

**Testado de medidas de detección térmica de
cachalotes con el fin de minimizar las colisiones de
cetáceos con el tráfico marítimo en la demarcación
marina canaria**

Informe Final

(Enero - Diciembre 2024)

LIFE IP INTEMARES

**Gestión integrada, innovadora participativa de la Red
Natura 2000 en el medio marino español**

LIFE15 IP ES012 – INTEMARES

C1.3 Proyectos demostrativos para evitar colisiones de cetáceos con el tráfico marítimo

1. RESUMEN EJECUTIVO	3
2. RESUMEN DE ACTUACIONES	5
2.1. ANTECEDENTES	5
2.2. CAMPAÑAS DE AVISTAMIENTO	6
2.3. CREACION DE DATASET PARA DETECCION	8
2.4. DISEÑO DE LOS MODELOS DE DETECCIÓN	35
2.5. DESARROLLO DEL SOFTWARE Y EVALUACIÓN	45
3. DESCRIPCIÓN DE LA SOLUCIÓN	48
3.1. ARQUITECTURA DE LA SOLUCIÓN	48
3.2. DEMOSTRADOR TECNOLÓGICO	50
4. RESULTADOS DE LA EVALUACIÓN CUALITATIVA INDEPENDIENTE Y CONCLUSIONES	57
4.1. RESULTADOS	57
4.2. CONCLUSIONES	59
5. MEDIDAS TÉCNICAS PARA SU APLICACION.....	62
5.1. SENSORES Y SISTEMA ÓPTICO	62
5.2. MÉTODOS Y ALGORITMOS	63
5.3. CAMPAÑAS Y CAPTURA DE DATOS	65
5.4. PASO A PRODUCCIÓN Y PILOTO	66
5.5. VALORACIÓN PRESUPUESTARIA	69
6. REFERENCIAS.....	72
ANEXO 1 – LISTADO DE ENTREGABLES.....	75

Autoría:

Arquimea Research Center y Universidad de La Laguna



Coordinación y revisión:

Fundación Biodiversidad del Ministerio para la Transición Ecológica y el Reto Demográfico

Edita:

El proyecto LIFE INTEMARES avanza hacia el objetivo de lograr una gestión eficaz de los espacios marinos de la Red Natura 2000, con la participación de los sectores implicados y con la investigación como herramientas básicas.

La Fundación Biodiversidad del Ministerio para la Transición Ecológica y el Reto Demográfico coordina el proyecto. Participan como socios el propio ministerio, a través de la Dirección General de Biodiversidad, Bosques y Desertificación; la Junta de Andalucía, a través de la Consejería de Consejería de Sostenibilidad, y Medio Ambiente, así como de la Agencia de Medio Ambiente y Agua; el Instituto Español de Oceanografía; AZTI; la Universidad de Alicante; la Universidad Politécnica de Valencia; la Confederación Española de Pesca, SEO/BirdLife y WWF-España. Cuenta con la contribución del Programa LIFE de la Unión Europea.



Coordina



Socios



Fecha de edición

20/12/2024

Agradecimientos

Universidad de Las Palmas de Gran Canaria, Plataforma Oceánica de Canarias.

1. RESUMEN EJECUTIVO

En Canarias, se ha detectado una preocupante problemática de colisiones entre ferris rápidos con cetáceos (especialmente cachalotes), un escenario que plantea riesgos tanto para la seguridad de los animales como para la operación marítima. En respuesta a este desafío, la Fundación Biodiversidad, en el marco del LIFE IP INTEMARES, convocó la licitación “Testado de medidas de detección térmica de cachalotes con el fin de minimizar las colisiones de cetáceos con el tráfico marítimo en la demarcación marina canaria” adjudicada a Arquimea Research Center quien, junto a sus colaboradores, inició una investigación centrada en la selección y mejora del desarrollo de una tecnología térmica para detectar los animales marinos desde embarcaciones de transporte de pasajeros y mitigar las mencionadas colisiones. Además, investigó también el empleo de algoritmos de Inteligencia Artificial en imagen térmica para detectar la presencia de animales marinos o sus soplos.

Fruto de la preocupación de FRED. OLSEN, naviera con ferris rápidos en su flota que realiza trayectos entre las islas canarias, por el reto de reducir las colisiones, ha colaborado en proyectos previos que buscaban este objetivo. En el marco de estas actuaciones, se ha comprometido a otorgar acceso a observadores de mamíferos marinos (MMO, por sus siglas en inglés) de la Universidad de La Laguna en sus dos ferris rápidos que recorren las rutas Santa Cruz de Tenerife – Agaete (Gran Canaria) y Las Palmas de Gran Canaria – Morro Jable (Fuerteventura), objeto de la mencionada licitación. Esta colaboración, además, ha facilitado alcanzar el objetivo previsto en la licitación de trabajos, ya que la naviera ha permitido la instalación de dos sistemas ópticos y de grabación de alta precisión en estos ferris, pieza clave para la recopilación de datos para el proyecto.

En la sección 2 se presenta un resumen de las actuaciones llevadas a cabo durante el año 2024, incluyendo el avance de las campañas de avistamientos, creación de los datasets para detección de cetáceos, y el diseño y validación de los modelos de detección basados en deep learning. Las actuaciones culminan con el desarrollo del software para realizar la prueba de concepto y evaluación cualitativa de forma independiente de la solución desarrollada.

En la sección 3 se realiza una descripción detallada de la arquitectura software de la solución y diseño algorítmico. Posteriormente se describen los diferentes bloques funcionales que implementa el demostrador tecnológico.

Los resultados y conclusiones se presentan en la sección 4, donde además de realizar una síntesis de los resultados de la evaluación cualitativa, se realiza una valoración general del desarrollo y resultados del proyecto. El trabajo realizado representa un avance significativo en el estado del arte de la detección térmica de cetáceos, validando técnicas novedosas de fusión multimodal basadas en deep learning y abordando desafíos técnicos como la alineación de imágenes y la escasez de datos. El conocimiento adquirido a través de esta prueba de concepto debería orientar futuras investigaciones y desarrollos en este campo.

Los resultados presentados reflejan objetivamente tanto los avances como las limitaciones del sistema evaluado. Mientras que la tasa de acierto general es alentadora, los desafíos asociados a la calidad de la imagen, las condiciones del entorno y la configuración de los algoritmos destacan áreas clave para mejorar.

En línea con este análisis, en la sección 5 se presentan una serie de posibilidades de mejora a nivel del sistema óptico, algoritmos y campañas de recolección de datos. Finalmente se detallan las medidas técnicas necesarias para el paso a producción y su estimación presupuestaria.

Executive summary

In the Canary Islands, a concerning issue has been identified involving collisions between high-speed ferries and cetaceans (especially sperm whales), creating risks for both the animals' safety and maritime operations. In response to this challenge, the Biodiversity Foundation, under the LIFE IP INTEMARES framework, announced the call for proposals titled "Testing Thermal Detection Measures for Sperm Whales to Minimize Cetacean Collisions with Maritime Traffic in the Canary Marine Demarcation." This project was awarded to Arquimea Research Center, which, together with its collaborators, initiated a study focused on the selection and further development of thermal technology to detect marine animals from passenger vessels and mitigate such collisions. Additionally, the study explored the use of artificial intelligence algorithms in thermal imaging to detect the presence of marine animals or their blows.

FRED. OLSEN, a shipping company operating high-speed ferries among the Canary Islands, has demonstrated concern about the challenge of reducing collisions. The company has participated in previous projects with this objective. Within these efforts, FRED. OLSEN has committed to granting access to marine mammal observers (MMOs) from the University of La Laguna aboard its two high-speed ferries operating on the Santa Cruz de Tenerife – Agaete (Gran Canaria) and Las Palmas de Gran Canaria – Morro Jable (Fuerteventura) routes, which are the focus of the aforementioned project. Furthermore, the company has facilitated the achievement of the project objectives by allowing the installation of two high-precision optical and recording systems on these ferries, a key component for data collection in the project.

Section 2 summarizes the activities carried out during 2024, including progress in sighting campaigns, the creation of datasets for cetacean detection, and the design and validation of detection models based on deep learning. These activities culminated in the development of software for proof-of-concept testing and the independent qualitative evaluation of the solution.

Section 3 provides a detailed description of the software architecture and algorithmic design of the solution. It also describes the different functional blocks implemented by the technological demonstrator.

The results and conclusions are presented in Section 4. This section synthesizes the qualitative evaluation results and provides a general assessment of the project's development and outcomes. The work represents a significant advancement in the state of the art for thermal cetacean detection, validating novel multimodal fusion techniques based on deep learning and addressing technical challenges such as image alignment and data scarcity. The knowledge gained through this proof of concept should guide future research and development in this field.

The results presented objectively reflect both the progress and the limitations of the system evaluated. While the overall accuracy rate is encouraging, challenges related to image quality, environmental conditions, and algorithm configuration highlight key areas for improvement.

In line with this analysis, Section 5 outlines a series of improvement opportunities regarding the optical system, algorithms, and data collection campaigns. Finally, it details the technical measures necessary for moving to production and provides an estimated budget for implementation.

2. RESUMEN DE ACTUACIONES

A continuación, se detallan brevemente las principales actuaciones realizadas en el periodo que comprende el presente informe tras una breve introducción que resume las actuaciones incluidas en los informes de seguimiento previos

2.1. ANTECEDENTES

Las Islas Canarias han sido identificadas por la Comisión Ballenera Internacional y la Organización Marítima Internacional como uno de los cuatro puntos calientes del mundo con respecto al problema de las colisiones de buques con cetáceos. En particular, en este ámbito geográfico, la especie más afectada es el cachalote (*Physeter macrocephalus*), especie caracterizada por largos periodos de gestación y desarrollo hasta alcanzar la madurez sexual.

El tráfico marítimo en el archipiélago, en constante crecimiento tanto en rutas como en densidad y velocidad de las embarcaciones, representa un riesgo importante para las poblaciones de cachalotes. Anteriormente, en otras partes del mundo, se han evaluado medidas para mitigar el número de colisiones entre cachalotes y embarcaciones, entre las que se encuentran la detección por imagen térmica infrarroja. Esta técnica persigue la detección del animal en superficie a través de la identificación de su soplo, cuya diferencia de temperatura con el medio es mayor que la del propio cuerpo del animal. Los retos de la detección por imagen térmica son varios, entre los que se encuentran conseguir esta detección con suficiente antelación como para alertar a la tripulación con tiempo suficiente para tomar las decisiones precisas para tratar de evitar el choque. El objetivo de este proyecto es evaluar y testar softwares desarrollados para conseguir la detección de cachalotes mediante esta tecnología térmica en aguas canarias.

En el primer informe (E.1.) para el periodo julio 2022 – diciembre 2022 se realizó una introducción de las medidas de detección y prevención de colisiones entre embarcaciones y cetáceos, dentro de las cuales se encuentra la detección térmica. Las medidas de detección térmica, objeto de esta licitación, se estudiaron con más detalle posteriormente, analizando los retos, ventajas e inconvenientes de este tipo de soluciones.

Asimismo, se incluyeron análisis de los sistemas ópticos estudiados durante la primera fase del proyecto y la justificación del sistema óptico propuesto en base a sus características y mejoras con respecto a lo especificado en las prescripciones técnicas de la licitación.

Por último, se realizó un análisis del estado del arte de trabajos previos de detección térmica de cetáceos, poniendo de manifiesto sus limitaciones técnicas y proponiendo como alternativa el uso de novedosas técnicas de *deep learning* para detección de objetos.

En el segundo informe (E.2.) para el periodo enero 2023 – diciembre 2023 se presentó en primer lugar el sistema de grabación elegido y validado para el registro de las imágenes captadas por las cámaras térmicas Sentry (comercializadas por la empresa SEA.AI) a bordo de los buques de Fred.Olsen. Gracias al acuerdo alcanzado con la naviera, se ha permitido la captura de los datos necesarios para el entrenamiento de los modelos de detección y la validación del software.

Diferentes problemas técnicos en la instalación y funcionamiento de las cámaras hicieron necesario una validación exhaustiva en el trayecto Santa Cruz de Tenerife – Agaete antes de la adquisición de la segunda cámara para el trayecto Las Palmas – Morroja. En cualquier caso, y una vez definidos los protocolos de observación y registro de datos, las campañas de avistamiento comenzaron en el último trimestre de 2023 como estaba previsto.

Los primeros meses de campaña pusieron de manifiesto el riesgo de la posible falta de datos necesarios para entrenamiento de los algoritmos de detección, en particular de soplos de cachalotes. Esto requirió ligeras modificaciones en los protocolos de campaña como medida de mitigación, así como la exploración de diferentes técnicas de aumentación de datos sobre las que se ha seguido trabajando en este último periodo.

Se analizaron otros riesgos relativos a la calidad de las imágenes y al sistema de estabilización de las cámaras, cuyo estudio y mitigación han continuado en este último periodo del proyecto, y de los que hablaremos en este informe, especialmente en lo relativo a las técnicas de preprocesamiento para la alineación de los diferentes sensores, y técnicas de postprocesamiento para la confirmación de detecciones incluso cuando la estabilización de la cámara no es óptima. En cualquier caso, se ha puesto de manifiesto la importancia del uso de técnicas de fusión RGB-T para mejorar la robustez de la detección en situaciones no óptimas para los sensores IR.

Los trabajos anteriores comprenden, asimismo, la selección del algoritmo de deep learning para la detección de objetos en tiempo real y el diseño preliminar de técnicas de fusión multimodal de datos RGB-T, siendo una de las principales innovaciones introducidas en la solución propuesta para detección de cetáceos.

Este es el punto de partida de las actuaciones realizadas en este último periodo, de enero de a diciembre de 2024, que se detallarán en el presente informe. Aunque no se recogía en los hitos de la planificación inicial, a mediados de 2024 se realizó un seguimiento con la Fundación Biodiversidad para comunicar resultados preliminares y asegurar que los trabajos se estaban realizando en la línea esperada.

2.2. CAMPAÑAS DE AVISTAMIENTO

Las campañas iniciadas en el último trimestre de 2023 continuaron durante 2024 hasta completar el número de embarques requerido, realizados por un equipo de 2 avisadoras entrenadas de la ULL. Las actividades de observación se efectuaron a bordo de dos navíos propiedad de la compañía Fred.Olsen (FO), Bajamar Express y Betancuria Express, que se equiparon con los equipos de cámaras Sentry como se ha descrito en informes anteriores.

La duración de los trayectos varía entre 1h 20 minutos (Bajamar Express, ruta Santa Cruz de Tenerife – Agaete) y 2h (Betancuria Express, ruta Las Palmas de Gran Canaria – Morro Jable), con una velocidad máxima de 38 nudos.

Durante la primera mitad de las campañas se pusieron de manifiesto dificultades en la logística y planificación de los viajes, especialmente en el trayecto de Las Palmas de Gran Canaria – Morro Jable (Fuerteventura) por su menor frecuencia de trayectos, así como trayectos en horario nocturno. Viendo que esto podía impactar en la planificación inicial, en mayo de 2024 se presentó un “Informe de Avance de las Campaña de Avistamientos” con una planificación alternativa más realista en la que, si bien no se conseguía un reparto del 50% entra las dos rutas, este era razonablemente equilibrado. Adicionalmente, se ofrecía un incremento de las rutas en un 5% lo que permite cubrir la diferencia de duración en tiempo de ambos trayectos. La planificación alternativa presentada y aprobada en mayo de 2024 por la dirección de los trabajos se resume en la Tabla 1.

Planificación original	BAJAMAR (Sta. Cruz de Tenerife-Agaete)	BETANCURIA (Las Palmas de Gran Canaria-Morro Jable)	TOTAL
Embarques	100	100	200
Planificación revisada	BAJAMAR (Sta. Cruz de Tenerife-Agaete)	BETANCURIA (Las Palmas de Gran Canaria-Morro Jable)	TOTAL
Embarques	130	80	210

Tabla 1. Planificación actualizada de campañas

El balance final de las campañas se resume en la Tabla 2, con lo que podemos concluir que se han cumplido con los objetivos en cuanto a las campañas de avistamiento.

Trayecto	Embarques realizados	Periodo	Duración
Bajamar (Sta. Cruz de Tenerife-Agaete)	132	1 noviembre 2023 a 31 octubre 2024	12 meses
Betancuria (Las Palmas de Gran Canaria-Morro Jable)	80	1 diciembre 2023 a 30 octubre 2024	11 meses
TOTAL	211		

Tabla 2. Balance final de campañas

2.2.1. Entregables de campañas de avistamiento

Los datos generados por las campañas de avistamientos, así como los registros adicionales proporcionados por el sistema Vigía de Fred.Olsen se incluyen en una archivo Excel adjunto a este informe (2023_EmbarquesFredOlsen_Diario_last) que incluye la siguiente información:

- Diario de embarques (ULL) (pestaña “Trayectos”).

Incluye el registro de los trayectos planificados (210) y realizados (211), incluyendo las cancelaciones por mal tiempo. En cada trayecto se indican las condiciones ambientales, de estado del mar y de visibilidad. Los avistamientos observados incluyen información completa conforme a los requerimientos de la licitación (posición GPS, distancia, especie, número de ejemplares, comportamiento, etc).

- Observaciones sistema Vigía (Fred.Olsen) (pestañas “Avistamientos”)-

Listado de observaciones adicionales proporcionado por Fred.Olsen y anotados por su tripulación tras recibir formación para el avistamiento de cetáceos. La información es más reducida que la que se recoge en las campañas de avistamientos de la ULL, y consiste principalmente en la posición del buque, distancia y rumbo de la observación, especie y número de ejemplares. Se incluye también información ambiental, aunque no puede garantizarse que esté siempre disponible. Estos datos son adicionales a los embarques requeridos en la licitación y no se tiene un control sobre su registro, por lo que se proporcionarán en el formato original proporcionado por Fred.Olsen.

- Volcado de datos SIG.

Archivo SIG generado que incluye los datos de los avistamientos de las campañas de la ULL según las especificaciones recogidas en el Protocolo SIG creado en el marco del proyecto LIFE IP INTEMARES. Estos datos han sido incluidos en la base de datos de carácter científico BIOTA, como alternativa a la base de datos REDPROMAR como mejora respecto a los requerimientos de la licitación, tal y como se acordó con la dirección de los trabajos.

2.3. CREACION DE DATASET PARA DETECCION

Una de las principales actuaciones durante el año 2024 ha sido la creación de un conjunto de datos, o dataset, para el entrenamiento, validación y test de los modelos de deep learning para detección de cetáceos.

Esta actuación ha requerido una gran cantidad de recursos no solo por lo laborioso de las tareas de etiquetado y curado de los datos, si no por los esfuerzos adicionales realizados para mitigar el problema conocido de falta de datos, que también se resumirán en este apartado.

2.3.1. Descripción general del dataset

El dataset consiste en un conjunto de imágenes etiquetadas mediante recuadros que indican la posición y clase de los objetos de interés. Para ello se han utilizado los videos obtenidos mediante el sistema de grabación instalado en los buques Bajamar y Betancuria Express, así como los registros de avistamientos proporcionados por Fred.Olsen y la ULL como referencia para localizar eventos de interés.

El dataset está formado por dos conjuntos independientes según el campo de visión o Field-of-View (FOV) de los sensores, de acuerdo con las características del sistema de captura y debido a que no siempre el mismo evento se registra en los dos campos de visión. Por tanto, tendremos un dataset NARROW (FOV nominal de 8º) y otro dataset WIDE (FOV nominal de 30º). Cada FOV contiene parejas de imágenes RGB-T (espectro visual e IR) del mismo fotograma, junto con las etiquetas o anotaciones y una serie de metadatos con información adicional. Las imágenes se extraen del video con una frecuencia de 10 fotogramas por segundo para maximizar la información disponible, por este motivo, un mismo objeto generará múltiples anotaciones en el dataset.

Hay que aclarar que en la forma en que están registrados los avistamientos, principalmente por su naturaleza y complejidad, estos deben interpretarse como un evento de interés que está formado por diferentes objetos en un determinado espacio de tiempo. Por ejemplo, si se avista un grupo de delfines lo habitual es que estos sean registrados en cámara durante un periodo de tiempo durante el que hay diferentes interacciones y, por tanto, pueden verse diferentes objetos de interés como aletas o cuerpos. La función del algoritmo de detección es detectar cada uno de estos objetos individuales que componen un evento o avistamiento.

Esto supone una diferencia entre la información proporcionada por los observadores MMO, afectados por la limitación humana para detectar y registrar en tiempo real múltiples objetos simultáneos o en un breve espacio de tiempo, y la proporcionada por los algoritmos de detección de objetos, que tienen una mayor capacidad para distinguir objetos individuales en un evento de avistamiento. Como veremos, esta diferencia será tenida en cuenta también en el proceso de evaluación del software de detección.

Por la forma en que funciona y se entrena un algoritmo de detección de objetos mediante aprendizaje supervisado, es necesario identificar y etiquetar individualmente estos objetos

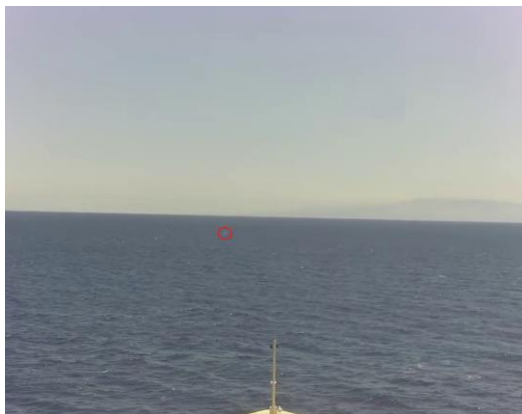
de interés, lo que conlleva un minucioso proceso de visionado, confirmación en video y etiquetado de los objetos correspondientes a un avistamiento, cuando solo se cuenta con la referencia general del evento y no de los objetos individuales que lo componen.

