convolutional neural networks for geo-localisation with a single aerial image. *Journal of Real-Time Image Processing*, 1–11.

Cai, M., Shen, C., & Reid, I. D. (2018). A hybrid probabilistic model for camera relocalization. In *BMVC*, volume 1, (pp.8̃).

Cai, Z. & Müller, M. (2023). Clnerf: Continual learning meets nerf. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 23185–23194).

Carrasco, P., Cuesta, F., Caballero, R., Pérez-Grau, F. J., & Viguria, A. (2021). Monte-carlo localization for aerial robots using 3d lidar and uwb sensing. In *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, (pp. 354–360). IEEE.

Caselitz, T., Steder, B., Ruhnke, M., & Burgard, W. (2016). Monocular camera localization in 3d lidar maps. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (pp. 1926–1931). IEEE.

Chathuranga, T. S. & Munasinghe, R. (2019). Aerial image matching based relative localization of a uav in urban environments. In *2019 Moratuwa Engineering Research Conference (MERCon)*, (pp. 633–637). IEEE.

Chee, J., Cai, Y., Kuleshov, V., & De Sa, C. M. (2023). Quip: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, *36*, 4396–4429.

Chen, K., Gong, S., & Xiang, T. (2011). Human pose estimation using structural support vector machines. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, (pp. 846–851). IEEE.

Chen, S., Wu, J., Lu, Q., Wang, Y., & Lin, Z. (2021). Cross-scene loop-closure detection with continual learning for visual simultaneous localization and mapping. *International Journal of Advanced Robotic Systems*, *18*(5), 17298814211050560.

Chow, J. F., Kocer, B. B., Henawy, J., Seet, G., Li, Z., Yau, W. Y., & Pratama, M. (2019). Toward underground localization: Lidar inertial odometry enabled aerial robot navigation. *arXiv preprint arXiv:1910.13085*.

Cimarelli, C., Cazzato, D., Olivares-Mendez, M. A., & Voos, H. (2019). Faster visual-based localization with mobile-posenet. In *International Conference on Computer Analysis of Images and Patterns*, (pp. 219–230). Springer.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., & Bengio, Y. (2016). Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*.

Craig, J. J. (2005). *Introduction to robotics: mechanics and control, 3/E*. Pearson Education India.

Cudrano, P., Luo, X., & Matteucci, M. (2024). The empirical impact of forgetting and transfer in continual visual odometry. *arXiv preprint arXiv:2406.01797*.

Cui, J. & Chen, X. (2023). Ccl: Continual contrastive learning for lidar place recognition. *IEEE Robotics and Automation Letters*, *8*(8), 4433–4440.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. *Advances in neural information processing systems*, *9*.

Dusmanu, M., Miksik, O., Schönberger, J. L., & Pollefeys, M. (2021). Cross-descriptor visual localization and mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 6058–6067).

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Gu, T., Wang, Z., Chi, Z., Zhu, Y., & Du, W. (2021). Unsupervised cycle optimization learning for single-view depth and camera pose with kalman filter. *Engineering Applications of Artificial Intelligence*, *106*, 104488.

Guo, E., Chen, Z., Zhou, Y., & Wu, D. O. (2021). Unsupervised learning of depth and camera pose with feature map warping. *Sensors*, *21*(3), 923.

He, J., Mao, R., Shao, Z., & Zhu, F. (2020). Incremental learning in online scenario. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 13926–13935).

He, Y., Cisneros, I., Keetha, N., Patrikar, J., Ye, Z., Higgins, I., Hu, Y., Kapoor, P., & Scherer, S. (2023). Foundloc: Vision-based onboard aerial localization in the wild. *arXiv preprint arXiv:2310.16299*.

Helgesen, H. H., Leira, F. S., Bryne, T. H., Albrektsen, S. M., & Johansen, T. A. (2019). Real-time georeferencing of thermal images using small fixed-wing uavs in maritime environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, *154*, 84–97.

Herrera, L., Kim, J. J., & Agrawal, B. N. (2025). Deep learning for unambiguous pose estimation of a non-cooperative fixed-wing uav. *Machine Vision and Applications*, *36*(1), 5.