En cuanto a los objetos de interés, se han definido 3 categorías o clases de etiquetas, que consideramos más relevante para abordar el problema de detección (en la Figura 1 se proporcionan ejemplos de cada clase):

- Soplos
- Cuerpos
- Splash



a) Ejemplo soplo dataset TELE.



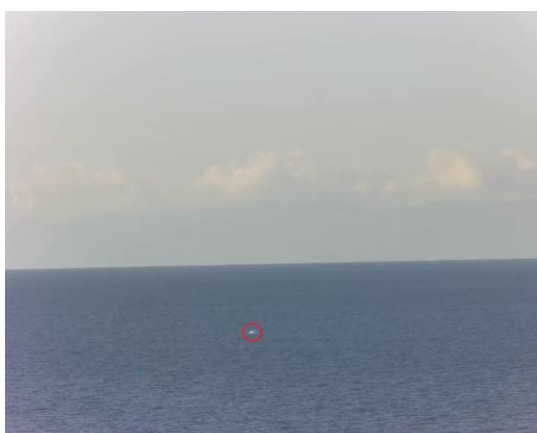
b) Ejemplo soplo dataset WIDE.



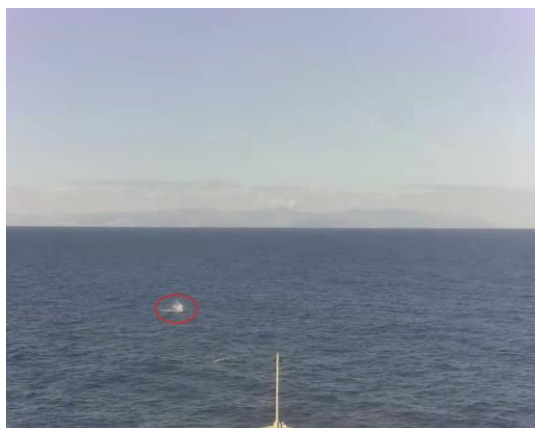
c) Ejemplo cuerpo dataset TELE.



d) Ejemplo cuerpo dataset WIDE.



e) Ejemplo splash dataset TELE.



f) Ejemplo splash dataset WIDE.

Figura 1. Ejemplo de objetos de las diferentes categorías: a) Soplo dataset TELE. b) Soplo dataset WIDE. c) Cuerpo dataset TELE. d) Soplo dataset WIDE. e) Splash dataset TELE. f) Splash dataset WIDE

En cuanto a los metadatos, se ha hecho la siguiente selección de metadatos que se consideran útiles tanto por trazabilidad como para el análisis del dataset y de los resultados:

- Marca de tiempo: fecha y hora.
- Archivo de video: subclip original que contiene la anotación.
- Fuente de la observación: Fred.Olsen, ULL, otros.
- Distancia de la observación (reportada por el observador).
- Temperatura del agua.
- Temperatura del aire.
- Estado del mar: plato, algunas ondas, pequeñas olitas, algunos borregos, borregos numerosos, marejada.
- Visibilidad del horizonte: buena, media, mala (horizonte no visible).
- Grupo de especie: cachalote, calderón, delfín, ballenas, otros.
- Facilidad de confirmación (apreciación del anotador): fácil, media, difícil, muy difícil.

El conjunto de datos está compuesto igualmente por cierto número de imágenes de fondo (Figura 2), es decir sin objetos de interés. Esto es necesario para que los modelos aprendan correctamente a diferenciar los objetos de interés, y con ello reducir el número de falsos positivos. El tipo de imágenes de fondo utilizadas contiene escenas con:

- Mar abierto con crestas de olas.
- Puerto y fondo de edificios o tierra firme.
- Embarcaciones.
- Pájaros.



Figura 2. Ejemplo imágenes de fondo.

2.3.2. Proceso de creación del dataset

Esta sección describe de forma resumida el flujo de trabajo para la creación del dataset, que se representa en la Figura 3.



Figura 3. Proceso de creación del dataset

Como se ha comentado, el flujo de trabajo se inicia con la visualización y confirmación de los avistamientos en video. Al contar con dos fuentes de avistamientos, las campañas de observadoras MMO de la ULL y los registros del sistema Vigía de Fred.Olsen, cada una con un formato de datos y variables diferentes, es necesario integrar dicha información mediante un preprocesado de los datos y una selección de metadatos para unificar en el dataset.

El resultado final es un registro de observaciones confirmadas en video con metadatos curados y unificados para las dos fuentes citadas. Este registro definirá el instante de inicio y de fin de los eventos a etiquetar, y se utilizará para generar subclips de los avistamientos. Los subclips son identificados por la marca de tiempo del avistamiento e incluidos en el registro de observaciones confirmadas lo que favorece la trazabilidad en la creación del dataset.

Una vez recortados los subclips por evento, se procede a su alineación temporal si fuese necesario y a su alineación espacial. La alineación espacial también se conoce como registro o co-registro, y es necesaria habitualmente cuando se dispone de imágenes multimodales o multicanal producidas por sensores diferentes, debido a las diferencias en las características del sensor (FOV, aberraciones ópticas, etc.) o en su ubicación (offset por paralaje). Trataremos este asunto por separado más adelante.

Finalmente, una vez que tenemos los subclips de los eventos (avistamientos) alineados, se procede al etiquetado de los diferentes objetos dibujando el recuadro alrededor de los mismos cada 10 imágenes o frames (Figura 4). Para el etiquetado y generación del dataset en el formato compatible con el detector de objetos utilizado, se ha utilizado el software de acceso abierto CVAT [1].

CVAT es un software de anotación avanzado para machine learning que permite trabajar con distintos formatos de datos y anotaciones. Su uso principal ha sido para el etiquetado de datos, integración de metadatos y control de versiones.

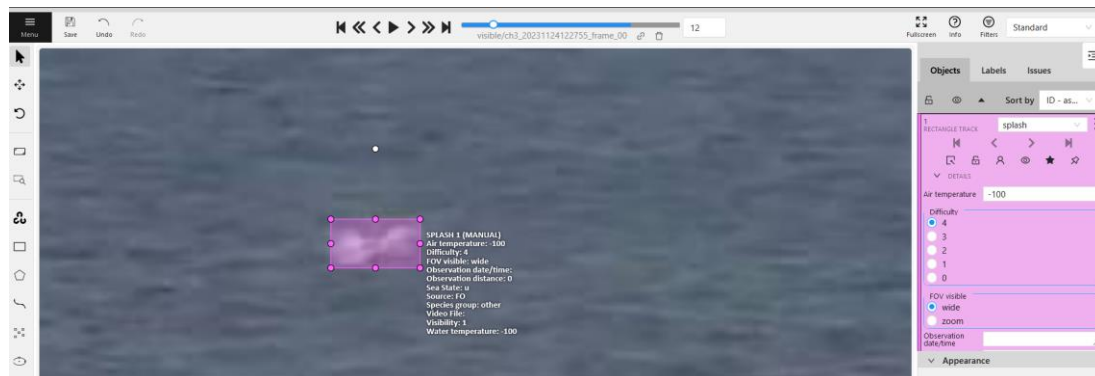


Figura 4. Etiquetado de objetos mediante CVAT

Una vez creadas todas las anotaciones es necesario dividir el conjunto de datos en tres subconjuntos, según su uso en el proceso de entrenamiento de los modelos:

- Entrenamiento. Utilizado para entrenamiento de los modelos.
- Validación. No utilizados para entrenar, pero usados en el proceso de entrenamiento para validar el rendimiento del modelo con un subconjunto independiente.
- Test o prueba. No utilizado durante el proceso de entrenamiento. Se usa para evaluar el rendimiento del modelo ya entrenado y reportar métricas cuantitativas.

Es muy importante destacar que el dataset se ha organizado en eventos o subclips, y no en frames. Esto permite mantener la independencia entre eventos en los diferentes subconjuntos, de forma que no se distribuyan frames del mismo evento en diferentes subconjuntos, evitando así el problema conocido como 'data leakage', o filtración de datos, que invalidaría las métricas de rendimiento en test ya que los frames de un mismo evento pueden ser muy similares entre sí. Es decir, el dataset de test estará formado por eventos que no se hayan visto durante el entrenamiento.

La división de los eventos en estos subconjuntos se ha hecho de forma automática y aleatoria, únicamente forzando el balanceo de cada clase entre los tres subconjuntos. Esto asegura que cada clase esté igualmente representada en el subconjunto de entrenamiento, validación y test, de forma que la evaluación en test sea más representativa.

2.3.3. Información detallada del dataset

La creación del dataset se ha realizado de forma incremental, para poder trabajar en paralelo en el desarrollo y mejora de los modelos de detección, y ampliándose a medida que se disponía de nuevos datos. Como referencia, se detallan las principales versiones y su evolución, incluyendo el número de ejemplos etiquetado en cada clase (Tabla 3):

- Versión 1.3. Diseño y refinamiento de los modelos de detección.
 - Versión mínima para una evaluación preliminar de los modelos.
- Versión 2.0. Validación del método de fusión frente a baselines.
 - Versión ampliada con las nuevas observaciones hasta la fecha.

- Aumento de ejemplos de fondo susceptibles de producir falsos positivos.
- Aumento de proporción de subconjunto de test para evaluación más precisa.
- Versión 3.0. Optimización final de los modelos de detección.
 - Incluye todos los datos disponibles de las campañas, excluyendo los datos reservados para la evaluación cualitativa.
 - Mejora de equilibrio de clases en los diferentes subconjuntos.
 - Re-etiquetado asistido por un modelo propio de detección de cetáceos.

TELE			
	V1.3	V2.0	V3.0
Body	2206	4008	4767
Blow	359	365	365
Splash	1293	1436	1445
WIDE			
	V1.3	V2.0	V3.0
Body	1635	1847	2424
Blow	83	83	83
Splash	406	406	406

Tabla 3. Número de etiquetas por clase para diferentes versiones del dataset

A continuación, se presenta un análisis detallado de la última versión del conjunto de datos mediante una serie de gráficas que resumen las características generales y la distribución en función de diferentes variables ambientales clave como la temperatura del aire, la temperatura del agua, la visibilidad del horizonte y el estado del mar. Además, se incluye información sobre los niveles de dificultad asociados a las anotaciones, y de la distribución de clases y etiquetas en el conjunto de datos.

El primer análisis es el relativo al número de eventos y anotaciones. Como se ha detallado en el apartado anterior, los avistamientos o eventos se organizan en subclips, de los cuales se extraen los frames que son etiquetados. Esto nos permite maximizar el dataset aumentando el número de anotaciones por evento, aunque tiene una limitación en cuanto a la diversidad de los datos ya que los frames próximos de un mismo evento tendrán características similares.

Como se puede observar en la Figura 5, en principio el número de anotaciones para cuerpos es suficiente para un problema de detección de objetos, donde se recomienda alrededor de 1000 ejemplos por clase para el ajuste fino de un modelo pre-entrenado en un dataset general como COCO [2]. Sin embargo, el número de subclips que se muestra en la Figura 6 es mucho más limitado. El caso de los soplos es más grave, porque ni siquiera se llega a un mínimo de 1000 anotaciones, con únicamente 12 eventos diferentes en el dataset TELE y 3 en WIDE. La clase 'splash' solo se ha incluido por referencia, ya que no se usó finalmente para la detección.

Estos resultados ponen de manifiesto uno de los principales desafíos para el desarrollo y evaluación de los algoritmos de detección, la escasez de datos y la falta de diversidad de estos. Esto puede producir dificultades para que el modelo de detección generalice en situaciones diferentes a las observadas durante el entrenamiento.

Asimismo, se observa un desequilibrio de clases, donde la clase cuerpo - body – es mayoritaria. En general este tipo de desequilibrios va a hacer que el modelo de detección está sesgado, dando mejores resultados para la clase mayoritaria. En nuestro caso, esto no supone un problema grave ya que el objetivo es la detección de eventos, independientemente del tipo.

Cabe destacar por último la diferencia entre los conjuntos de datos TELE y WIDE, siendo el primero considerablemente más amplio en términos de volumen de datos.

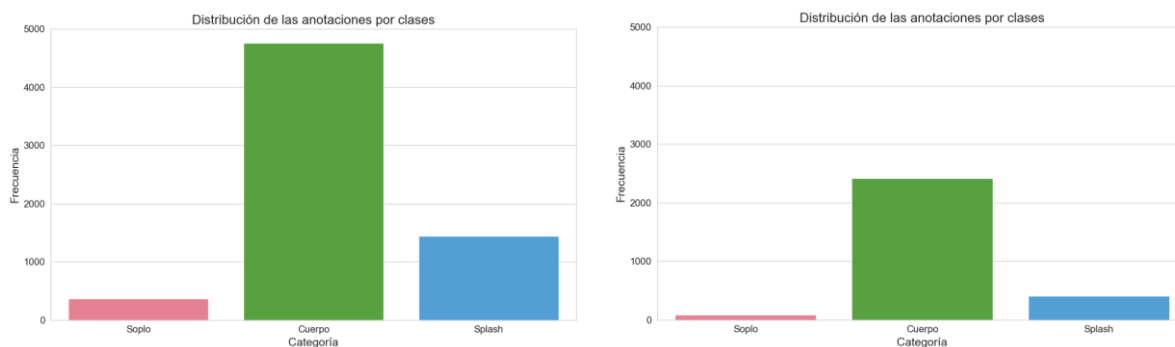


Figura 5. Distribución de las anotaciones por clase. Dataset TELE a la izquierda y WIDE a la derecha.

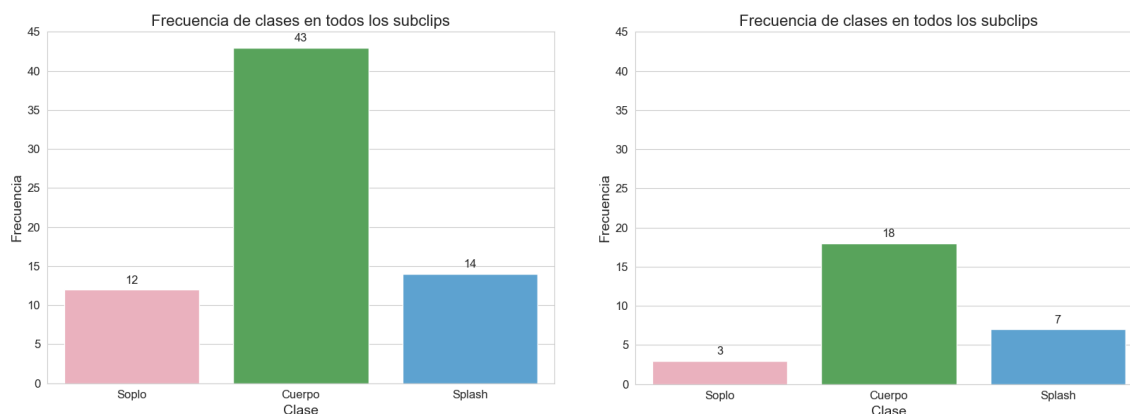


Figura 6. Distribución de los subclips por clase. Dataset TELE a la izquierda y WIDE a la derecha.

El análisis del tamaño de las etiquetas también es significativo para entender la complejidad del problema de detección de cetáceos en ferris. La Figura 7 se presenta un diagrama de dispersión que muestra el tamaño de las anotaciones. Se pueden observar tres patrones:

- Los cuerpos tienden a ser objetos pequeños o extremadamente pequeños.
- Los soplos son los elementos más altos dentro del dataset, lo cual era de esperar.
- Los splash superan al resto de las clases en su dimensión horizontal.

La conclusión principal, es que el tamaño de los objetos de interés es relativamente pequeño, lo que supone un gran desafío para los algoritmos de detección.

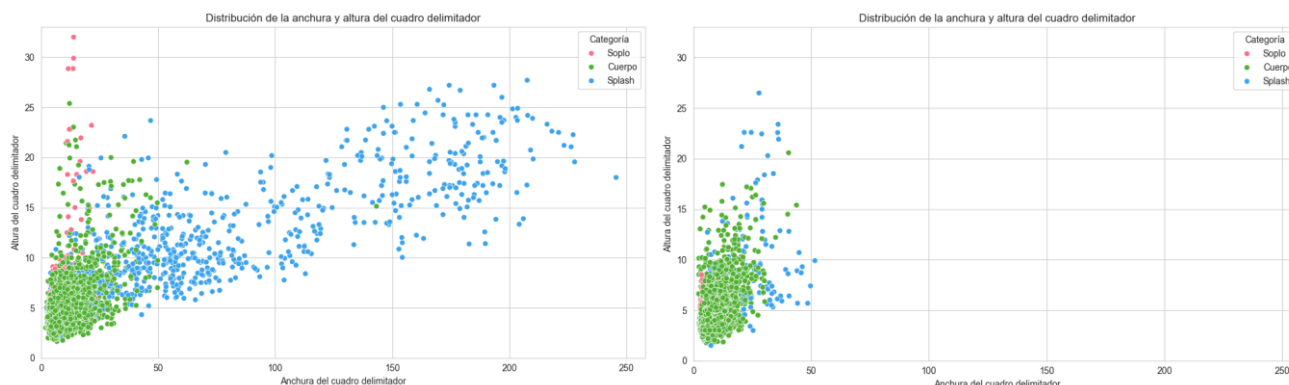


Figura 7. Distribución del tamaño de los recuadros de detección por clase.
Dataset TELE a la izquierda y WIDE a la derecha.

A continuación, se realiza un breve análisis de la distribución de los eventos en función de variables ambientales. La distribución de la temperatura del aire (Figura 8) y del agua (Figura 9) muestran lo que esperábamos para la zona de Canarias, temperaturas cálidas durante todo el año, lo que supone un desafío para la detección de cetáceos al existir un menor contraste térmico entre su temperatura corporal y la exterior. Como es lógico, los resultados del presente trabajo deben limitarse al contexto geográfico y estacional del estudio, ya que las condiciones específicas de las islas Canarias pueden no ser extrapolables a otras áreas donde las variaciones térmicas o ambientales sean diferentes.

Dentro de este rango de temperaturas cálidas, la distribución no es uniforme para los eventos observados, con una distribución monomodal en la temperatura del aire y un pico en 24 grados, y una distribución multimodal en la temperatura del agua con varios picos. Esto podría producir igualmente sesgos en el algoritmo de detección hacia esos picos de temperatura, aunque no consideramos que sean relevantes ya que es un rango de temperaturas relativamente pequeño y las características de los eventos no deberían variar tanto en ese rango, aunque habrá que tenerse en cuenta en los casos más extremos.

Nótese igualmente que la gráfica de temperatura del agua presenta un pico inusualmente alto en 27 grados. Un análisis más detallado revela que estos datos corresponden con mediciones realizadas por Fred. Olsen en los meses de septiembre/octubre, donde no tenemos control sobre los instrumentos de medida. Dado que la compañía no ha podido garantizar su exactitud, podríamos considerarlo como un error de medida. Hemos confirmado esta hipótesis contrastando con datos de temperatura del agua publicados por Puertos del Estado (Figura 10) para un punto SIMAR (modelo con asimilación de datos) en medio de la ruta Santa Cruz - Agaete, donde se confirma que la temperatura no sube por encima de 25 grados, presentando una distribución bimodal consistente con nuestras mediciones.

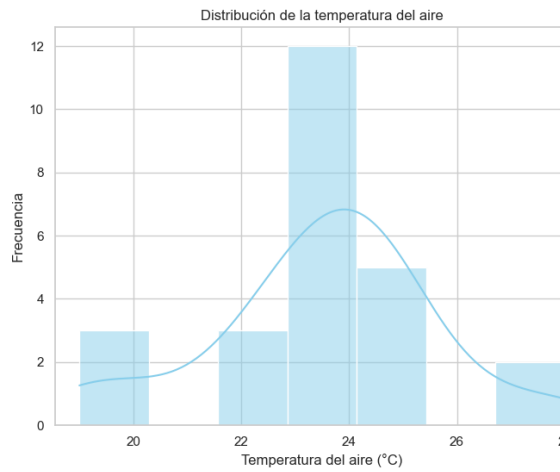


Figura 8. Distribución de la temperatura del aire.

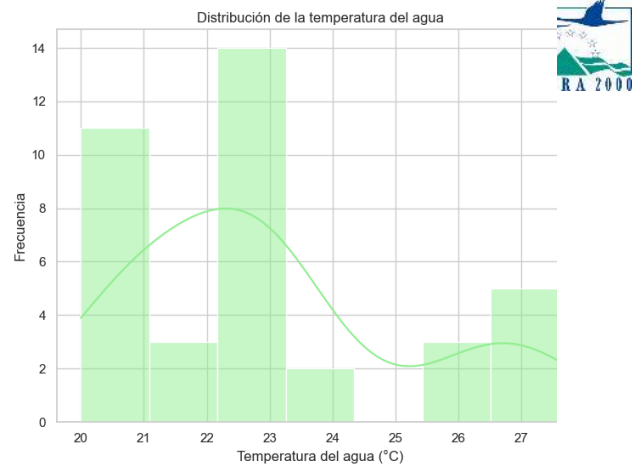


Figura 9. Distribución de la temperatura del agua.



Figura 10. Distribución de la temperatura del agua (Puertos del Estado)

La distribución de la visibilidad del horizonte (Figura 11) muestra que la mayoría de los avistamientos ocurrieron en condiciones de buena visibilidad, mientras que los avistamientos en condiciones de visibilidad media y mala fueron considerablemente menores. Esta situación si puede tener un mayor impacto en las capacidades de detección del algoritmo, sobre todo cuando hemos evidenciado en las grabaciones realizadas un peor funcionamiento del sensor IR en condiciones de niebla y calima. En este sentido podemos esperar un mejor comportamiento del modelo de detección en condiciones de buena visibilidad. Por otro lado, estos datos están condicionados por las propias capacidades de los propios observadores para detectar los avistamientos en condiciones de mala visibilidad.

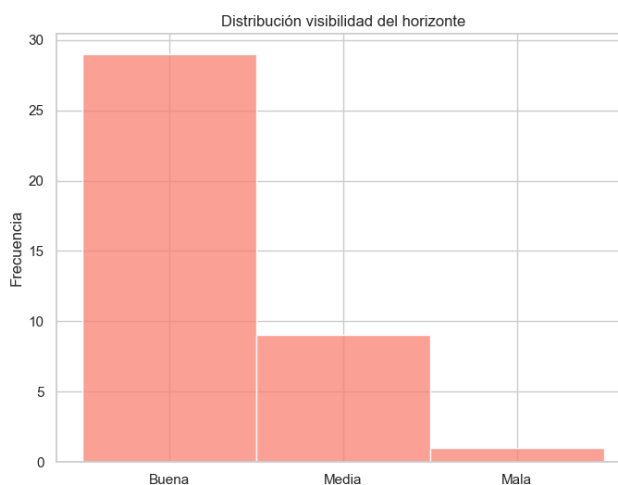


Figura 11. Distribución visibilidad del horizonte.

La distribución del estado del mar (Figura 12) durante los avistamientos muestra que la mayoría ocurrieron en condiciones de mar ligeramente agitado, con condiciones desde “pequeñas olitas” hasta “borregos”. Esto indica que los cetáceos fueron detectados en una variedad de estados del mar, siendo las condiciones de mayor oleaje las más comunes. Aunque intuitivamente los días de mar plato deberían ser más favorables para el avistamiento de cetáceos, estos resultados podrían deberse a una combinación de factores, entre ellos la mayor ocurrencia de días con mar agitado en las zonas observadas y la influencia del diseño de muestreo. Dado que el muestreo puede haberse realizado en zonas o en momentos donde el mar tiende a estar más agitado, esto pudo sesgar los resultados hacia este tipo de condiciones.

En cualquier caso, la detección en días de mar agitado presenta un mayor desafío para el algoritmo de detección ya que como se ha evidenciado en estudios anteriores [3], [4] las crestas de olas son causas habituales de falsos positivos, por lo que contar con un mayor número de ejemplos en estas condiciones puede ser beneficioso.

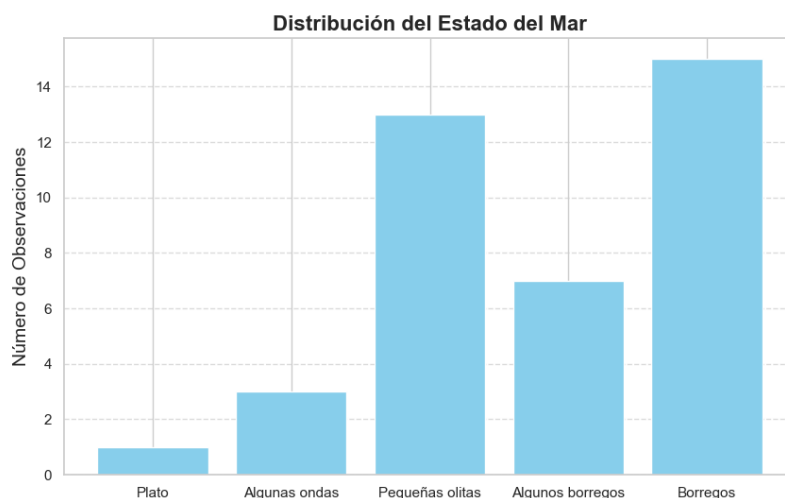


Figura 12. Distribución del estado del mar.

Por último, se incluye una evaluación cualitativa de la dificultad de anotación según la apreciación subjetiva del anotador durante la creación del dataset (Figura 13, datos extraídos del anexo en Excel "Evaluacion_cualitativa"). Aunque se trata de una evaluación subjetiva, proporciona información relevante sobre el esfuerzo requerido para cada anotación y posible dificultad de la detección. La mayor proporción de anotaciones corresponde a un nivel de dificultad 'media', con un 38.5%, mientras que los niveles 'fácil', 'difícil' y 'muy difícil' se distribuyen equitativamente, representando cada uno un 20.5% del total. En general, observamos una distribución bastante uniforme en cuanto al grado de dificultad, lo cual es positivo.



Figura 13. Proporción de niveles de dificultad.

2.3.4. Dificultades en la confirmación visual

Se han identificado dificultades en la confirmación visual de los avistamientos en video, un paso que, como se ha mencionado, es fundamental en el proceso de creación del dataset. Tras realizar un análisis detallado de las posibles causas, se concluyó que estas dificultades podrían deberse a una combinación de diversos factores:

1. **Retraso en el apuntado de la cámara tras una observación.** Desde el momento en que el observador detecta un avistamiento hasta que se comunica con el puente para que se apunte la cámara en su dirección, puede suceder que el evento observado ya no esté visible o la posición haya cambiado fuera del FOV de la cámara debido al movimiento del buque. Este retraso es especialmente crítico en el caso de los soplos, ya que pueden transcurrir varios minutos hasta que el cetáceo emita otro soplo, lo que reduce las oportunidades de capturar en video la interacción.
2. **Desajustes de tiempo.** Posible desajuste entre el instante de tiempo registrado en las observaciones y las grabaciones, que dificultarían la identificación de las observaciones en los videos.
3. **Desventaja de los prismáticos frente a las cámaras:** Se utilizan dos tipos de prismáticos en las campañas de avistamiento, unos pequeños (7x, 7-8º FOV aprox.) y otros de mayor aumento los Big Eye (x15, 4º FOV aprox.) que permiten observar

detalles a mayor distancia. Sin embargo, las cámaras tienen un campo de visión más amplio lo que se traduce en un menor zoom, lo que podría afectar a la confirmación visual en video de las observaciones más lejanas. A esto hay que sumar las limitaciones observadas en los sensores IR para eventos lejanos debido a la atenuación en el espectro infrarrojo, especialmente en días de niebla o calima.

Tras este análisis se aplicaron una serie de acciones tanto para mitigar el problema como para ayudar en el proceso de confirmación visual de las observaciones en video:

1. **Mejorar la comunicación con el puente de mando.** Para reducir los retrasos en el ajuste de la cámara tras una observación se modificó el protocolo de comunicación con el puente de mando para hacerlo más rápido y directo. Esto incluye la formación de la tripulación para seguir el protocolo y para el manejo de la cámara. Estos cambios ayudaron a reducir los tiempos para el apuntado de la cámara.
2. **Revisar la sincronización de los equipos.** Se modificó el procedimiento de operación para incluir una revisión sistemática de la sincronización de los equipos implicados en las campañas, anotando posibles diferencias para trazabilidad: GPS de las observadoras, hora del sistema vigía de Fred. Olsen y hora del sistema de grabación NVR.
3. **Anotar el tipo de prismático usado para el avistamiento.** Con el fin de proporcionar más información para la confirmación visual en video, junto con la información de distancia anotada, y poder identificar aquellos ejemplos difíciles de confirmar.
4. **Valoración del observador.** Se añade un campo con la valoración subjetiva del observador sobre si el evento ha sido capturado en cámara con un objetivo similar al del punto anterior.
5. **Asistencia en el proceso de visualización.** Se desarrolló una implementación del algoritmo STA/LTA utilizado por Zitterbart en trabajos previos (REF) para pre-detección mediante cambios de contraste. Este algoritmo permite obtener candidatos de baja calidad con alta sensibilidad, que deben ser confirmados posteriormente. Para la implementación se utilizó la librería Obspy [5] y se utilizó como herramienta de asistencia en el proceso de confirmación visual de observaciones al preseleccionar y filtrar automáticamente los datos con mayor probabilidad de ser una detección, lo que reduce significativamente la cantidad de información a procesar y almacenar.

Pese a estos esfuerzos, el ratio de confirmaciones en video no mejoró significativamente. A nuestro entender, esto pone de manifiesto las limitaciones de captura de datos mediante cámaras instaladas a bordo de ferris rápidos durante el curso de su ruta comercial. Entendemos que el tiempo de operación de la cámara para el apuntado hacia observaciones inicialmente fuera de su campo de visión es crítico, debido a la corta duración de los eventos y a la velocidad de desplazamiento de los buques. Esto es especialmente sensible en avistamientos en los laterales del barco, donde el apuntado de la cámara es realmente difícil, ya que el rumbo de la observación respecto al barco cambia rápidamente.

Hay que tener en cuenta que en trabajos previos del estado del arte [3], donde era necesario recopilar un elevado número de ejemplos en video, se utilizaron barcos científicos cuyo objetivo principal era la observación de cetáceos, que podían ajustar el rumbo y velocidad según lo requiriese la observación y que podían seguir a grupos de individuos para obtener

mayor cantidad de ejemplos. En otros trabajos, se utilizaron posiciones fijas desde tierra en zonas de paso frecuente de cetáceos [4].

2.3.5. Co-registro de imágenes

Uno de los principales desafíos en la detección multimodal de objetos es el desalineamiento entre las imágenes captadas por los diferentes sensores o modalidades, en este caso RGB (visual) e IR (infrarrojo). Un ejemplo de este problema puede observarse en la Figura 14. Este desalineamiento ya fue identificado en el informe previo (E.2), y posteriormente se realizó un análisis más detallado, así como esfuerzos específicos para mitigarlo.

El desalineamiento es un problema común cuando se emplean diferentes sensores, particularmente si el diseño del sistema óptico no ha sido optimizado específicamente para evitarlo. Las causas pueden ser diversas:

- Efecto de paralaje: Debido a la diferente ubicación de los sensores RGB e IR, los objetos aparecen ligeramente desplazados entre fotogramas, especialmente aquellos cercanos a las cámaras. Este efecto es mayor cuanto más separados estén los sensores.
- Diferencias en campo de visión (FOV) y Longitud Focal: Variaciones en el FOV y las longitudes focales provocan diferencias de escala en los objetos, afectando la consistencia de tamaño y alineación entre fotogramas.
- Distorsión de lentes y limitaciones de calibración: Las distorsiones varían según el lente, y aunque la calibración las reduce, persisten pequeñas inconsistencias, particularmente en los bordes de las imágenes.
- Diferencias de resolución: La ampliación de imágenes térmicas de menor resolución a la resolución de RGB introduce imprecisiones a nivel de píxeles, especialmente en los bordes de los objetos.
- Variaciones ambientales: factores ambientales pueden afectar de forma diferente a señales en luz visible (RGB) y espectro infrarrojo (térmico).



Figura 14. Ejemplo de desalineación espacial. A la izquierda la imagen RGB, en el centro la imagen IR y a la derecha ambas imágenes superpuestas.

A continuación, se detallan los diferentes esfuerzos realizados para resolver el problema y los resultados obtenidos en cada caso. El estudio comenzó con una aproximación simple, seguida por la aplicación de técnicas clásicas de homografía, y por último incorporando soluciones del estado del arte basadas en deep learning.

- Aproximación simple: escalado con puntos distantes

El primer enfoque aplicado con el objetivo de realizar el co-registrado de las imágenes fue una aproximación simple centrada en el **escalado a partir de puntos distantes de referencia**, técnica que se basa en utilizar puntos lejanos en la escena como referencia para ajustar el tamaño de la imagen. La idea es seleccionar puntos que estén a una distancia considerable de la cámara, donde las variaciones de profundidad tengan un impacto mínimo. De esta manera, al utilizar estos puntos distantes como referencia, se puede escalar la imagen garantizando que las posiciones relativas de dichos puntos se mantengan tras el escalado. El primer paso en este proceso es la extracción manual de los puntos de referencia en las imágenes IR y RGB (Figura 15).

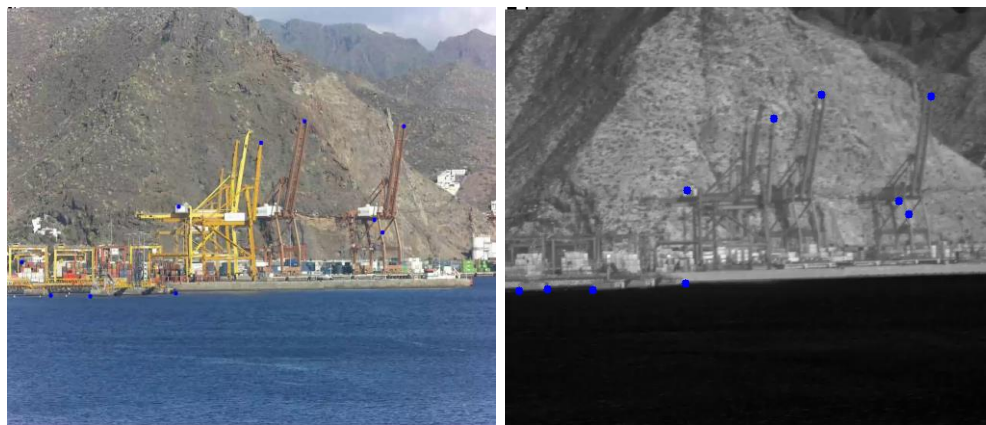


Figura 15. Selección manual de puntos en ambas modalidades.

Una vez han sido tomados los puntos de referencia, se calcula la media de la distancia entre los puntos en distintas modalidades, obteniendo el factor de escala para la imagen RGB en los ejes x e y. Tras escalar la imagen RGB, quedaría un desplazamiento entre los puntos y sus posiciones respectivas, por lo que el siguiente paso consistiría en calcular dicho desplazamiento para ambos ejes (Figura 16).

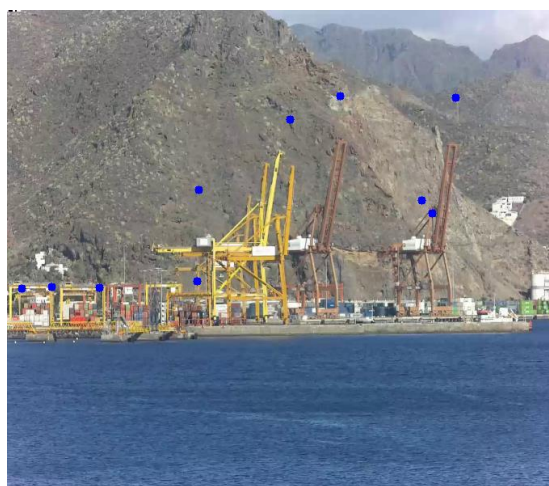


Figura 16. Imagen RGB escalada, pero con los puntos desplazados.
Los puntos azules representan las referencias en el modo IR

Tras aplicar el desplazamiento calculado a cada uno de los puntos, estos pueden ser reajustados. En los ejemplos de la Figura 17 se aprecia como los puntos de referencia del modo IR (en azul) coinciden con las referencias en RGB.

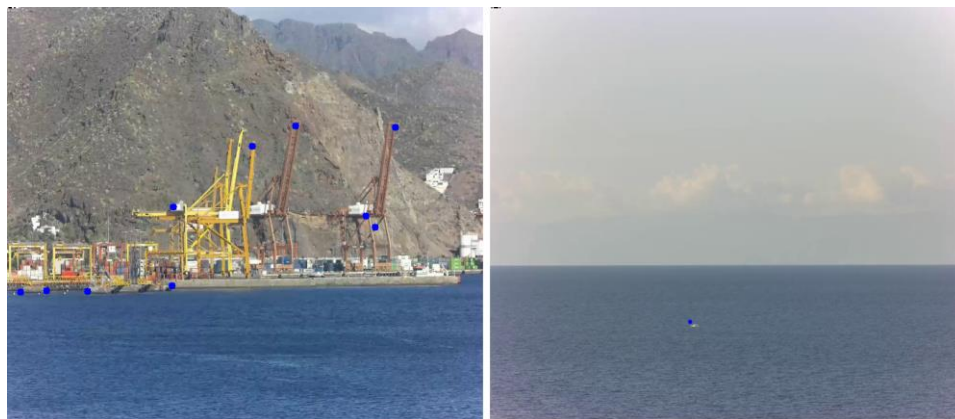


Figura 17. Imágenes RGB tras escalar y eliminar el desplazamiento.
Los puntos azules representan las referencias en el modo IR tras la corrección

Aunque el factor de escalado y los valores de desplazamiento parecen ser extrapolables a otras imágenes, los resultados del co-registro obtenidos mediante este método no logran alcanzar el nivel de precisión esperado cuando se evalúa con un conjunto más amplio de imágenes.

- Aproximación clásica basada en características:

Una vez descartado el enfoque más básico se decidió aplicar un flujo de trabajo clásico de co-registro de imágenes que implica cuatro pasos principales:

1. **Extracción de características:** Se identifican puntos de interés mediante características distintivas en la imagen, que pueden ser detectados manual o automáticamente. En el caso de aplicar detección automática se utilizan algoritmos robustos para la extracción de características como ORB [6] o SIFT [7].
2. **Emparejamiento de características:** Se establecen correspondencias entre puntos a través de las características extraídas. Para realizar dicha labor se suelen utilizar algoritmos como RANSAC o sus variantes.
3. **Estimación del modelo de transformación:** Se diseña una función de mapeo que ajusta la imagen para superponerla correctamente sobre la imagen de referencia a partir del emparejamiento de puntos previo. Esta función es habitualmente una matriz de homografía o bien una matriz de transformación afín.
4. **Transformación de la imagen:** La imagen capturada se ajusta a la imagen de referencia utilizando técnicas de interpolación, minimizando los artefactos y asegurando una alineación precisa.

El principal problema que se ha encontrado para aplicar este flujo de trabajo es la extracción de características distintivas y emparejamiento de puntos en mar abierto. Mientras que en zonas de puerto o cerca de costa, los algoritmos clásicos son capaces de extraer características distintivas y hacer emparejamiento de puntos, en mar abierto apenas se emparejan puntos o estos emparejamientos son erróneos.

Si analizamos este tipo de imágenes, se observa que son escenas muy monótonas con pocos objetos o características destacables. A esto hay que añadir la dificultad añadida de emparejamiento de puntos entre distintas modalidades, ya que las características y texturas en RGB a IR son muy diferentes, y en este caso la imagen IR es habitualmente más ruidosa y de peor calidad. Aunque el registro multimodal es un problema resuelto en algunos campos como la teledetección por satélite, la distancia de los objetos al sensor varía poco, en general es un problema abierto y un objeto de estudio en la literatura actual [8].

Para mejorar el emparejamiento de puntos se prueban diferentes alternativas:

1. **Detección de bordes.** En vez de utilizar las imágenes originales para la detección de características, se utiliza una representación simplificada tras un procesado de detección de bordes. Se utilizan técnicas como el algoritmo de Canny [9] para identificar los contornos de los objetos en la imagen. Esta técnica destaca los cambios en la intensidad de los píxeles, resaltando los bordes más importantes (Figura 19). Sin embargo, como se aprecia en la Figura 18 la información de bordes por lo general es muy diferente en cada modalidad, lo que hace que esta solución no dé buenos resultados.

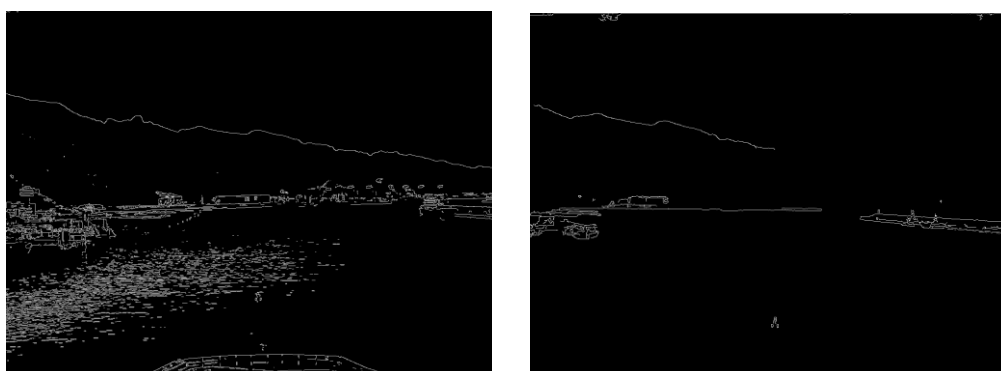


Figura 18. Detección de bordes. RGB (izquierda), IR (derecha)

2. **Optical Flow.** El Optical Flow es una técnica que estima el movimiento de píxeles entre dos imágenes consecutivas, modelando cómo cambia la posición de los objetos en el tiempo. Se calcula como un campo de vectores que indica la dirección y magnitud del desplazamiento de cada píxel. Aunque su aplicación principal es el análisis de video, esta técnica puede usarse también para el registro de imagen, indicando correspondencia entre pares de puntos si lo que recibe como entrada son las dos imágenes para registrar.

Se hicieron diversas pruebas aplicando la técnica de Optical Flow mediante el método de Lucas-Kanade [10], tanto directamente como para puntos concretos extraídos mediante un detector de esquinas mediante el algoritmo de Shi-Tomasi [11]. Sin embargo, como en el caso del emparejamiento mediante extractores de características, los resultados no fueron satisfactorios en mar abierto.