Hofmann-Wellenhof, B., Lichtenegger, H., & Collins, J. (2012). *Global positioning system: theory and practice*. Springer Science & Business Media.

Ionescu, C., Bo, L., & Sminchisescu, C. (2009). Structural svm for visual localization and continuous state estimation. In *2009 IEEE 12th International Conference on Computer Vision*, (pp. 1157–1164). IEEE.

Jantos, T. G., Hamdad, M. A., Granig, W., Weiss, S., & Steinbrener, J. (2023). Poet: Pose estimation transformer for single-view, multi-object 6d pose estimation. In *Conference on Robot Learning*, (pp. 1060–1070). PMLR.

Jin, R., Jiang, J., Qi, Y., Lin, D., & Song, T. (2019). Drone detection and pose estimation using relational graph networks. *Sensors*, *19*(6), 1479.

Jokic, P., Emery, S., & Benini, L. (2018). Binaryeye: A 20 kfps streaming camera system on fpga with real-time on-device image recognition using binary neural networks. In *2018 IEEE 13th International Symposium on Industrial Embedded Systems (SIES)*, (pp. 1–7). IEEE.

Jondhale, S. R., Mohan, V., Sharma, B. B., Lloret, J., & Athawale, S. V. (2022). Support vector regression for mobile target localization in indoor environments. *Sensors*, *22*(1), 358.

Kendall, A. & Cipolla, R. (2016). Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, (pp. 4762–4769). IEEE.

Kendall, A. & Cipolla, R. (2017). Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 5974–5983).

Kendall, A., Grimes, M., & Cipolla, R. (2015). Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, (pp. 2938–2946).

Kim, D. & Ko, K. (2022). Camera localization with siamese neural networks using iterative relative pose estimation. *Journal of Computational Design and Engineering*, *9*(4), 1482–1497.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521–3526.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Li, B., Yan, Z., Wu, D., Jiang, H., & Zha, H. (2024). Learn to memorize and to forget: A continual learning perspective of dynamic slam. In *European Conference on Computer Vision*, (pp. 41–57). Springer.

Li, H., Jiang, H., Gu, X., Peng, J., Li, W., Hong, L., & Tao, C. (2020). Clrs: Continual learning benchmark for remote sensing image scene classification. *Sensors*, *20*(4), 1226.

Li, M., Qin, J., Li, D., Chen, R., Liao, X., & Guo, B. (2021). Vnlstm-posenet: A novel deep convnet for real-time 6-dof camera relocalization in urban streets. *Geo-Spatial Information Science*, *24*(3), 422–437.

Li, R., Wang, S., Long, Z., & Gu, D. (2018). Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, (pp. 7286–7291). IEEE.

Li-Chee-Ming, J. & Armenakis, C. (2018). Uav navigation system using line-based sensor pose estimation. *Geo-spatial information science*, *21*(1), 2–11.

Lima, R., Das, K., & Ghose, D. (2019). Support vector regression based sensor localization using uav. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, (pp. 938–945).

Lin, Y., Liu, Z., Huang, J., Wang, C., Du, G., Bai, J., & Lian, S. (2019). Deep global-relative networks for end-to-end 6-dof visual localization and odometry. In *Pacific Rim International Conference on Artificial Intelligence*, (pp. 454–467). Springer.

Liu, Y., Zhao, C., & Wei, Y. (2022). A robust localization system fusion vision-cnn relocalization and progressive scan matching for indoor mobile robots. *Applied Sciences*, *12*(6), 3007.

Lomonaco, V. & Maltoni, D. (2017). Core50: a new dataset and benchmark for continuous object recognition. *arXiv preprint arXiv:1705.03550*.

Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J., & Wei, F. (2024). The era of 1-bit llms: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, *1*.

Madhukaro, J. R. & Rao, D. (2024). Svm-ga based a novel technique for the detection of the vehicle in an optimized overlapped multi-camera system. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(1), 810–818.

Mandal, D. & Jain, A. (2022). Unsupervised learning of depth, camera pose and optical flow from monocular video. *arXiv preprint arXiv:2205.09821*.