3. **Emparejamiento manual y transformación afín.** Ante la dificultad de los métodos anteriores para el emparejamiento automático de puntos, se analiza una solución de compromiso mediante un emparejamiento manual en varias escenas y el cálculo de una única matriz de transformación. En este caso se obtienen mejores resultados con una matriz transformación afín en lugar de una matriz de homografía.

No es una solución óptima ya que idealmente el emparejamiento debe calcularse en cada frame, o cada ciertos frames, pero es la única forma de hacer un emparejamiento manual y fiable. Los resultados son variables, con algunos ejemplos donde se mejora considerablemente el alineamiento y otros con resultados menos favorables. La Figura 19 y Figura 20 muestran varios ejemplos en las ópticas Tele y Wide respectivamente, el recuadro verde indica la posición del objeto de interés en cada uno de los dos modos antes del registro (izquierda) y después (derecha).

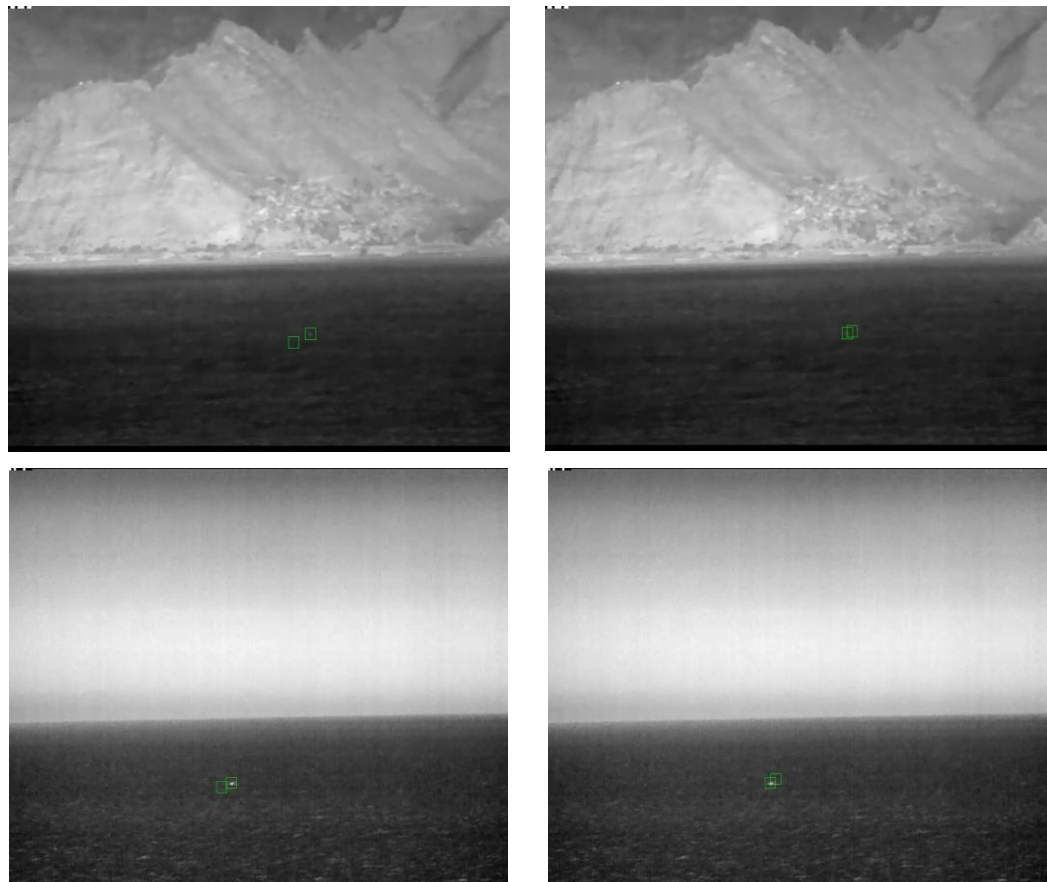


Figura 19. Imagen transformada aplicando homografía (TELE)
Antes del registro (izquierda) y después (derecha)

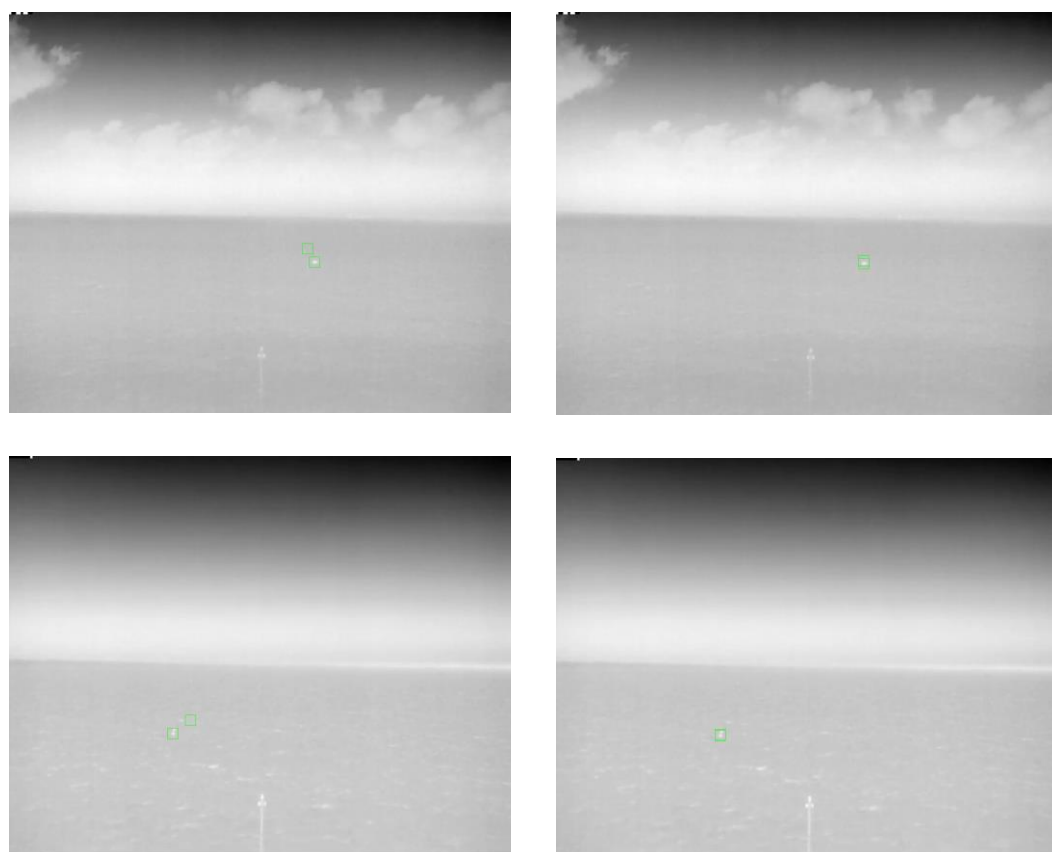


Figura 20. Imagen transformada aplicando homografía (WIDE)
Antes del registro (izquierda) y después (derecha)

- Co-registro en el dominio de Fourier

El método de registro de imágenes en el dominio de Fourier [12] se utiliza para alinear imágenes basado en un análisis de las componentes en frecuencia de las mismas (Figura 21). Este proceso es automatizado y no depende de características específicas de la imagen, lo que lo hace útil en una amplia gama de situaciones. Se utiliza principalmente cuando se requiere corregir traslaciones, rotaciones y escala de forma global en la imagen, mientras que la transformación mediante homografía permite hacer correcciones más complejas de perspectiva. Los resultados fueron variables, con algunos casos en los que se mejoraba el alineamiento tras el registro (Figura 22 a) y otros donde se empeoraba (Figura 22 b).

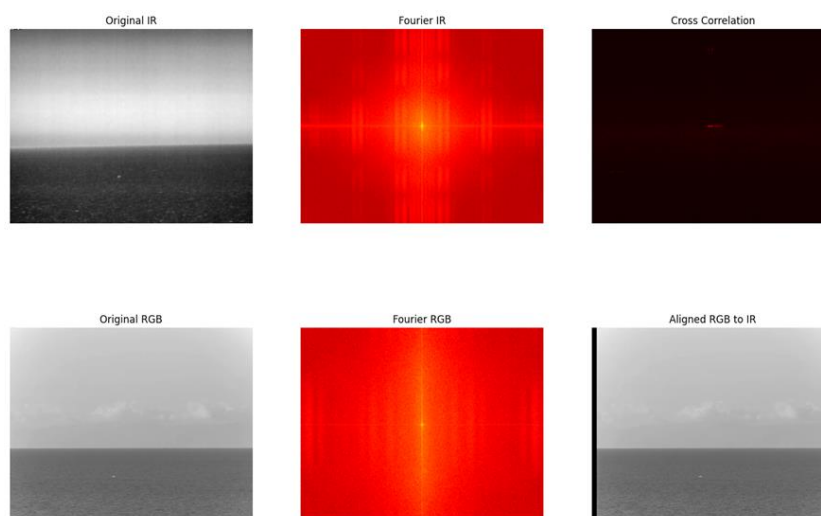
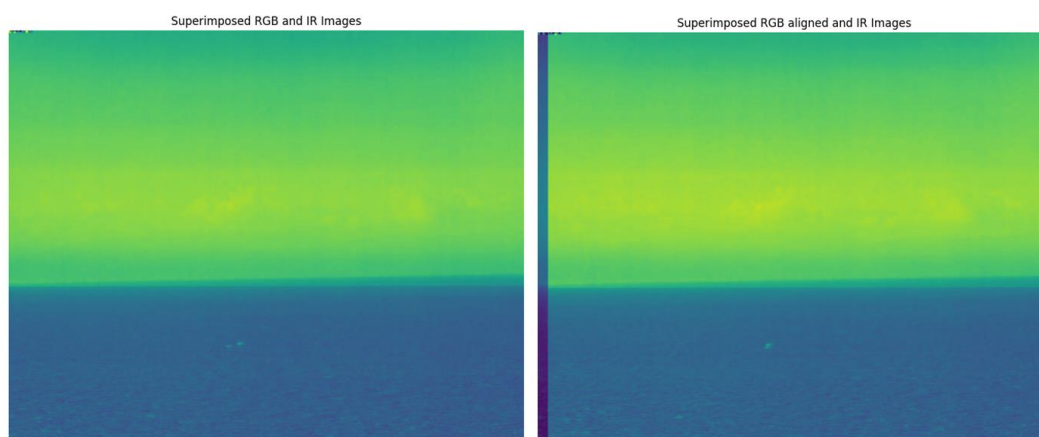


Figura 21. Registrado en el dominio de Fourier.



a) Soplo en mar abierto



b) Imagen de puerto

Figura 22. Registrado en el dominio de Fourier.
Imágenes RGB e IR superpuestas. Antes del registro (izquierda) y después (derecha)

- Métodos avanzados mediante deep learning

Un enfoque común es el uso de redes neuronales para el emparejamiento de puntos clave. Estas redes se entrenan para detectar automáticamente características relevantes en las imágenes, sustituyendo los métodos tradicionales de detección de puntos, como SIFT o ORB. Posteriormente, una vez emparejados dichos puntos, se aplica el flujo de trabajo clásico para homografía. Este enfoque ha mostrado mejoras significativas en la robustez y precisión del registro, sobre todo en escenarios complejos, donde los métodos clásicos fallan debido a la falta de características fácilmente detectables.

En este sentido se evaluó el método SSG-Net [13]. Este sistema utiliza una red neuronal basada en la arquitectura SuperGlue [14] para localizar puntos clave en las imágenes y generar un registro inicial. Luego, ajusta estos puntos de forma más precisa utilizando un algoritmo especializado, lo que mejora la exactitud del proceso (Figura 23).

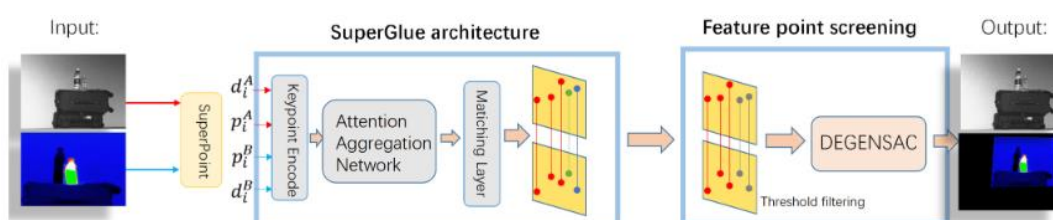


Figura 23. Arquitectura de SSG-Net.

La publicación original del método afirma que es eficaz en escenarios complicados, como cuando hay un gran paralaje o interferencias entre el objeto y el fondo de la imagen. Sin embargo, en las pruebas realizadas se obtuvieron resultados variables y no logró resolver el emparejamiento en escenas de mar abierto. Este problema se observó especialmente cuando las condiciones de iluminación o calidad de la imagen IR eran malas. También se comprobó que aun en escenas donde se logra emparejar suficientes puntos, muchos eran erróneos, lo que puede indicar un funcionamiento no optimizado para registro multimodal. La Figura 24 y Figura 25 muestran dos ejemplos donde SSG-Net tuvo problemas, en mar abierto y en puerto respectivamente. Se muestra con diferentes colores la confianza en el emparejamiento, una línea roja indica alta confianza, mientras que una línea verde señala baja confianza.

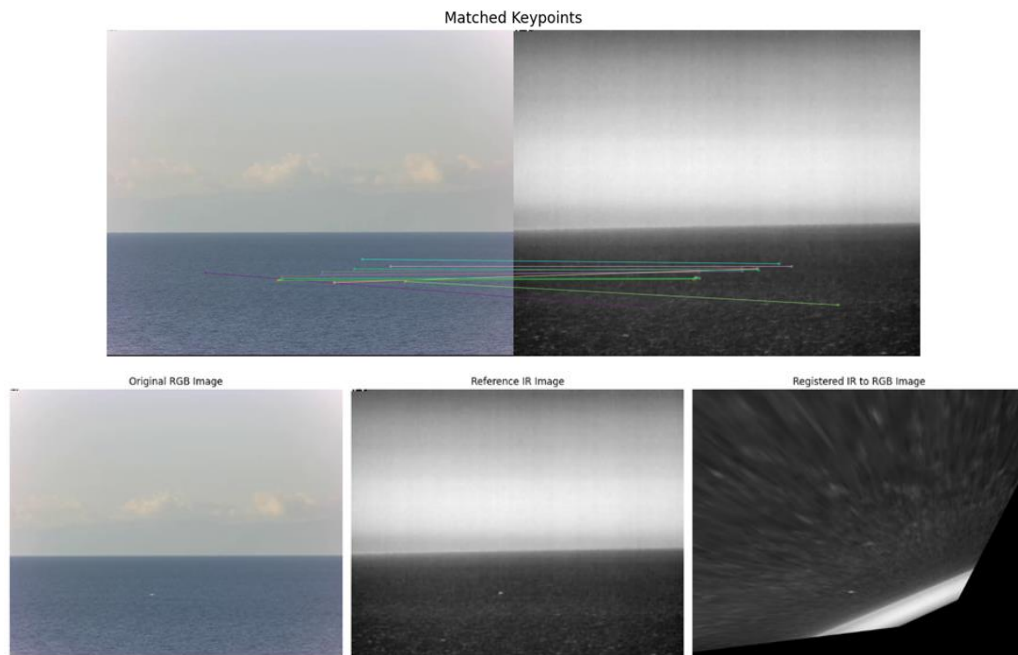


Figura 24. Emparejamiento y registrado en mar abierto usando SSG-Net.

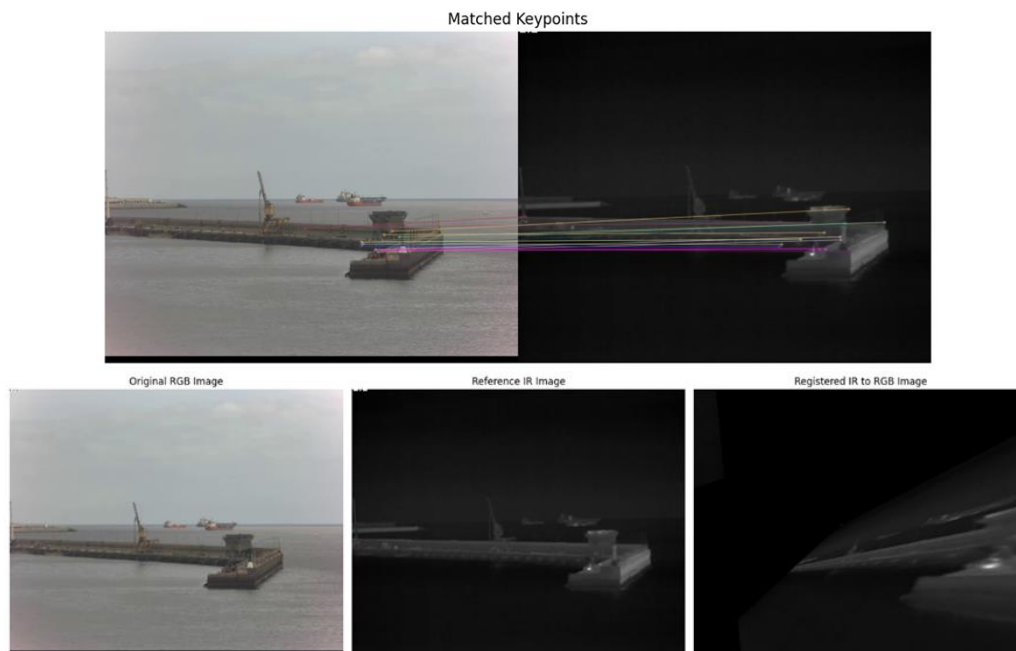


Figura 25. Emparejamiento y registrado en puerto usando SSG-Net.

Otra línea de investigación dentro del aprendizaje profundo para el co-registro se centra en el uso de redes neuronales que aplican el proceso de co-registro de manera directa, sin la necesidad de detectar previamente puntos clave. Estas arquitecturas aprenden a generar las transformaciones necesarias para alinear las imágenes a partir de los datos de entrenamiento, contando con un mayor grado de libertad que una transformación homeográfica. Sin embargo, estos métodos habitualmente requieren re-entrenamiento o ajuste fino para adaptarse a casos de uso particulares, lo cual puede ser un desafío al requerir un conjunto de imágenes registradas ground-truth.

En este sentido, se estudió el método del estado del arte llamado C-MPDR [15] para el registro de imágenes multimodales, en particular en espectro infrarrojo y visible. Los resultados obtenidos en un dataset RGB-T de detección de peatones son prometedores, aunque existen dudas sobre su aplicabilidad a nuestro caso de uso. Una de las principales inquietudes radica en su solución para reducir las discrepancias entre modalidades, la cual genera una imagen pseudo-infrarroja a partir del modo RGB utilizando la red CPSTN, una red de tipo 'style-transfer' propuesta por [16]. Este enfoque generativo, aunque innovador, generalmente produce resultados cuestionables, generando imágenes que visualmente se asemejan a una imagen infrarroja pero que carecen de coherencia con la información térmica real de esta modalidad. En cualquier caso, no fue posible realizar pruebas ya que el modelo pre-entrenado no estaba disponible y entrenar este tipo de métodos desde cero habría requerido un esfuerzo significativo con una alta incertidumbre.

En conclusión, el método seleccionado para el co-registro de imagen fue la **transformación afín con selección manual de puntos**, al ser la que obtuvo mejores resultados en general. No obstante, es importante destacar que este sigue siendo un problema abierto, y la solución aplicada debe considerarse como una medida de mitigación. La alineación conseguida con esta solución no es perfecta y depende tanto de la distancia como de la posición de los objetos en la imagen, habiéndose identificado casos donde los recuadros de detección de los objetos de interés en ambos modos no llegaban a solaparse.

Una alineación espacial incorrecta puede afectar negativamente la efectividad de los métodos de detección con fusión RGB-T. Aunque los resultados razonablemente buenos obtenidos con los métodos desarrollados sugieren cierta robustez frente al desalineamiento, consideramos que este sigue siendo un factor crucial de mejora para futuras investigaciones y desarrollos.

2.3.6. Mitigación de escasez de datos

Sin duda, el principal desafío encontrado durante el desarrollo del proyecto es la escasez de datos, en concreto ejemplos de imágenes de soplos y cuerpos, que son necesarios para el entrenamiento de los algoritmos de inteligencia artificial. Este problema tiene aún más impacto por la gran diversidad de condiciones meteorológicas y de iluminación, que requieren un conjunto de datos para entrenamiento mucho más amplio y diverso.

La escasez inherente de los eventos de interés y las dificultades de adquisición de imágenes mediante cámaras a bordo de ferris rápidos agravan igualmente el problema. Asimismo, el uso de sensores RGB-T complica la utilización de imágenes disponibles de otras campañas o de acceso público, ya que, generalmente, estas solo se encuentran en formato RGB.

A continuación, se presenta un resumen de las medidas técnicas en las que se ha trabajado durante esta fase para mitigar el problema, con mayor o menor grado de éxito.

- Múltiples frames por ejemplo

Una de las decisiones clave al crear el conjunto de datos fue la de extraer múltiples frames o fotogramas por evento para maximizar la cantidad de ejemplos disponibles. Finalmente, se optó por extraer los fotogramas a una tasa de 10 fotogramas por segundo (FPS), en lugar de los 30 fotogramas por segundo de los vídeos originales para conseguir un equilibrio entre la cantidad de ejemplos y la diversidad de los datos. Entrenar con 30 FPS no aporta una mejora considerable por la similitud entre fotogramas consecutivos y el aumento de tiempos de entrenamiento y esfuerzo de etiquetado.

Este enfoque ha permitido obtener suficientes datos para entrenar al modelo de manera efectiva, al mismo tiempo que reduce significativamente la carga de trabajo en las tareas de etiquetado, ya que se procesan manualmente tres veces menos imágenes, optimizando tanto el tiempo como los recursos necesarios.

- **Incluir imágenes de cuerpos**

Uno de los principales desafíos que se identificó fue la escasez de datos para la clase "soplo". Al no contar con el número mínimo de ejemplos recomendados para entrenar adecuadamente un modelo YOLO, y dada la baja variabilidad de estos ejemplos, que provenían de solo unos pocos subclips, el modelo no lograba un entrenamiento eficaz. La solución implementada fue aprovechar los datos de cuerpos de cetáceos, de los cuales disponemos de una mayor cantidad de ejemplos.

Aunque en trabajos previos del estado del arte el evento preferido para detección térmica es el soplo, con dudas sobre la efectividad de detección térmica de cuerpos, en nuestro caso, esta decisión ha dado muy buenos resultados, seguramente debido a la fusión RGBT.

Durante la fase de desarrollo, se hicieron diferentes experimentos en los que se comprobó una mejora considerable de los resultados incluyendo la clase 'cuerpo', pero un empeoramiento si se usaba la clase 'splash', seguramente por la semejanza de estos últimos con otros eventos como crestas de olas.

- **Métodos generativos**

La generación sintética de datos puede ser una herramienta valiosa para superar la escasez de datos, ya que permite crear grandes volúmenes de datos artificiales que simulan las condiciones y características de los escenarios reales. Sin embargo, algo que normalmente pasa desapercibido es que los métodos generativos, como las redes GAN, necesitan una gran cantidad de datos para ser entrenados por lo que no es una solución ideal para nuestro caso, donde los ejemplos son escasos.

Para resolver este problema surgen diferentes enfoques conocidos como *data-efficient* GANs o *few-shot* GANs que permiten generar imágenes sintéticas con menos ejemplos, aunque con ciertas limitaciones. Sin embargo, en el estado del arte para estos métodos el número de referencia es de 100 ejemplos. Esta cantidad mínima está aún lejos de la cantidad de soplos disponibles, aún más teniendo en cuenta la complejidad añadida por la diversidad de condiciones de observación.

Existen otras dificultades para nuestro caso de uso que complican el empleo de estas soluciones: (1) la generación GAN de pares de imágenes RGB-T con consistencia entre los modos es un problema abierto en el estado del arte; (2) la generación GAN para detección de objetos es más compleja ya que requiere condicionar el resultado para incluir objetos de interés (Conditional GAN).

En este contexto, se realizaron experimentos con el generador de datos sintéticos GenCo [17], una arquitectura basada en StyleGAN2-ADA [18] mejorada para generar imágenes con datos limitados. El objetivo fue explorar su capacidad para generar imágenes térmicas y RGB, así como datos condicionados con etiquetas para detección de objetos (por ejemplo, soplos de cetáceos). Se probaron diferentes configuraciones, incluyendo la generación de imágenes RGB de cetáceos y la fusión de imágenes RGB y térmicas concatenadas en una sola. Sin embargo, los resultados no fueron satisfactorios y no pudieron utilizarse para ampliar el conjunto de datos.

- IA generativa comercial

Adobe Firefly, un producto de Adobe Creative Cloud [19], es un modelo de inteligencia artificial generativa que se utiliza en el campo del diseño y actualmente se encuentra en fase de prueba beta pública. Firefly ha sido desarrollado a partir de la plataforma Sensei de Adobe, y su entrenamiento se basa en imágenes provenientes de Creative Commons, Wikimedia, Flickr Commons, así como en 300 millones de imágenes y videos de Adobe Stock y del dominio público. Utiliza estos conjuntos de datos para generar diversos diseños, aprendiendo y ajustando sus resultados a partir de la retroalimentación de los usuarios.

Una función especialmente interesante de Firefly es su capacidad para añadir objetos a una imagen base, similar a un copy-paste generativo. Esto la hace destacar frente a otras alternativas, como Midjourney o DALL-E, ya que podríamos usar una imagen de fondo real. Sin embargo, en nuestras pruebas los resultados generados por Adobe Firefly han sido en general pobres, produciendo imágenes ficticias que no cumplen con los estándares esperados. Además, el número de ejecuciones permitidas es limitado, lo que restringe su uso en nuestro proyecto. Otro inconveniente importante es la falta de opciones para ajustar finamente los diseños, lo que limita su capacidad de personalización y precisión en la generación de contenidos.

- Aumentaciones clásicas

La aumentación de datos es una técnica que incrementa artificialmente el tamaño de un conjunto de datos aplicando transformaciones a los datos existentes. Estas transformaciones crean nuevas versiones de los datos originales manteniendo su relevancia para el problema que se quiere resolver. La implementación original de YOLO7 presenta una serie de aumentaciones básicas. Entre las aumentaciones clásicas que se utilizaron para entrenar los modelos, se encuentran rotaciones, escalados, cambios de brillo y contraste, traslaciones y volteos horizontales (Figura 26).



Figura 26. Imagen con ejemplo de cada transformación.

Otro de los métodos empleados para el entrenamiento de los modelos YOLO es la técnica conocida como mosaico [20], la cual consiste en combinar cuatro imágenes en una sola. Para ello se redimensionan las imágenes, uniéndolas y tomando un recorte aleatorio de las mismas con la que se obtiene la imagen final (Figura 27). Este tipo de aumentación suele dar buenos resultados en problemas de detección de objetos.

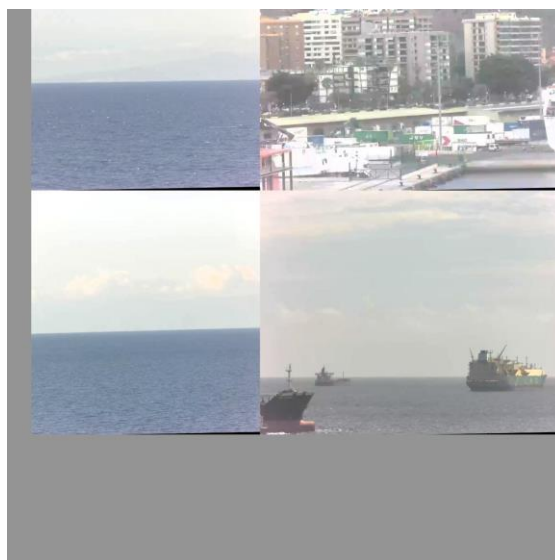


Figura 27. Ejemplo imagen mosaico con traslación y escalado usada en entrenamiento.

- Aumentaciones climáticas

Una de las principales limitaciones que se pueden encontrar respecto a la escasez de datos, es la falta de avistamientos en condiciones desfavorables como pueden ser con lluvia o niebla, esto se puede ver reflejado en la Figura 11 donde se aprecia que la mayoría de las observaciones se realizaron en condiciones de buena visibilidad, esto puede afectar al rendimiento del modelo en condiciones adversas. Para ello se empleó la librería *Albumentations* [21], la cual permite realizar de manera eficiente y flexible diversas transformaciones sobre imágenes (Figura 28), como la simulación de condiciones climáticas (lluvia, niebla, sombras, destellos solares, entre otras), facilitando así la creación de datos aumentados para mejorar la robustez y generalización de los modelos en entornos dinámicos y desafiantes. Aunque como se puede ver en los ejemplos, estas aumentaciones tienen ciertas limitaciones en cuanto al realismo de las mismas.



Figura 28. Imagen con ejemplo de aumentaciones climáticas.

- Aumentación *copy-paste*

La técnica de aumentación de datos *copy-paste* [22] consiste en tomar objetos o regiones de una imagen, copiarlos, y pegarlos en otra imagen para crear nuevas muestras (Figura 29). Aunque puede parecer un método muy rudimentario, por lo general da buenos resultados en problemas de detección de objetos.

En nuestro proyecto, hemos adaptado la implementación de Copy-Paste-Augmenter [23] que ajusta el tamaño los ejemplos pegados utilizando la línea del horizonte como referencia, además de aplicar la técnica poisson blend [24], lo que asegura una transición suave de los píxeles entre la imagen original y la región copiada. Esto permite que las imágenes generadas se integren de manera más natural, creando una apariencia coherente y consistentes con la geometría.



Figura 29. Segmentación de soplo usado en *copy-paste*.

Sin embargo, en nuestra experiencia, la aplicación de esta técnica no ha mejorado de manera significativa la detección de cetáceos, en particular en la identificación de soplos, probablemente debido a la limitada variedad de datos con la que se han trabajado. La técnica

únicamente ha podido ser aplicada sobre varios fotogramas provenientes de solo cuatro subclips de video, lo que reduce la variabilidad que el modelo necesita para aprender de manera efectiva. Esto subraya la importancia de disponer de un conjunto de datos más amplio y diverso para obtener mejores resultados en este tipo de detección.

- Otros esfuerzos: Cachete challenge

El *Cachete Challenge* fue una iniciativa de crowdsourcing interno en la que el personal del ARQUIMEA Research Center participó voluntariamente para identificar y etiquetar eventos de interés, como soplos y cuerpos de cetáceos, en videos grabados a bordo de ferris rápidos. Los participantes registraban sus observaciones mediante un formulario web (Figura 30), y obtenían puntos según el tipo de evento detectado. Los resultados se evaluaban para premiar a los voluntarios con las mejores puntuaciones, y el desafío se extendió hasta junio de 2024.

Sin embargo, la iniciativa solo pudo hacerse de forma interna por evitar la cesión datos y no incumplir el acuerdo alcanzado con Fred.Olsen al inicio del proyecto. Esta limitación, unida a la dificultad de visualizar los eventos en video y su baja frecuencia de aparición, ha hecho que en la práctica el número de eventos detectados de esta forma fuese bajo.

Avistamientos CACHETE Challenge

[Descripción]

1. Nombre del fichero *

Escriba su respuesta

2. Fecha del avistamiento *

Especifique la fecha (d/M/yyyy)

3. Hora del avistamiento (hh:mm:ss) *

Escriba su respuesta

4. Tipo avistamiento *

☐ Soplo

☐ Cuerpo

☐ Otros

Figura 30. Formulario de envío de datos del challenge

2.4. DISEÑO DE LOS MODELOS DE DETECCIÓN

En esta sección se describen las actuaciones referentes al diseño y entrenamiento de los modelos de detección de deep learning en los que está basado el software desarrollado para detección de cetáceos. Este ha sido un exhaustivo proceso de diseño, experimentación y validación, por lo que, por cuestiones prácticas, solo será posible presentar un breve resumen de las actuaciones llevadas a cabo.

La metodología que se ha seguido en líneas generales es la siguiente:

- Diseño preliminar de métodos de fusión (Dataset KAIST)
- Refinamiento de los modelos (Dataset 1.3)
- Validación y selección de métodos de fusión (Dataset 2.0)
- Optimización de los modelos seleccionados (Dataset 3.0)

En informes anteriores, se realizó un estudio del estado del arte de algoritmos de detección basados en deep learning, donde se destacó que la detección en tiempo real en imágenes ha madurado significativamente, mientras que la detección espaciotemporal en video (VOD) sigue siendo un campo experimental y menos consolidado. Debido a los desafíos y la alta incertidumbre asociados con los métodos VOD, se optó por centrarse en la detección de características espaciales en imágenes en tiempo real, recurriendo a detectores "de una etapa". Dentro de esta categoría, se seleccionó YOLOv7 [25] como la mejor opción para nuestro proyecto, ya que ofrece resultados del estado del arte en cuanto a precisión y velocidad de inferencia. YOLOv7 destaca dentro de la familia YOLO por su capacidad para procesar imágenes en tiempo real, con una arquitectura eficiente (backbone, neck y head) que permite detectar objetos con alta precisión y con un 40% menos de parámetros que otros modelos similares, lo que lo convierte en una solución ideal para aplicaciones embarcadas que requieren alta eficiencia y bajos tiempos de procesamiento.

2.4.1. Evaluación cuantitativa de los modelos

La comparación y selección de los modelos de detección requiere realizar una evaluación cuantitativa. Esta evaluación de los modelos entrenados se realiza con datos etiquetados pertenecientes al dataset de test y mediante el uso de métricas estándar en tareas de detección de machine learning.

El dataset de test está formado por imágenes de eventos que no han sido vistos durante el proceso de entrenamiento para garantizar que la evaluación es representativa. La independencia se asegura al nivel de evento y no de frames por la similitud que existe entre frames consecutivos de un mismo evento que podría producir un problema de filtración de datos incluso si diferentes frames de un evento se utilizan para entrenamiento y test.

Se ha realizado una evaluación independiente de los modelos entrenados con el dataset WIDE y TELE, al tratarse de dos conjuntos de datos diferentes y sensores con diferentes características. Las métricas empleadas para realizar la evaluación cuantitativa son: precisión, sensibilidad y la métrica F1. Especialmente se tomará en consideración la métrica F1 por ser un promedio de las dos primeras.

- La precisión mide la proporción de predicciones correctas (TP) entre todas las instancias clasificadas como positivas, es decir, indica qué tan eficaz es el modelo al evitar falsos positivos (FP), o lo que es lo mismo, cuando predice que algo es positivo, qué tan seguro estamos de que realmente lo es.

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

- La sensibilidad refleja la capacidad del modelo de identificar correctamente todas las instancias positivas, enfocándose en reducir los falsos negativos (FN).

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

- La métrica F1 es una medida combinada que evalúa el equilibrio entre precisión y sensibilidad, proporcionando una única métrica cuando se necesita considerar ambas perspectivas.

$$F_1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

Además, para considerar una detección como positiva en tareas de detección de objetos, es necesario tener en cuenta una referencia espacial. Para ello, se define la métrica IoU o Intersection-over-Union, que calcula la superposición entre el área de la predicción y el área del recuadro real (ground truth). Una detección se considera positiva si el IoU supera un umbral predefinido dependiendo de los requisitos de precisión de la aplicación. En nuestro caso se ha utilizado el umbral de IoU que maximiza la métrica F1.

$$\text{IoU} = \frac{\text{Área de Intersección}}{\text{Área de Unión}}$$

En el contexto de este proyecto, la diferenciación entre cuerpos y soplos de cetáceos no es tan relevante como la capacidad del modelo para detectar de manera precisa la presencia de eventos en general. El objetivo principal es identificar de forma efectiva la aparición de cetáceos en las imágenes, independientemente de si lo que se observa es su cuerpo o el soplo que emiten. Por lo tanto, el enfoque se centra en maximizar la tasa de detección de cualquiera de estos indicadores en lugar de realizar una clasificación específica entre cuerpo o soplo, para ello fue necesario implementar una métrica de evaluación multiclase que considerara a ambas como una misma clase “Evento”.

2.4.2. *Diseño preliminar de métodos de fusión.*

Como se introdujo en el primer informe de seguimiento, se planteó un sistema de detección multimodal que utiliza sensores en el espectro visible (RGB) y en el espectro térmico (IR) para mejorar la capacidad de detección mediante la combinación de diferentes bandas del espectro óptico.

Este tipo de soluciones, conocidas en la literatura como fusión multimodal RGB-T, presentan un desafío particular, ya que la mayoría de los métodos de detección de objetos basados en deep learning están diseñados para detección mediante imágenes RGB y el uso de RGB-T sigue siendo limitado. Las redes neuronales disponibles están principalmente diseñadas y optimizadas para procesar imágenes RGB con solo tres canales de entrada, y sus redes troncales o *backbones* suelen estar preentrenadas en conjuntos de datos RGB, lo que presenta un reto adicional para trabajar con imágenes multimodales.

Por estas razones, gran parte del esfuerzo de este proyecto se ha dirigido a investigar y aplicar técnicas de fusión multimodal RGB-T, con el fin de adaptarlas a la red YOLOv7, especializada en detección en tiempo real, para la detección de cetáceos, en particular cachalotes. Es importante señalar que no se han identificado trabajos previos que utilicen la fusión RGB-T para la detección de soplos de cetáceos, por lo que las principales referencias provienen de otros campos, como la conducción autónoma y la teledetección aérea.

Como se comentó en el informe previo (E.2), ante la ausencia inicial de datos RGB-T específicos, la primera fase de diseño se realizó utilizando conjuntos de datos RGB-T públicos diseñados para la detección de peatones en entornos urbanos [26]. Aunque estos datos no corresponden directamente a nuestro caso de uso, han permitido desarrollar una versión preliminar de los algoritmos de fusión. A pesar de que el dominio es distinto, se cree que estos datos pueden replicar condiciones similares a las que se encuentran en la detección de cetáceos, como fondos complejos con múltiples fuentes de falsos positivos, condiciones de iluminación variadas y objetos a diferentes distancias y tamaños.

En la detección mediante redes neuronales, la fusión de datos se puede realizar en a distintos niveles de la red, y cada estrategia requiere una integración específica con el algoritmo de detección. En el estudio preliminar, se implementaron y analizaron tres enfoques principales para la fusión de datos [27] [28], clasificados en:

- Fusión temprana (early-fusion) a nivel de píxeles.
- Fusión intermedia (middle-fusion) a nivel de características, mediante métodos de atención CFT (Cross-modality Fusion Transformer) [29].
- Fusión tardía (late-fusion) a nivel de decisión.

Para evaluar el impacto de estas técnicas de fusión, se definieron dos métodos base sin fusión de datos: detección monomodal RGB, utilizando únicamente imágenes RGB, y detección monomodal térmica, empleando solo imágenes IR.

Los conjuntos de datos empleados para realizar el análisis inicial de rendimiento de los modelos fueron (1) el conjunto de datos LLVIP [30] compuesto por 15.488 pares de imágenes RGB e infrarrojas (IR), caracterizado por capturas en entornos de baja visibilidad y una estricta alineación espacio-temporal entre las imágenes, y el *Multispectral Pedestrian Detection Benchmark Dataset* [26] de KAIST compuesto por 95,000 pares de imágenes con más de 103,000 anotaciones detalladas de peatones, ciclistas y grupos de personas y mayor variedad de condiciones de iluminación. En el caso de KAIST, no todas las imágenes están perfectamente alineadas como se muestra [31] y en la Figura 31 de ejemplo.



Figura 31. Ilustración del problema de desalineación débil de la modalidad.

Los experimentos realizados en esta fase muestran que los modelos de fusión intermedia CFT han obtenido los mejores resultados en términos de la métrica F1, superando al modelo baseline IR. Le sigue el modelo de fusión tardía y en último lugar, el modelo de fusión temprana evidenciando las limitaciones de los modelos de fusión a nivel de pixel, al combinar características de los datos térmicos y visuales (RGB).

2.4.3. Refinamiento de los modelos

En la siguiente fase del proyecto, los experimentos fueron efectuados sobre la versión 1.3 del dataset TELE, la cual dispone de un total de 359 y 2197 etiquetas para las clases “soplo” y “cuerpo” respectivamente. El dataset WIDE no disponía de suficientes eventos (Figura 32) causando que los modelos tuvieran rendimientos poco satisfactorios, por lo que se decidió no realizar más pruebas con esta versión del dataset WIDE.

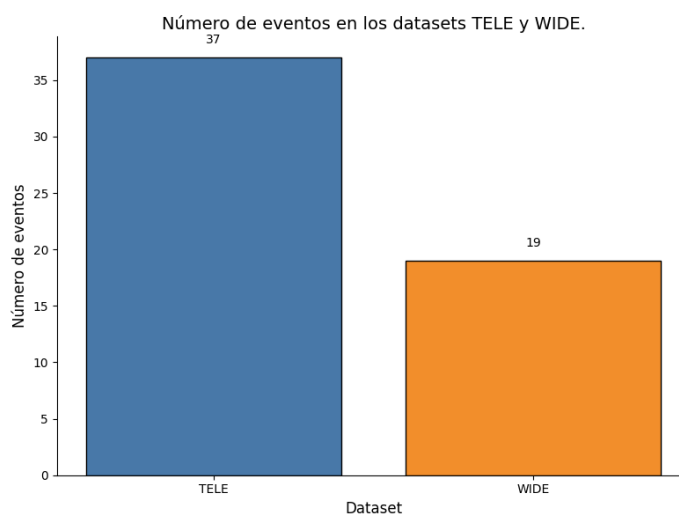


Figura 32. Número de eventos por dataset (v1.3).

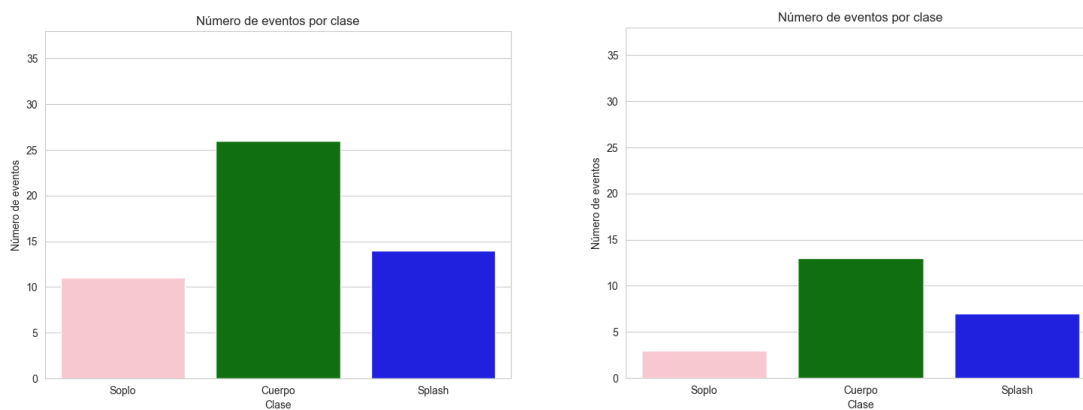


Figura 33. Número de eventos por clase TELE y WIDE (v2.0).