Martínez-Carranza, J., Bostock, R., Willcox, S., Cowling, I., & Mayol-Cuevas, W. (2016). Indoor mav auto-retrieval using fast 6d relocalisation. *Advanced Robotics*, *30*(2), 119–130.

Mascaro, R., Teixeira, L., Hinzmann, T., Siegwart, R., & Chli, M. (2018). Gomsf: Graph-optimization based multi-sensor fusion for robust uav pose estimation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, (pp. 1421–1428). IEEE.

Matrice, D. (2017). 100. *On-line: www. dji. com/matrice100, Accessed*, *19*, 12.

Matteo, B., Tsesmelis, T., James, S., Poiesi, F., & Del Bue, A. (2024). 6dgs: 6d pose estimation from a single image and a 3d gaussian splatting model. In *European Conference on Computer Vision*, (pp. 420–436). Springer.

Meng, L., Chen, J., Tung, F., Little, J. J., & de Silva, C. W. (2016). Exploiting random rgb and sparse features for camera pose estimation. In *BMVC*.

Mermillod, M., Bugaiska, A., & Bonin, P. (2013). The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects.

Mohanty, V., Agrawal, S., Datta, S., Ghosh, A., Sharma, V. D., & Chakravarty, D. (2016). Deepvo: A deep learning approach for monocular visual odometry. *arXiv preprint arXiv:1611.06069*.

Moreau, A., Piasco, N., Bennehar, M., Tsishkou, D., Stanciulescu, B., & de La Fortelle, A. (2023). Crossfire: Camera relocalization on self-supervised features from an implicit representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 252–262).

Müller, M., Urban, S., & Jutzi, B. (2017). Squeezeposenet: Image based pose regression with small convolutional neural networks for real time uas navigation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*, 49–57.

Mur-Artal, R. & Tardós, J. D. (2017). Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, *33*(5), 1255–1262.

Ott, F., Feigl, T., Loffler, C., & Mutschler, C. (2020). Vipr: visual-odometry-aided pose regression for 6dof camera localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (pp. 42–43).

Pan, Y., Zhou, W., Cao, Y., & Zha, H. (2024). Adaptive vio: Deep visual-inertial odometry with online continual learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 18019–18028). IEEE.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural networks*, *113*, 54–71.

Pirinen, A., Samuelsson, A., Backsund, J., & Aström, K. (2022). Aerial view goal localization with reinforcement learning. *arXiv preprint arXiv:2209.03694*.

Pita Fernández, S. & Pértegas Díaz, S. (2002). Investigación cuantitativa y cualitativa.

Ponce, A. A. C., Pérez, L. O. R., Ortiz, M. I. M., & Martinez-Carranza, J. (2024). Binary networks and continual learning for pose estimation from a single aerial image.

In Richardson, T. (Ed.), *15th annual International Micro Air Vehicle Conference and Competition*, (pp. 101–109)., Bristol, United Kingdom. Paper no. IMAV2024-11.

Qian, J., Chen, K., Chen, Q., Yang, Y., Zhang, J., & Chen, S. (2021). Robust visual-lidar simultaneous localization and mapping system for uav. *IEEE geoscience and remote sensing letters*, *19*, 1–5.

Qiao, C., Xiang, Z., Fan, Y., Bai, T., Zhao, X., & Fu, J. (2023). Transapr: Absolute camera pose regression with spatial and temporal attention. *IEEE Robotics and Automation Letters*, *8*(8), 4633–4640.

Qiao, Y., Cappelle, C., & Ruichek, Y. (2015). Place recognition based visual localization using lbp feature and svm. In *Advances in Artificial Intelligence and Its Applications: 14th Mexican International Conference on Artificial Intelligence, MICAI 2015, Cuernavaca, Morelos, Mexico, October 25-31, 2015, Proceedings, Part II 14*, (pp. 393–404). Springer.

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., & Ng, A. Y. (2009). Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, (pp.5̃). Kobe, Japan.

Rabiee, S. & Biswas, J. (2021). Iv-slam: Introspective vision for simultaneous localization and mapping. In *Conference on Robot Learning*, (pp. 1100–1109). PMLR.