Durante los experimentos realizados, se observa una variabilidad considerable de los resultados, más acusada en los modelos RGB e IR, que depende de los eventos seleccionados para los conjuntos de validación y test. Esto es debido a las limitaciones de la versión del dataset, principalmente por la falta de datos. En cualquier caso, estos experimentos nos han permitido refinar los modelos y obtener una serie de conclusiones importantes.

Se llevaron a cabo dos tipos de experimentos con el fin de analizar la capacidad del modelo para detectar una o ambas clases de interés. En el primer experimento, se entrenó el modelo exclusivamente con la clase "soplo", mientras que en el segundo se utilizaron tanto la clase "soplo" como la clase "cuerpo". El objetivo de esta estrategia fue evaluar el rendimiento del modelo al detectar estas características, especialmente considerando que solo se dispone de un volumen de datos suficiente para la clase "cuerpo". Observamos que la detección utilizando ambas clases, 'cuerpos' y 'soplos', muestra un desempeño significativamente mejor, en lugar de solo utilizar la clase 'soplos' como se planteaba en el objetivo inicial del proyecto. Por este motivo, en adelante se decide aplicar detección con ambas clases.

Los resultados incluyendo la clase 'splash' no fueron satisfactorios. Por este motivo se decidió eliminar la clase "splash" en el futuro y considerarla como fondo en el proceso de entrenamiento debido a la ambigüedad que este elemento puede generar. Al tratarlo como fondo, evitamos que el modelo se confunda al identificar eventos no relevantes, como las perturbaciones en el agua causadas por embarcaciones o crestas de olas, y nos enfocamos en mejorar la precisión en la detección de las clases más representativas y distintivas.

En cuanto a las estrategias de entrenamiento, o la forma de entrenar los modelos, se realizaron diferentes experimentos para encontrar las estrategias óptimas. La primera estrategia consistió en aplicar aprendizaje por transferencia, o *transfer-learning* [32] para los modelos base RGB e IR, utilizando el dataset COCO [2], ampliamente reconocido en el ámbito de la visión por computador. En este enfoque, se optó por congelar las capas iniciales del modelo, que contienen características generales aprendidas previamente, mientras que las capas superiores se entrenaron con una tasa de aprendizaje reducida para adaptar el modelo a nuestro conjunto de datos específico.

Para la segunda estrategia, relativa al entrenamiento de los modelos de fusión intermedia, se optó por aplicar la estrategia de ajuste fino, o *fine-tuning*, que consiste en entrenar toda la red utilizando una tasa de aprendizaje menor. Esta decisión se fundamenta en las diferencias de dominio (*domain gap*) entre RGB y RGBT ya que el dataset COCO es RGB.

La tercera estrategia de entrenamiento también se centra en los modelos de fusión intermedia y se basa en hacer un pre-entrenamiento en el dataset de KAIST, para posteriormente realizar un aprendizaje por transferencia con el dataset final de cetáceos. Dado que KAIST incluye datos en formatos RGB y térmico (RGBT), similares a nuestro caso de uso, esta aproximación permite aprovechar la baja diferencia entre dominios para mejorar los resultados respecto a un pre-entrenamiento común con COCO, que solo contiene imágenes RGB. Sin embargo, en esta fase el modelo entrenado con dicha estrategia obtiene resultados relativamente parecidos al pre-entrenamiento con COCO.

Por otro lado, como se indica anteriormente, la fusión temprana quedó descartada debido a que este método es el más afectado por los problemas de co-registro, ya que requiere una alineación precisa entre los datos RGB y térmicos desde las primeras capas del modelo. Cualquier desajuste en dicha alineación puede degradar significativamente el rendimiento del modelo y provocar errores en la detección, haciendo que sea menos robusto frente a las diferencias entre las modalidades de entrada.

Aunque el modelo IR obtuvo resultados cuantitativos ligeramente mejores, se ha comprobado en ejemplos con videos adicionales, que el modelo IR es muy sensible a otros eventos con alto contraste térmico, produciendo una alta tasa de FP en determinadas situaciones. Esto nos ha llevado a mejorar el dataset en versiones posteriores incluyendo una mayor cantidad de ejemplos de fondo y ejemplos susceptibles de generar falsos positivos, tanto en entrenamiento como en test. El modelo RGB-T middle-fusion demostró ser más robusto frente a falsos positivos en estas circunstancias.

Los resultados con el modelo RGB-T mediante la técnica late-fusion fueron significativamente peores, lo que nos llevó a pensar que había algún error de implementación, que efectivamente fue detectado y corregido en fases siguientes.

En cualquier caso, todos estos resultados son preliminares y han sido interpretados con cautela debido a ciertas limitaciones del conjunto de datos a varios niveles:

- Escasez de datos. No se dispone del número mínimo de etiquetas recomendado en problemas de detección de objetos para la clase soplo, unas 500 aproximadamente.
- Baja diversidad. Las etiquetas para la clase soplo corresponden exclusivamente a unos pocos eventos y no representan diferentes condiciones meteorológicas, de distancia, tamaño, etc.
- Desbalanceo en el dataset. Existe un desequilibrio considerable en la distribución de las clases y en el posicionamiento y tamaño de las etiquetas, entre las diferentes divisiones del dataset, es decir, el conjunto de entrenamiento puede no ser representativo de los conjuntos de test y validación.

2.4.4. Validación de métodos de fusión

Una vez realizado el refinamiento de los métodos y estrategias de entrenamiento, y después de realizar una mejora en el dataset con la versión 2.0, se procedió a realizar una serie de experimentos para validar los métodos de fusión y elegir el método que será utilizado para la evaluación cualitativa final. Los métodos que se compararon son baseline RGB, baseline IR, middle-fusion CFT y late-fusion.

La versión 2.0 del conjunto de datos presenta una distribución balanceada a nivel de clases, con ejemplos de fondo para reducir el número de falsos positivos y para una evaluación más justa. Además, aprovechando el aumento del número de ejemplos, se aumentó el tamaño del subconjunto de test del 10% original al 16% para realizar una evaluación más representativa en test. El número de etiquetas es de 365 y 4160 etiquetas respectivamente para las clases soplo y cuerpo en el dataset TELE y 83 y 2228 respectivamente en la versión wide. Como se aprecia en la Figura 34, aunque ha aumentado el tamaño del dataset WIDE, sigue habiendo una diferencia importante de eventos respecto al TELE.

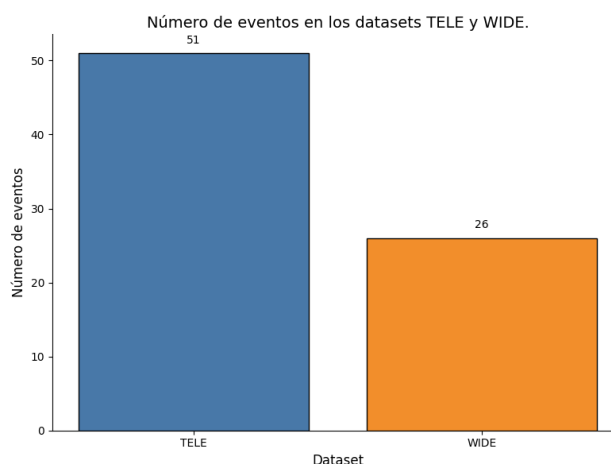


Figura 34. Número de eventos por dataset (v2.0).

Para una comparación lo más justa posible de los diferentes métodos, se definieron las siguientes condiciones del experimento:

- Misma arquitectura del modelo Yolov7 basada en variante P5, con tres cabezas de detección diferentes (P3, P4, P5). Estas cabezas permiten la detección multiescala, mejorando la precisión en distintos tamaños de objetos, donde **P3** (1/8 de la resolución) detecta objetos pequeños, **P4** (1/16) se enfoca en objetos medianos y **P5** (1/32) en objetos grandes. Un mayor número de cabezas mejora la detección multiescala, pero aumenta la complejidad y reduce la velocidad de inferencia, por lo que es conveniente buscar un equilibrio según la aplicación.
- Aumentaciones por defecto de Yolov7 para todos los métodos.
- Ejemplos de fondo en entrenamiento y test.
- Estrategia de entrenamiento mediante transfer-learning
 - Modelos baselines están pre-entrenados con COCO
 - Modelos middle-fusion CFT pre-entrenados con COCO y posteriormente KAIST.

Aunque el dataset WIDE cuenta con una mayor cantidad de datos, al menos para la clase "cuerpos", estos aún no son suficientes para lograr un rendimiento óptimo en general. Por ello, la estrategia adoptada difiere ligeramente, empleando modelos pre-entrenados en el conjunto de datos TELE. A pesar de las diferencias en el campo de visión (FOV), la similitud en el dominio entre ambos datasets ha permitido obtener resultados significativamente mejores, que si se hubiese utilizado directamente COCO o KAIST para el pre-entrenamiento.

La Tabla 4 muestra un resumen de los resultados obtenidos. Las métricas utilizadas en la evaluación son las definidas en la sección 2.4.1 para un valor estándar de solapamiento IUO del 50%. Se observa que el modelo de fusión multimodal 'Middle-fusion CFT' presenta los mejores resultados en los experimentos realizados para todas las métricas. En particular, nos interesan los resultados de *F1 score*, que representa un equilibrio entre precisión y sensibilidad, proporcionando una única métrica cuando se necesita considerar ambas perspectivas como es nuestro caso.

DATASET TELE			
	Precisión	Sensibilidad	F1 score
Baseline RGB	0.678	0.518	0.587
Baseline IR	0.2	0.251	0.227
Middle-fusion CFT	0.824	0.529	0.644
Late-fusion	0.737	0.427	0.541
DATASET WIDE			
	Precisión	Sensibilidad	F1 score
Baseline RGB	0.515	0.451	0.480
Baseline IR	0.00542	0.018	0.008
Middle-fusion CFT	0.602	0.522	0.559
Late-fusion	0.209	0.158	0.18

Tabla 4. Resultados de validación de métodos de fusión

En esta fase se realizaron además experimentos de sensibilidad a la inicialización del entrenamiento. Esto implicó la ejecución de 10 entrenamientos para cada uno de los tres modelos, empleando diferentes semillas aleatorias en una misma división del conjunto de datos. El propósito de este experimento fue estudiar la variabilidad de los resultados al modificar las inicializaciones del modelo y otros generadores aleatorios usados durante el entrenamiento, lo que puede influir en el rendimiento final de los modelos, especialmente en cuanto a la convergencia y estabilidad. Este análisis permitió obtener una visión más amplia sobre la robustez de los modelos frente a variaciones en el proceso de entrenamiento.

Con el objetivo de cuantificar la variabilidad en los resultados se emplea una métrica conocida como coeficiente de variación (CV) [33]. El CV es una medida estadística que expresa la dispersión de los datos en relación con su media. Se calcula dividiendo la desviación estándar por la media y expresando el resultado como un porcentaje. Un valor bajo de CV indica que la variabilidad entre los resultados es pequeña en comparación con la media, mientras que un valor alto señala una mayor dispersión. Esta métrica permite comparar la estabilidad de los modelos ante diferentes configuraciones de los conjuntos de entrenamiento.

$$C_V = \left(\frac{\sigma}{\mu} \right) \times 100$$

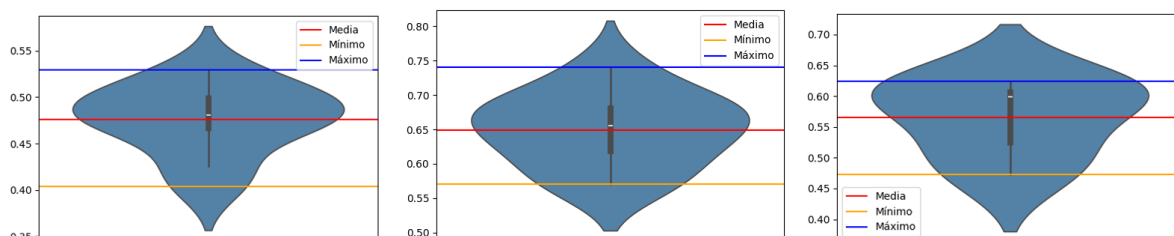


Figura 35. Variabilidad de los resultados obtenidos en los entrenamientos de los modelos RGB (izquierda), IR (medio) y fusión intermedia (derecha) según la inicialización.

Los diagramas de violín en la Figura 35 representan la distribución de los resultados obtenidos en los entrenamientos de cada uno de los modelos (RGB, IR y fusión intermedia), permitiendo visualizar tanto la dispersión de los datos como la concentración de los mismos en ciertas zonas. Estos diagramas muestran la variabilidad de los resultados en función de las diferentes inicializaciones utilizadas durante los entrenamientos. Al calcular el coeficiente de variación para cada uno de los modelos, se obtienen valores de 7.49% para el modelo RGB, 7.86% para el modelo IR y 10.12% para el modelo de fusión intermedia. Un coeficiente de variación cercano al 10% indica que la variabilidad causada por las diferencias en la inicialización de los modelos es relativamente baja, lo que sugiere que este factor no tiene un impacto significativo en el rendimiento final de los modelos.

2.4.5. Optimización final de los modelos

De acuerdo con los resultados previos, se elige el método middle-fusion CFT para continuar con el proceso de ajuste y optimización para obtener los modelos finales usados en el software de detección. Para esta fase, se utilizó una nueva versión del conjunto de datos, con todos los ejemplos disponibles, excluyendo los últimos dos meses de campañas reservados para validación. Además, este conjunto de datos fue revisado y re-etiquetado mediante un proceso basado en detección asistida por el modelo CFT de la fase anterior y confirmación manual de las detecciones. Esto hizo que aumentase el número de eventos y de etiquetas, por lo que los resultados cuantitativos son independientes de los obtenidos en fases anteriores.

En esta fase el objetivo no es llevar a cabo una optimización exhaustiva, sino lograr una mejora general de los modelos hasta alcanzar un rendimiento aceptable. Este proceso es complejo y suele requerir cientos de horas de computación. Las exploraciones se centraron en tres áreas clave:

- Aumentación de datos.
- Ajuste de la tasa de aprendizaje.

Como se mencionó previamente en el apartado 2.3.6, con el objetivo de mitigar la escasez de datos y optimizar el desempeño de los modelos, se implementaron técnicas de aumentación climática mediante la herramienta *Albumentations*, que incorpora variaciones en la iluminación, niebla, lluvia, entre otros factores. Sin embargo, el uso de aumentaciones climáticas no dio buenos resultados, posiblemente por la falta de realismo o por no poder aplicarse de forma consistente en el par RGB-T ya que esta implementación está diseñada para imágenes RGB únicamente. Por ese motivo se utilizaron únicamente aumentaciones clásicas consistentes en: volteo horizontal, rotaciones, escalado, desplazamiento, equalización de histograma, desenfoque, ajustes gamma, cambios de brillo y contraste.

El ajuste de la tasa de aprendizaje se centra en la experimentación con los valores de *learning rate*, un parámetro crítico en el entrenamiento de modelos de aprendizaje profundo. La tasa de aprendizaje controla el tamaño de los pasos que el modelo da al actualizar sus parámetros en cada iteración, lo que tiene un impacto directo en la velocidad y estabilidad del entrenamiento. Por esta razón, se llevaron a cabo varios experimentos con diferentes valores de *learning rate* para encontrar un equilibrio adecuado, permitiendo que el modelo aprenda de manera eficiente sin comprometer la estabilidad. La tasa de entrenamiento que obtuvo mejores resultados fue de 0.01 con un optimizador tipo SGD utilizado por defecto en YoloV7.

Los resultados del mejor modelo obtenido en esta fase, con la versión 3.0 del dataset para TELE y WIDE se presentan en la Tabla 5 a continuación.

DATASET TELE					
IOU 25%			IOU 50%		
Precisión	Sensibilidad	F1	Precisión	Sensibilidad	F1
0.649	0.454	0.534	0.635	0.454	0.529
DATASET WIDE					
IOU 25%			IOU 50%		
Precisión	Sensibilidad	F1	Precisión	Sensibilidad	F1
0.578	0.291	0.387	0.578	0.291	0.387

Tabla 5. Evaluación cuantitativa de los modelos optimizados

Por último, debido a la escasez de datos, no ha sido posible realizar una segmentación de los resultados de la evaluación. Hay que tener en cuenta que el número de subclips o eventos diferentes que contienen los subconjuntos de test es bajo, por lo que este tipo de segmentación no sería significativa.

2.5. DESARROLLO DEL SOFTWARE Y EVALUACIÓN

Una vez completado el diseño, entrenamiento, selección y optimización de los modelos de detección se procedió al desarrollo de un software de evaluación y de demostración de la tecnología. Este software implementa la arquitectura de la solución descrita en la sección 3, integrando los elementos necesarios tanto para la detección como para la visualización y análisis de resultados.

El demostrador proporciona una interfaz gráfica que permite al usuario ejecutar el software de detección a la vez que simula la visualización y operación de forma similar a como se haría en un sistema en producción. Esto posibilita hacer una evaluación cualitativa de software de detección, que proporciona información complementaria a la evaluación cuantitativa de los modelos de detección de objetos basados en Yolov7 (sección 2.4.1).

Como se explica en la sección 3, la solución completa combina diferentes elementos además de los modelos de detección, entre los que destaca el uso de un *tracker* para asociar detecciones del mismo objeto en múltiples frames, lo que ayuda a confirmar detecciones y reducir el número de falsos positivos. Al incorporar estos elementos, la evaluación cualitativa proporciona una información más completa de la solución, en la cual los modelos de detección son solo una parte.

La evaluación cualitativa ha sido realizada de manera independiente por observadores MMO de la ULL. Esta evaluación permite validar la tecnología desde un punto de vista práctico, con mayor relevancia para determinar su utilidad en la mitigación de colisiones producidas por ferris. La complejidad del problema, donde lo importante es la detección del evento o avistamiento (a menudo consistente en varias interacciones) y no tanto de la detección individual de todas estas interacciones u objetos, hace que este tipo de evaluación sea más significativa e interpretable que una evaluación cuantitativa.

Para realizar dicha evaluación, se ha hecho una selección los avistamientos registrados en los últimos 2 meses de campaña. El criterio para esta selección ha sido que la dirección o rumbo del avistamiento estuviese dentro del campo de visión (FOV) de la cámara. Además, se han añadido los fragmentos de video correspondientes a los eventos utilizados en el subconjunto de datos de test durante el entrenamiento de los modelos de detección. De esta forma podemos evaluar el funcionamiento en diferentes condiciones ambientales durante todo el

año, a la vez que aseguramos una evaluación rigurosa con datos que no han sido utilizados para el entrenamiento de los modelos.

Para reproducir las condiciones de funcionamiento en producción, y para poder confirmar si el avistamiento está dentro del FOV de la cámara, se optó por mantener la cámara fija apuntando en la dirección de la proa del barco. Esto refleja una situación más realista en contraste con el modo de operación durante la captura de datos donde se decidió que la cámara se moviese apuntándola para captar avistamientos fuera de su campo de visión (FOV). En un entorno real, la cámara no se ajustaría constantemente para seguir los avistamientos, por lo que esta configuración permite evaluar el rendimiento del sistema en condiciones más representativas.

Los videos tienen una duración aproximada de 10 minutos, lo que permite un margen suficiente de tiempo para posibles imprecisiones en la anotación de los avistamientos, y proporciona imágenes de fondo variadas para evaluar el software en condiciones donde no hay eventos. Por cuestiones prácticas, es posible que dentro de estos 10 minutos algunos videos incluyan varios avistamientos según la anotación del registro de campañas.

La relación de videos utilizados para la evaluación se presenta en la Tabla 6. El nombre del video está formado por la fecha-hora de inicio y fin del video lo que permite su trazabilidad con el registro de campañas. Dichos fragmentos de video, así como los resultados del procesado (subclips con detecciones, logs de resultados, y video completo procesado) se han proporcionado anexo a este informe en formato digital (hoja de cálculo “Evaluación cualitativa”). Esto permitirá examinar los resultados y reproducirlos mediante el software de demostración y los modelos de detección proporcionados.

Video	Estado del mar	Visibilidad	T_Aire	T_Agua
20240916171400_20240916172600	2	2	23	24
20240918100000_20240918101200	1	1	23	21
20240918104228_20240918105200	1	1	23	21
20240918125500_20240918130700	2	2	24	23
20240919125500_20240919130700	0	1	23	21
20240919141100_20240919142300	1	1	26	23
20240919144228_20240919145100	1	1	26	23
20240919164400_20240919165600	0	1	26	23
20240919165700_20240919170900	0	1	26	23
20240923171400_20240923172600	0	1	26	23
20241007100000_20241007101200	1	1	24.7	23.9
20241007102900_20241007104100	1	1	24.7	23.9
20241007125400_20241007130600	2	1	25.6	24.8
20241007165000_20241007170200	2	1	26.6	24.5
20241008103000_20241008104200	2	1	24.2	24
20241008125000_20241008130200	3	1	27	23
20241008140100_20241008141300	4	1	24.5	24.5
20241008143400_20241008144600	4	1	24.5	24.5
20241008164228_20241008165000	4	1	26.1	25
20241008165000_20241008170200	4	1	26.1	25
20241015102200_20241015103400	3	1	25	21

20241031081100_20241031082300	0	1	24	23
20241031081800_20241031083000	0	1	24	23
20241031141800_20241031143000	0	1	23.5	21

Tabla 6. Videos usados para evaluación cualitativa (nuevos)

Estado del mar: (0, plato), (1, algunas ondas), (2, pequeñas olitas), (3, algunos borregos), (4, borregos), (5 marejada). Visibilidad (del horizonte): (1, buena), (2, media), (3, mala)

Adicionalmente, se han incluido en la evaluación cualitativa fragmentos de video correspondientes a los subclips del conjunto de datos de test (Tabla 7). Estos datos se utilizaron para la evaluación cuantitativa de los algoritmos de detección, pero no para su entrenamiento, por lo cual su inclusión en la evaluación cualitativa no está comprometida. Como los datasets TELE y WIDE se crearon de forma independiente, es posible que algunos eventos visibles en los dos FOV se hayan utilizado para entrenamiento solo en uno de ellos. En ese caso, el FOV usado para entrenamiento ha sido excluido de la evaluación.

Video	Estado del mar	Visibilidad	T_Aire	T_Agua
20231019082900_20231019083230	0	1	-	26.4
20231128162233_20231128162833	2	1	24	23
20231213120800_20231213122000	4	3	25	23
20231214175300_20231214175332	4	1	-	22.4
20231225141600_20231225141800	4	1	-	22.4
20240307082200_20240307083400	4	1	22	21
20240308121700_20240308121718	4	1	22	21
20240703163000_20240703164200	3	1	23	21

Tabla 7. Videos usados para evaluación cualitativa (conjunto de test)

Estado del mar: (0, plato), (1, algunas ondas), (2, pequeñas olitas), (3, algunos borregos), (4, borregos), (5 marejada). Visibilidad (del horizonte): (1, buena), (2, media), (3, mala)

Los resultados de la evaluación cualitativa y las conclusiones se presentan en la sección 4. Debe tenerse en cuenta que, como es habitual en los sistemas de detección automática, es posible ajustar diferentes parámetros que establecen un equilibrio entre precisión y sensibilidad. El software de evaluación permite realizar estos ajustes mediante la configuración avanzada, descrita en el manual de usuario.

Inicialmente, se utilizó la misma configuración para todos los fragmentos de video. Sin embargo, tras identificar algunos casos límite, estos se agruparon y procesaron con una configuración optimizada por grupos (Tabla 8, Tabla 9). Estos casos particulares se proporcionaron a la ULL para su evaluación y comparación, indicando claramente la configuración utilizada.

Subclip	Configuración
20231213120800_20231213122000	Conf2
20231214175300_20231214175332	Conf2
20240308121700_20240308121718	Conf2
20240918100000_20240918101200	Conf1
20240918104228_20240918105200	Conf1
20240918125500_20240918130700	Conf1

20241008103000_20241008104200	Conf1
20241008140100_20241008141300	Conf1
20241008143400_20241008144600	Conf3
20241015102200_20241015103400	Conf1

Tabla 8. Videos procesados con configuración optimizada

Configuración	Descripción
Conf0	Configuración estándar con un equilibrio entre sensibilidad y precisión.
Conf1	Configuración optimizada para reducir el número de falsos positivos.
Conf2	Configuración optimizada para mejorar la detección de eventos difíciles
Conf3	Configuración optimizada para la detección de soplos

Tabla 9. Descripción de las configuraciones

3. DESCRIPCIÓN DE LA SOLUCIÓN

En esta sección se presentará una descripción detallada de la solución desarrollada y evaluada para la detección de cetáceos, destacando los distintos componentes y los detalles más relevantes. La solución se ha desarrollado para el hardware específico de captura de imagen que fue evaluado y seleccionado en fases anteriores del proyecto (ver informe E.1). En concreto, el sistema óptico Sentry del fabricante SEA.AI, con las siguientes características:

- Modelo para barcos a motor.
- Pan 360°, Tilt +/-20°
- Banda IR: LWIR 8 µm – 14 µm
- 2x sensores IR FLIR Boson. No refrigerados.
 - o 640 x 512 px, FOV 32°x25.8°
 - o 640 x 512 px, FOV 8°x6.4°
- 2x sensores low light RGB (HD).
 - o 2560 x 1944 px , FOV 34°x25.8°
 - o 2560 x 1944 px , FOV 8°x6.4°
- Frecuencia actualización: 30 FPS

3.1. ARQUITECTURA DE LA SOLUCIÓN

La Figura 36 contiene un diagrama de bloques a alto nivel de la arquitectura de la solución software con los componentes más relevantes. La arquitectura está compuesta por dos ramas de detección independientes, cada una para un Field-Of-View (FOV), o campo de visión, coincidiendo con las características del sistema óptico. FOV_WIDE se refiere a la óptica gran angular (FOV nominal de 30°), mientras que FOV_TELE se refiere a la óptica con zoom (FOV nominal de 8°). Los resultados de ambas ramas se combinarán posteriormente en el componente de fusión multifocal o MULTIFOCAL FUSION.

Como se discutió en la selección del sistema óptico del informe E.1, el uso de dos FOV complementarios va a permitir mejorar las capacidades de detección. El FOV WIDE

proporciona un amplio campo de visión para la detección de objetos cercanos, mientras que el FOV TELE habilita la detección a larga distancia con el compromiso de un campo de visión más reducido, pero con mayor margen de maniobra.

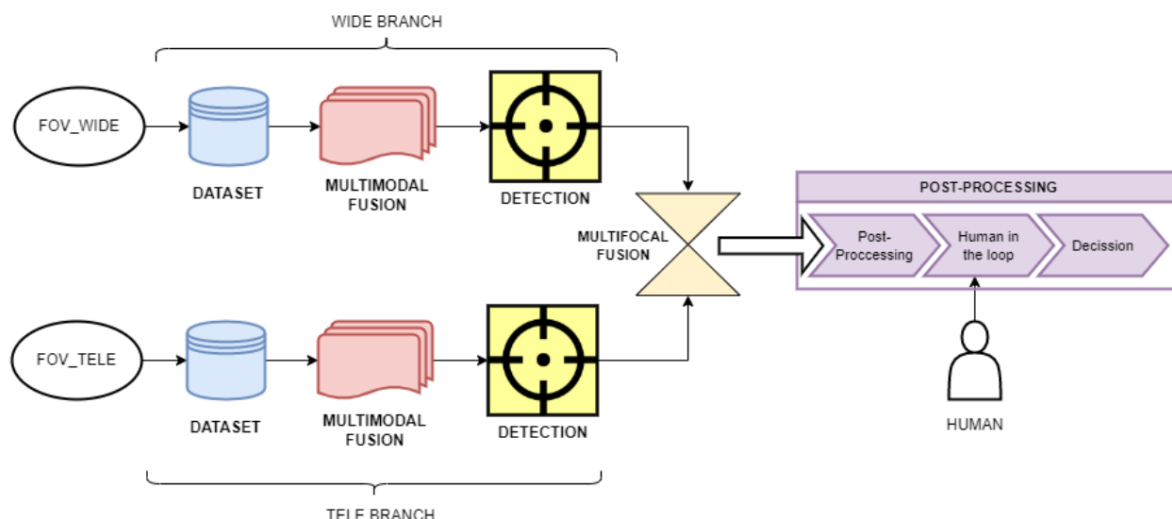


Figura 36. Arquitectura de la solución software

Para la detección se emplean modelos de detección de deep learning mediante las técnicas de fusión multimodal descritas ampliamente en este informe y en informes previos. Por fusión multimodal se entiende la combinación de diferentes modos de imagen, en particular modos visual (RGB) y térmico (IR). En cuanto a los modelos de detección, estos están basados en la arquitectura de redes neuronales YOLOv7 [34] para detección de objetos en tiempo real. Adicionalmente, el componente de detección realizará una confirmación de detecciones automática mediante la asociación de detecciones correspondientes a un mismo objeto en frames contiguos utilizando un algoritmo de tracking. Esto permite reducir el número de falsos positivos mejorando la precisión.

La detección en cada rama se realiza de forma independiente de la otra, y ambas son similares salvo por el elemento DATASET que será específico de cada rama. El término DATASET debe entenderse de forma amplia, y se refiere tanto a los conjuntos de datos utilizados durante el proceso de entrenamiento de los modelos, como los videos utilizados para la evaluación cualitativa, o la entrada de video en tiempo real si se tratase de un sistema en producción. En definitiva, se refiere a los datos de entrada del modelo de detección.

Para la fusión multifocal, es decir, de las diferentes ramas de FOV, se ha preferido aplicar un enfoque simple pero efectivo que consiste en la unión de las detecciones individuales, de forma que ambas ramas puedan generar alertas de eventos independientemente, lo que contribuirá a la mejora de la sensibilidad del sistema.

La parte final del sistema es el módulo de post procesamiento POST-PROCESING HITL, que incluye la toma de decisión con intervención humana. El módulo de post procesado se encarga de diferentes funciones que permitirán mostrar los resultados e información relevante a un usuario como puede ser un operador en un sistema en producción.

Entendemos que el usuario es parte fundamental del sistema y, por tanto, la solución se ha enfocado como una herramienta que debe facilitar la toma de decisiones de forma rápida y efectiva a un operador humano, y no como un sistema de decisión automático.

La inclusión del humano en el bucle de toma de decisiones se conoce como Human-in-the-loop (HITL). Es un enfoque en el desarrollo y uso de aplicaciones de inteligencia artificial (IA) donde la intervención y supervisión humana juegan un papel central en el ciclo de evaluación, refinamiento y toma de decisiones del sistema. En lugar de confiar completamente en la automatización, los sistemas HITL integran la participación humana en puntos clave para asegurar que la IA sea más precisa, segura y éticamente responsable. Este concepto reconoce que, aunque los algoritmos de IA pueden procesar y analizar grandes cantidades de datos rápidamente, la intervención humana es crucial para manejar la ambigüedad, corregir errores y aportar juicio contextual. Este enfoque es especialmente relevante en aplicaciones críticas, donde las decisiones tienen un alto impacto, y en entornos dinámicos con alta incertidumbre, como es el caso de uso que nos ocupa.

3.2. DEMOSTRADOR TECNOLÓGICO

El software entregado consiste en un demostrador tecnológico para la evaluación de la tecnología y las medidas de detección desarrolladas. La finalidad del demostrador es estrictamente la de evaluar la tecnología y proporcionar una herramienta que permita realizar un testeo más exhaustivo con nuevas grabaciones por parte de la entidad licitadora si así lo considera oportuno, así como para la reproducción de los resultados presentados en el presente informe.

No debe entenderse, por tanto, dicho demostrador como un software para producción ni para uso comercial. La aplicación de la tecnología analizada requerirá una serie de medidas y mejoras que se describen en la sección 5.

El demostrador software implementa la arquitectura descrita en la sección previa, constando de los siguientes componentes, que serán detallados en los apartados siguientes:

- Modelos de detección
- Librería de tracking
- Módulo de post-procesado
- Interfaz gráfica

3.2.1. Modelos de detección

El demostrador se entrega con los siguientes modelos pre-entrenados basados en la arquitectura YOLOv7, tanto para la rama FOV_WIDE como FOV_TELE:

- Baseline RGB. Detección mediante imágenes RGB a color, sin fusión de datos RGB-T.
- Baseline IR. Detección mediante imágenes IR térmicas, sin fusión de datos RGB-T.
- RGB-T CFT middle-fusion. Detección mediante fusión multimodal RGB-T tipo middle-fusion a la salida del backbone de la red, con mecanismos de atención CFT (Cross-Modality Fusion Transformer) inspirado en [29].

- RGB-T late-fusion. Detección mediante fusión multimodal RGB-T tipo late-fusion a la salida del detector mediante un módulo PMS (Prediction Matching System) inspirado en [35].

El demostrador permite elegir el modelo de detección con el que realizar el procesamiento de los videos de evaluación, lo que permite comparar el rendimiento de diferentes modelos y técnicas de fusión de datos.

Para facilitar la integración en el demostrador, los modelos se proporcionan en el formato universal y multi-plataforma ONNX [36] y se ejecutan mediante el intérprete de ejecución de redes neuronales ONNX Runtime [37].

El modelo de detección se ejecuta de forma offline con cada frame de los videos de entrada, que han sido previamente grabados por el sistema de captura Sentry instalado en los ferris. El resultado para cada frame es una lista de bounding-boxes o recuadros para cada objeto detectado. Dicha lista incluye la siguiente información:

- Clase: clase a la que pertenece el objeto detectado (soplo, cuerpo, splash).
- Recuadro: posición y dimensiones en coordenadas normalizadas XYWH, donde (X,Y) se refiere a la posición centro del recuadro y (W,H) al ancho y alto de la imagen.
- Confianza: puntuación de confianza asignado a la detección. Valor entre (0,1) con 0 la menor confianza y 1 la mayor confianza.

3.2.2. Librería de tracking

Como se ha comentado, resulta conveniente asociar detecciones correspondientes a un mismo objeto en frames consecutivas. Esto permite confirmar las detecciones, reduciendo el número de falsos positivos, y reducir el número de alertas al generar una alerta por asociación y no por cada detección individual.

Este tipo de asociación es conocida como tracking [38], y es un problema de visión por computador tan amplio como lo es la detección de objetos propiamente dicha. Entre los diferentes métodos y enfoques disponibles se ha optado por una solución basada en el método DeepSORT [39]. DeepSORT es un método de tracking ligero para aplicaciones en tiempo real que combina deep learning y técnicas clásicas, y que realiza una asociación de detecciones mediante métricas de distancia basadas en:

- Movimiento. Distancia con predicción de posición basada en filtros Kalman [40], que tienen en cuenta el movimiento del objeto en frames pasados y predicen su posición futura.
- Apariencia. Distancia entre representaciones vectoriales de apariencia (*embeddings*) extraídas mediante una red neuronal.

DeepSORT es un método empleado ampliamente, que permite un alto grado de configuración para adaptarse a distintos tipos de aplicaciones. El método ha sido adaptado para detección multimodal RGB-T, pudiendo seleccionar el modo para tracking: RGB o IR. El demostrador utiliza una configuración por defecto que ha sido probada y optimizada durante el desarrollo de la solución, aunque puede proporcionarse una configuración diferente si se desea.

Otra de las mejoras introducidas en la implementación original es la confirmación de detecciones mediante un esquema "N-out-of-M". Según este método, para confirmar una asociación o track es necesario N asignaciones no consecutivas de detecciones en M frames

consecutivos tras la inicialización del track (con $N \leq M$). Solo los tracks confirmados serán considerados como una detección y generarán una alarma.

El esquema “N-out-of-M” permite eliminar detecciones espurias por lo general correspondientes a falsos positivos, mientras que proporciona cierta flexibilidad al permitir M-N pérdidas de track iniciales, que pueden ser producidas por oclusiones o detecciones fallidas al inicio del track.

El resultado de aplicar el algoritmo de tracking a la salida del detector es una lista de tracks o asociaciones, cada una correspondiente a un objeto de interés confirmado y que generará una alarma. Para cada track se indican los frames del video donde se ha asociado con una detección y el recuadro correspondiente a dicha detección.

3.2.3. *Módulo de post-procesado HITL*

El módulo de post procesado se encarga de diferentes funciones tras la confirmación de las detecciones realizadas por el tracker. Estas funciones permitirán mostrar los resultados e información relevante a un usuario de forma similar a la que se presentaría a un operador en un sistema en producción.

Para el demostrador, las funciones implementadas están principalmente enfocadas a facilitar la evaluación de la tecnología de detección. Como se discutirá en la sección 5, para la aplicación con éxito de la tecnología será necesario desarrollar otras funciones de HITL para conseguir un ciclo continuo de colaboración entre humanos y algoritmos, en el que los humanos no solo supervisan, sino que también contribuyen a la mejora de los sistemas durante todo el ciclo de vida del producto. Por ejemplo, la confirmación de detecciones puede proporcionar realimentación y ejemplos útiles con lo que reentrenar y mejorar los algoritmos de detección de forma continua, lo que se considera clave para el caso que nos ocupa.

En concreto, el módulo de post procesado del demostrador realiza las siguientes funciones:

- **Filtrado de detecciones no admisibles.**

En particular, se filtran las detecciones por encima de la línea del horizonte, que en ningún caso pueden corresponder a un evento de interés.

La detección de la línea del horizonte se realiza mediante algoritmos clásicos de visión por computador mediante la librería OpenCV [41], con un método simple pero ligero pensando en su aplicación en tiempo real, basado en la aplicación del método Otsu [42] de umbralización para separar el cielo del resto de la imagen.

- **Generación de alertas.**

Una vez se ha confirmado la detección con múltiples frames y se han filtrado las detecciones no admisibles, se genera una alerta. Esto permite reducir el número de alertas, evitando una tasa de alertas excesiva que podría comprometer la usabilidad.

Además de su visualización en la interfaz gráfica, las alertas serán registradas por el software de demostración y serán incluidas en un informe final de resultados generado por el demostrador.

- **Cálculo de rumbo y distancia de observación.**

Esta información es relevante para la confirmación de la observación y evaluación del riesgo de colisión.

Para el cálculo del rumbo se asume que la cámara está apuntada en la dirección de la proa del barco, es decir, los valores estimados son relativos al centro óptico de la cámara. Conocidos el valor de FOV y resolución del sensor de referencia (en este caso RGB), el cálculo consiste en calcular el FOV por píxel, o iFOV, en la dirección horizontal, para después estimar el ángulo de la observación contabilizando el número de píxeles en horizontal desde el centro de la imagen hasta el centro del recuadro del objeto detectado. Para aplicar este método previamente ha sido necesario estimar la línea del horizonte y corregir la posible rotación de la imagen.

Para la estimación de la distancia se utiliza el método de triangulación esférica [43] utilizado por Zitterbart [3]. Es un cálculo geométrico teniendo en cuenta la curvatura media terrestre que se basa en la medida del ángulo respecto a la línea del horizonte, para la cual se empleará el método de detección del horizonte comentado previamente.

Es conocido que este método presenta limitaciones cuando la línea del horizonte no es visible en la imagen, bien por la orientación de la cámara, por el movimiento del barco, por ejemplo, o bien cuando lo que se observa en la imagen es realmente la línea de costa, cuando la costa es más cercana que la distancia del horizonte. Para el primer problema, es necesario contar con mecanismos efectivos de estabilización de la cámara y un encuadre con margen suficiente para compensar pequeñas oscilaciones. Mientras que para resolver el segundo problema es necesario integrar cartografía y posición GPS en el software de detección, de forma que se pueda estimar la distancia a la línea de costa y aplicar el método alternativo descrito en [REF] para estos casos.

- **Estimación de riesgo de colisión.**

Se realiza una estimación del riesgo de colisión basada en la distancia y rumbo con respecto a la embarcación (Figura 37) que define diferentes sectores con 4 grados de riesgo: muy alto, alto, medio y bajo.

Para definir los límites de los sectores se han tenido en cuenta las características de campo de visión de los sensores, la distancia máxima de detección, las características de los ferris y la distancia de detección mínima deseable según conversaciones con Fred.Olsen para tener margen suficiente para confirmar, notificar y actuar frente a una detección (1km).

Los umbrales de ángulo también tienen en cuenta un margen mínimo de 50m a cada lado del ferry, superior al ancho del buque (25-30m), en el caso peor de cada sector, es decir al inicio del sector. De forma que, si el ferry avanza en línea recta y el objeto detectado no se desplaza, al quedar fuera del umbral de ángulo hay margen suficiente con el ferry para evitar la colisión y reduzca el riesgo.

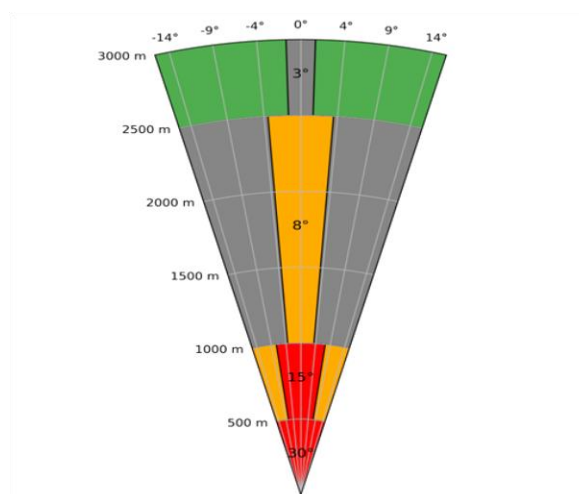


Figura 37. Áreas de riesgo de colisión

- Presentación de resultados.

En el demostrador desarrollado, este componente está más enfocado a la visualización de resultados para la evaluación cualitativa de la solución. La presentación de resultados se realiza a diferentes niveles y está muy asociada al funcionamiento de la interfaz gráfica (GUI), que se tratará de forma específica en la sección 3.2.4.