Rojas-Perez, L. O. & Martinez-Carranza, J. (2023). Deeppilot4pose: a fast pose localisation for mav indoor flight using the oak-d camera. *Journal of Real-Time Image Processing*, *20*(1), 8.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Ruping, S. (2001). Incremental learning with support vector machines. In *Proceedings 2001 IEEE International Conference on Data Mining*, (pp. 641–642).

Rus Arias, E. (2021). Investigación cuantitativa.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., & Hadsell, R. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Safa, A., Verbelen, T., Ocket, I., Bourdoux, A., Sahli, H., Catthoor, F., & Gielen, G. G. (2022). Learning to encode vision on the fly in unknown environments: A continual

learning slam approach for drones. In *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, (pp. 373–378).

Samano, N., Zhou, M., & Calway, A. (2021). Global aerial localisation using image and map embeddings. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, (pp. 5788–5794). IEEE.

Sarlin, P.-E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Hammarstrand, L., Kahl, F., et al. (2021). Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 3247–3257).

Scaramuzza, D. & Fraundorfer, F. (2011). Visual odometry [tutorial]. *IEEE robotics & automation magazine*, *18*(4), 80–92.

Seifi, S. & Tuytelaars, T. (2019). How to improve cnn-based 6-dof camera pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, (pp. 0–0).

Shen, S., Yu, G., Zhang, L., Yan, Y., & Zhai, Z. (2025). Landnet: Combine cnn and transformer to learn absolute camera pose for the fixed-wing aircraft approach and landing. *Remote Sensing*, *17*(4), 653.

Shetty, A. & Gao, G. X. (2019). Uav pose estimation using cross-view geolocalization with satellite imagery. In *2019 International Conference on Robotics and Automation (ICRA)*, (pp. 1827–1833). IEEE.

Su, W., Ravankar, A., Ravankar, A. A., Kobayashi, Y., & Emaru, T. (2017). Uav pose estimation using ir and rgb cameras. In *2017 IEEE/SICE International Symposium on System Integration (SII)*, (pp. 151–156). IEEE.

Sucar, E., Liu, S., Ortiz, J., & Davison, A. J. (2021). imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 6229–6238).

Syed, N. A., Liu, H., & Sung, K. K. (1999). Handling concept drifts in incremental learning with support vector machines. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 317–321).

Taguchi, S. & Hirose, N. (2022). Unsupervised simultaneous learning for camera re-localization and depth estimation from video. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (pp. 6840–6847). IEEE.

Tang, S., Tang, C., Huang, R., Zhu, S., & Tan, P. (2021). Learning camera localization via dense scene matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 1831–1841).

Thrun, S., Burgard, W., & Fox, D. (2005). Probabilistic robotics cambridge. *MA: MIT Press [Google Scholar]*.

Truong, T.-D., Helton, P., Moustafa, A., Cothren, J. D., & Luu, K. (2024). Conda: Continual unsupervised domain adaptation learning in visual perception for self-driving cars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 5642–5650).

Valada, A., Radwan, N., & Burgard, W. (2018). Deep auxiliary learning for visual localization and odometry. In *2018 IEEE international conference on robotics and automation (ICRA)*, (pp. 6939–6946). IEEE.

Vallone, A., Warburg, F., Hansen, H., Hauberg, S., & Civera, J. (2022). Danish airs and grounds: A dataset for aerial-to-street-level place recognition and localization. *arXiv preprint arXiv:2202.01821*.

Vödisch, N., Cattaneo, D., Burgard, W., & Valada, A. (2023a). Continual slam: Beyond lifelong simultaneous localization and mapping through continual learning. In *Robotics Research* (pp. 19–35). Springer.

Vödisch, N., Cattaneo, D., Burgard, W., & Valada, A. (2023b). Covio: Online continual learning for visual-inertial odometry. *arXiv preprint arXiv:2303.10149*.

Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., Yang, F., Wang, R., Wu, Y., & Wei, F. (2023). Bitnet: Scaling 1-bit transformers for large language models. arxiv. *arXiv preprint arXiv:2310.11453*, *3*.

Wang, J., Han, J., Dong, R., & Kan, J. (2024). Binvpr: Binary neural networks towards real-valued for visual place recognition. *Sensors*, *24*(13), 4130.

Wang, M., Xu, J., Zhang, J., & Cui, Y. (2024). An autonomous navigation method for orchard rows based on a combination of an improved a-star algorithm and svr. *Precision Agriculture*, 1–25.

Wang, S., Clark, R., Wen, H., & Trigoni, N. (2017). Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, (pp. 2043–2050). IEEE.

Wang, S., Laskar, Z., Melekhov, I., Li, X., & Kannala, J. (2021). Continual learning for image-based camera localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 3252–3262).

Wang, Y., Liu, E., & Wang, R. (2020). Camera re-localization by training multi-dataset simultaneously via convolutional neural network. In *Proceedings of the 2020 3rd International Conference on Signal Processing and Machine Learning*, (pp. 35–39).

Wattanacheep, B. & Chitsobhuk, O. (2020). Camera pose estimation using cnn. In *Proceedings of the 3rd International Conference on Control and Computer Vision*, (pp. 84–88).

Wickramasuriya, M., Yu, B., Lee, T., & Snyder, M. (2025). Vision-in-the-loop simulation for deep monocular pose estimation of uav in ocean environment. In *2025 International Conference on Unmanned Aircraft Systems (ICUAS)*, (pp. 749–756). IEEE.

Wong, D., Deguchi, D., Ide, I., & Murase, H. (2017). Single camera vehicle localization using feature scale tracklets. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *100*(2), 702–713.

Xing, X., He, X., Liu, K., Chen, Z., Song, G., Hao, Q., Zhang, L., & Mao, J. (2025). Long-duration uav localization across day and night by fusing dual-vision geo-registration with inertial measurements. *Drones*, *9*(5), 373.

Yan, Z., Tian, Y., Shi, X., Guo, P., Wang, P., & Zha, H. (2021). Continual neural mapping: Learning an implicit scene representation from sequential observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 15782–15792).

Yang, J.-C., Lin, C.-J., You, B.-Y., Yan, Y.-L., & Cheng, T.-H. (2021). Rtlio: Real-time lidar-inertial odometry and mapping for uavs. *Sensors*, *21*(12), 3955.

Yang, L., Bai, Z., Tang, C., Li, H., Furukawa, Y., & Tan, P. (2019). Sanet: Scene agnostic network for camera localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, (pp. 42–51).

Yang, Z., Yu, X., & Yang, Y. (2021). Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 3907–3916).

Yin, P., Abuduweili, A., Zhao, S., Liu, C., & Scherer, S. (2022). Bioslam: A bio-inspired lifelong memory system for general place recognition. *arXiv preprint arXiv:2208.14543*.

Yoon, H.-S., Hwang, J.-H., Kim, C., Son, E. I., Yoo, S.-W., & Seo, S.-W. (2024). Adaptive robot traversability estimation based on self-supervised online continual learning in unstructured environments. *IEEE Robotics and Automation Letters*, *9*(6), 4902–4909.

Zaffar, M., Nan, L., & Kooij, J. F. P. (2023). Copr: Toward accurate visual localization with continuous place-descriptor regression. *IEEE Transactions on Robotics*, *39*(4), 2825–2841.

Zhang, R., Luo, Z., Dhanjal, S., Schmotzer, C., & Hasija, S. (2018). Posenet++: A cnn framework for online pose regression and robot re-localization. *Tech. Rep*.

Zhang, Z., Sun, X., Dang, Y., & Yin, J. (2023). Bihrnet: A binary high-resolution network for human pose estimation. *arXiv preprint arXiv:2311.10296*.

Zhao, C., Fan, B., Hu, J., Tian, L., Zhang, Z., Li, S., & Pan, Q. (2017). Pose estimation for multi-camera systems. In *2017 IEEE International Conference on Unmanned Systems (ICUS)*, (pp. 533–538). IEEE.

Zhi, H., Yin, C., Li, H., & Pang, S. (2022). An unsupervised monocular visual odometry based on multi-scale modeling. *Sensors*, *22*(14), 5193.