1. Resultados en la interfaz gráfica GUI. Se muestran las detecciones confirmadas con el color correspondiente al nivel de riesgo estimado (Figura 39). La forma de presentar estos resultados es similar a como sería en un sistema en producción.
2. Subclips para re-visualización de las detecciones. Con cada detección se recorta un pequeño subclip de video en el FOV donde fue identificada, para su observación y análisis en retrospectiva. Al finalizar el procesado, todos los subclips estarán disponibles en la carpeta de resultados.
3. Archivo de registro de detecciones. Al completarse el procesado de los videos, se genera un archivo en formato JSON (texto plano) con un listado de las detecciones confirmadas, donde se incluye la siguiente información (Figura 38):
 - Identificador único por detección
 - Estimación de rumbo y distancia
 - Listado detecciones asociadas: frames, recuadros (bbox) y clase del objeto. La información de los recuadros se proporciona mediante coordenadas normalizadas XYWH, donde (X,Y) se refiere a la posición centro del recuadro y (W,H) al ancho y alto de la imagen.

```

1  events_data = {
2      "Events": [
3          {
4              "id": 0,
5              "distance": 100.0,
6              "angle": 45.0,
7              "bboxes": [
8                  {
9                      "frame_number": 1,
10                     "bbox": {
11                         "class": 0,
12                         "x_center": 0.5,
13                         "y_center": 0.5,
14                         "width": 0.2,
15                         "height": 0.3
16                     }
17                 },
18                 {
19                     "frame_number": 2,
20                     "bbox": {
21                         "class": 0,
22                         "x_center": 0.6,
23                         "y_center": 0.4,
24                         "width": 0.25,
25                         "height": 0.35

```

Figura 38. Ejemplo de listado de detecciones

3.2.4. Interfaz gráfica de usuario (Graphic user interface, GUI)

La interfaz gráfica (GUI) ofrece a los usuarios una plataforma para ejecutar el demostrador tecnológico, simulando la visualización y operación de manera similar a como se realizaría en un sistema en producción. Junto con el software, se entrega un manual de instalación y uso que incluye una descripción detallada de la interfaz y su funcionamiento.

La interfaz permite cargar hasta cuatro videos previamente grabados, correspondientes a los sensores del sistema óptico Sentry, los cuales se utilizarán para demostrar la tecnología de detección. Una vez cargados los videos y seleccionada la configuración de la prueba, se inicia el procesamiento offline mediante el sistema de detección, un procedimiento que puede tardar varios minutos.

Este procesamiento genera todos los resultados asociados a la prueba, incluyendo videos completos con los recuadros de detección superpuestos, subclips con detecciones y un archivo JSON que detalla el listado de todas las detecciones.

Posteriormente, el usuario puede simular la visualización de los videos en tiempo real, con la visualización de las detecciones en la GUI. Por defecto, estas detecciones irán apareciendo en la GUI a medida que ese vaya reproduciendo el video, pero al haberse procesado realmente offline previamente, es posible mostrar todas las detecciones desde el principio y avanzar o retroceder en el video usando el control del 'timeline'.

La pantalla principal de la GUI (Figura 39) consta de los siguientes elementos:

1. Área de visualización de los videos (Figura 39-1). Las detecciones se indican mediante un identificador único y un recuadro con el color según el riesgo estimado.

- Listado de detecciones 'Event List' (Figura 39-2). Incluye datos relevantes de la cada detección y haciendo clic se abre la ventana con los subclips correspondientes.
- Barra temporal o timeline (Figura 39-3). Permite desplazarse por la grabación e incluye marcas en las posiciones donde hay detecciones.

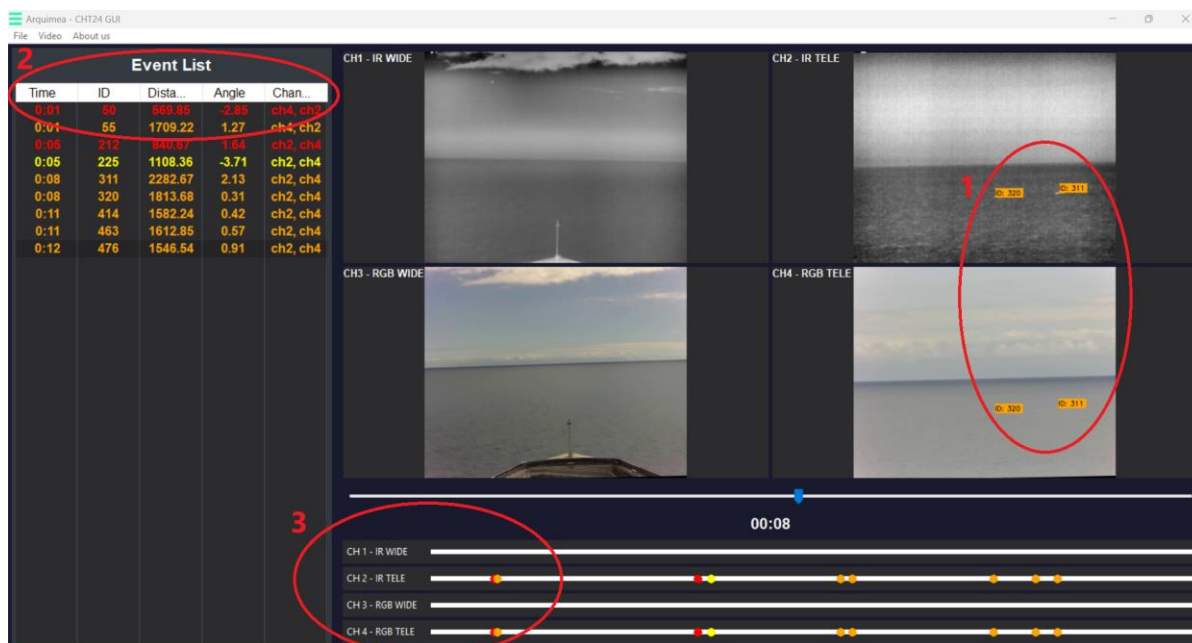


Figura 39. Pantalla principal de la GUI

Durante la demostración, es posible acceder a cada subclip directamente desde el listado de detecciones. Esto abrirá una ventana emergente (Figura 38) que muestra información adicional, como el instante en el que se produjo la detección, la estimación de la distancia y el rumbo, además de la ruta donde se almacena el subclip.

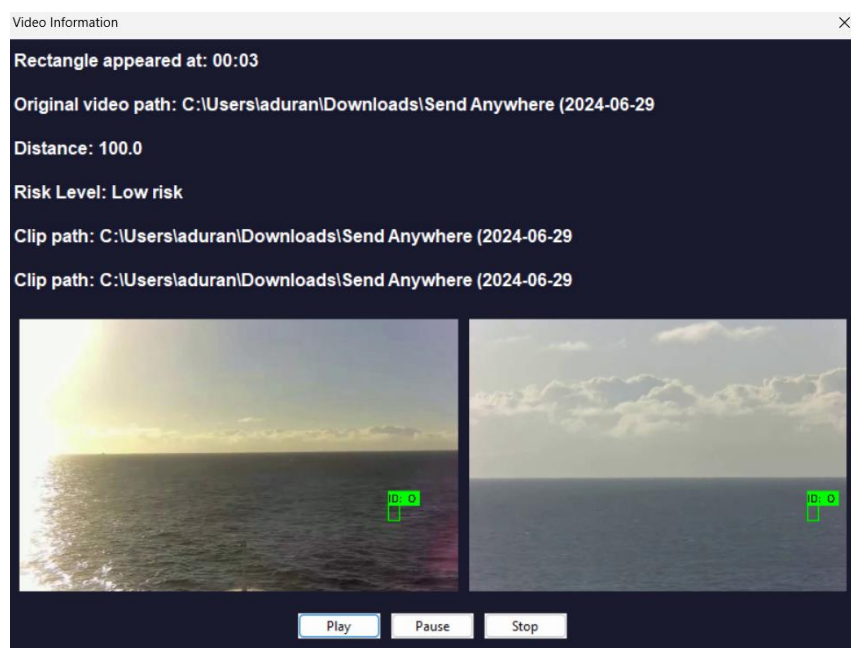


Figura 40. Ventana emergente con información de la detección

4. RESULTADOS DE LA EVALUACIÓN CUALITATIVA INDEPENDIENTE Y CONCLUSIONES

A continuación, se presentan los resultados de la evaluación cualitativa realizada de forma independiente por la ULL. Posteriormente, se expondrán las conclusiones basadas en estos resultados y en otros factores observados a lo largo del proyecto.

Como se ha mencionado previamente, tanto los resultados de la evaluación cualitativa como los videos originales y procesados utilizados para la evaluación, serán proporcionados como entregables para permitir una verificación detallada de los mismos (ver ANEXO 1).

4.1. RESULTADOS

La evaluación cualitativa completa se incluye como entregable en formato Excel (ver ANEXO 1, entregable E3.7, correspondiente a la hoja de cálculo “Evaluación_cualitativa”). Este archivo incluye un análisis detallado de cada uno de los videos seleccionados según se describe en la sección 3.2.5. Para cada video se han definido una serie de eventos a criterio de los revisores, independientes del registro de campañas, como si estos fuesen avistamientos de una campaña ‘virtual’ realizada con la visualización de los videos y asistencia del software de detección. Para cada uno de estos eventos se indica el inicio y fin aproximado, y si se trata de un verdadero positivo (TP), un falso positivo (FP) o un falso negativo (FN). Además, se especifica en qué sensores del sistema óptico ha sido detectado cada evento por el software o confirmado visualmente (en caso de no haber sido detectado). Finalmente se incluye información adicional, como la calidad de la imagen y comentarios detallados.

Con esta información, se han planteado dos formas de sintetizar los resultados con métricas que permitan una visión general.

1. CRITERIO 1:

- Métricas de TP, FP y FN directamente de la evaluación cualitativa.
- Debe considerarse una aproximación, ya que algunos eventos agrupan varias interacciones o detecciones diferentes, como puede ocurrir con FP seguidos.
- No evalúa directamente la capacidad de detectar los avistamientos definidos en el registro de campañas.
- Los resultados se resumen en la Tabla 10.

	TP	FP	FN	TOTAL
TOTAL	30	109	11	150
TASA	0.2	0.73	0,07	1

Tabla 10. Métricas de evaluación CRITERIO 1

2. CRITERIO 2:

- Evalúa la tasa de acierto en la detección de los avistamientos registrados en las campañas de avistamiento correspondientes a los videos de evaluación.

- b. Se han tenido en cuenta únicamente los fragmentos de videos correspondientes a campañas de la ULL (excluyendo una pequeña cantidad de videos de test provenientes de observaciones de Fred.Olsen)
- c. Solo considera los avistamientos con ángulo de observación anotado dentro del FOV de la cámara. Puntualmente incluye varios ejemplos que a pesar de estar anotados fuera del FOV se han podido confirmar visualmente.
- d. Para asociar los avistamientos con los eventos anotados en la evaluación cualitativa con el fin de computar las métricas, se ha definido un tiempo +/- 3 minutos. Un minuto por ser la resolución de la anotación, con un margen de 2 minutos para tener en cuenta posibles retrasos en la anotación de los avistamientos o errores en la sincronización de los instrumentos de medida.
- e. Se definen 3 métricas para cada avistamiento:
 - i. ACIERTO_SYS (acierto del sistema). Si el evento ha sido detectado teniendo en cuenta el sistema completo de cámaras y software de detección. Corresponde a un verdadero positivo (TP).
 - ii. ACIERTO_SW (acierto del software). Se considera un acierto cuando el evento ha sido detectado correctamente, sin ser penalizado cuando no es posible su confirmación visual en video. Corresponde a un verdadero positivo (TP) o un verdadero negativo al no haberse podido confirmar visualmente (TN).
 - iii. CONFIRMACIÓN. Indica si el avistamiento ha sido confirmado en video por los MMO durante la evaluación.
- f. Se considera un TN cuando el avistamiento no ha sido confirmado visualmente y cuando no existen falsos positivos próximos al evento.
- g. La contabilización de TP y FP se ha hecho independientemente. Al resumir por avistamiento la actividad de múltiples interacciones, se ha permitido contabilizar FP y TP en el mismo avistamiento si la detección se realiza correctamente, pero aparecen falsos positivos. Bajo nuestro criterio, esta situación permite la detección del avistamiento y será considerada un acierto.
- h. En caso de que para un mismo avistamiento coincida un evento TP y FN de la evaluación, se da prioridad al TP clasificándolo como un acierto al considerar que el avistamiento ha sido detectado.
- i. Los resultados detallados de este criterio de evaluación se adjuntan junto con el entregable de evaluación cualitativa (E3.7). Para cada video de evaluación se indican los avistamientos asociados y sus métricas.
- j. Los resultados se resumen en la Tabla 11.

	ACIERTO SYSTEM	ACIERTO SOFTWARE	CONFIRMACION VISUAL
TOTAL	17	24	21
TASA	0,4857143	0,685714	0,6

Tabla 11. Métricas de evaluación CRITERIO 2

4.2. CONCLUSIONES

En vista de los resultados presentados podemos concluir que los avistamientos anotados en las campañas han sido detectados por el software con una tasa de acierto del 68,6%. Si excluimos los videos de test, esta tasa aumenta al 71.4%. Recordemos que los avistamientos de test no fueron capturados en las mismas condiciones que los de evaluación.

La tasa de acierto del software tiene en cuenta los avistamientos no detectados que además no han podido ser confirmados en video por los MMO, considerándolos verdaderos negativos (TN). Esto excluye de la evaluación posibles errores en la anotación del tiempo o la dirección del avistamiento, así como errores de sincronización de las referencias de tiempo de los equipos involucrados en la toma de datos. También se excluyen las limitaciones del sistema óptico y su calidad de imagen, en tanto pueden dificultar la confirmación visual en video. Si queremos tener todos estos factores en cuenta y evaluar el sistema completo, habría que considerar la métrica de tasa de detecciones, algo menor (48.5%).

La tasa de confirmaciones visuales nos da precisamente la información del porcentaje de avistamientos que se han podido confirmar en video, un 60%. Considerando la información de la calidad de imagen incluida en la evaluación cualitativa (Figura 41) y los problemas constatados en este informe al respecto, cabe pensar que este puede ser un factor limitante tanto en la detección como en la confirmación visual. Por otro lado, se ha constatado que un número considerable de avistamientos a corta distancia no han podido ser confirmados, lo que hace pensar que puede haber una zona ciega próxima al barco debida a la geometría de la captura: FOV, altura y ángulo de elevación.

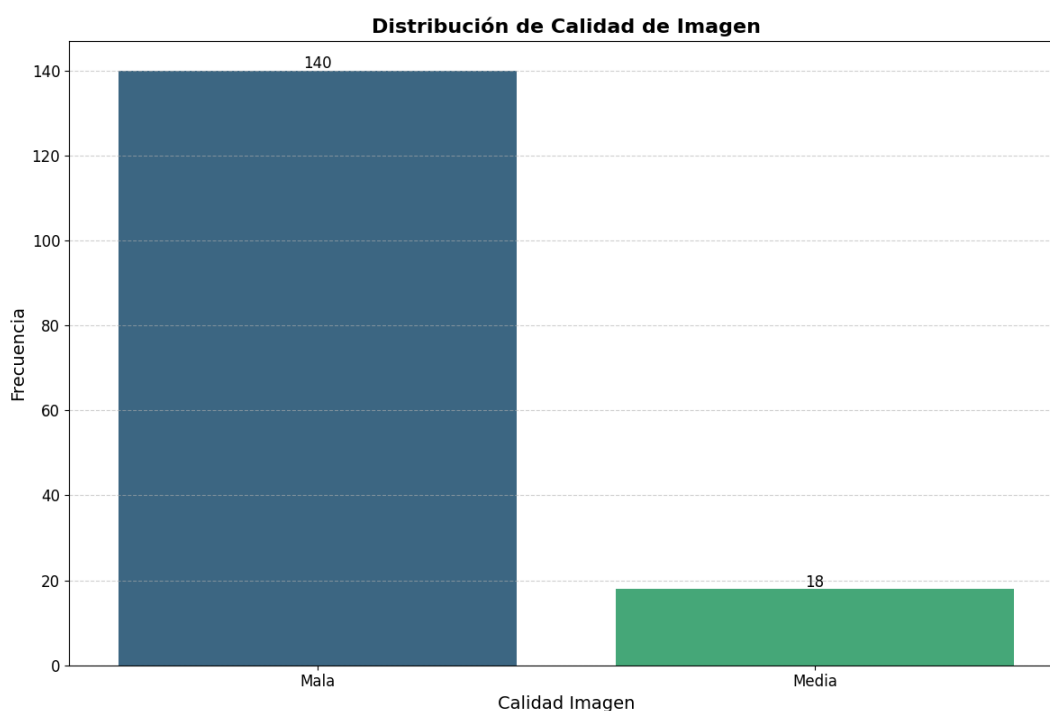


Figura 41. Distribución de calidad de la imagen en la evaluación cualitativa

Nótese, que en los resultados de la evaluación cualitativa (E3.7) se hace referencia a avistamientos no confirmados, aunque en este caso la mayoría corresponde a avistamientos fuera del FOV que no se han tenido en cuenta en la métrica de confirmación visual. La tabla

incluida en el cálculo de las métricas del CRITERIO 2, en el entregable E3.7, detalla para cada video cuales son los avistamientos asociados y si están dentro del FOV.

En cuanto a las métricas extraídas por evento del informe de evaluación cualitativa (CRITERIO 1) resultan más rigurosas al analizar en detalle los falsos positivos (FP). Se observa una baja precisión en los resultados globales, motivada por algunos ejemplos concretos que producen un numero de falsos positivos elevado (73%). Se han identificado como principales causas de FP:

- Olas. El más importante por frecuencia de aparición. Especialmente en condiciones de iluminación en las que las olas producen sombras, que se confunden visualmente con cuerpos de cetáceos especialmente en la imagen RGB. En IR encontramos imágenes ruidosas donde las crestas de las olas son muy visibles, lo que no ayuda.
- Aves. Estos eventos producen FP de forma ocasional y son seguidas por el algoritmo de tracking. Esto podría ayudar a filtrarlas a partir de la velocidad o tipo de movimiento.
- Edificios. De forma muy puntual se han observado FP por edificios en la línea de costa. Esto es poco habitual y normalmente deberían ser filtradas con la estimación de la línea del horizonte.

Es interesante comentar la diferencia en cuanto el número de FP con la evaluación cuantitativa de los algoritmos de detección, que obtenían mucho mejor resultado. Creemos que esto puede ser debido a dos factores. Por un lado, la variabilidad de los resultados evidenciada respecto a la distribución de los datos usados para evaluación, debido al problema de escasez de datos y a la gran diversidad de condiciones. A esto hay que sumar la dificultad de las condiciones ambientales, observándose en los videos de la evaluación cualitativa un número considerable de días con mal estado del mar.

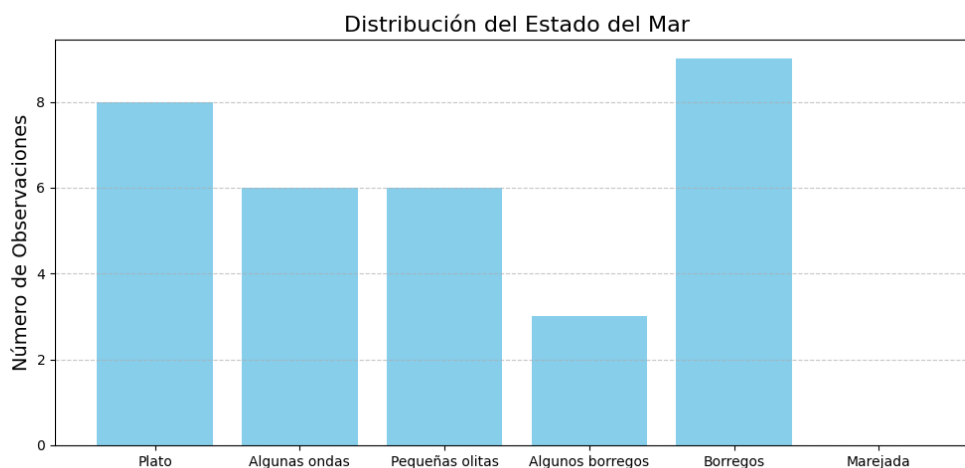


Figura 42. Distribución del estado del mar en la evaluación cualitativa

Por otro lado, un ajuste subóptimo de los umbrales de detección y algoritmos de tracking puede ser responsable de este aumento de FP. Inicialmente se pensó en una configuración para conseguir un balance entre precisión y sensibilidad, pero a la vista de los resultados parece que se ha aumentado demasiado la sensibilidad. Como nota positiva, las configuraciones optimizadas aplicadas a un grupo de videos han mostrado mejores resultados en cuanto al número de falsos positivos como se aprecia en el informe de

evaluación. Sería conveniente en el futuro realizar un análisis más exhaustivo en este sentido, y explorar soluciones para configuración adaptativa de los umbrales de detección con el fin de mejorar la precisión y no solo la sensibilidad.

Otro factor que puede estar impactando en la precisión de los resultados, es el hecho de haber ampliado el problema a resolver. Inicialmente, el objetivo era la detección de cachalotes mediante el soplo, pero ante la escasez de datos se decidió ampliar el problema a la detección de cuerpos. Aunque esto ha permitido poder entrenar los algoritmos y realizar la detección de cetáceos ante la escasez de soplos, estamos utilizando un evento que según trabajos previos es más difícil de detectar, cuyo contraste térmico en IR ha sido cuestionado. Además, se ha pasado de la detección de grandes cetáceos a todo tipo de especies de mucho menor tamaño, lo que supone un gran desafío.

Resulta positiva la tasa de falsos negativos de la evaluación cualitativa en el CRITERIO 1, que se mantiene bastante baja, en un 7% e incluso menor si consideramos las configuraciones optimizadas. Esto pone de manifiesto las capacidades de detección y el potencial para su uso en la práctica a medida que se trabaje para mejorar la precisión.

Más allá de las métricas, este proyecto representa un avance significativo en el estado del arte de la detección térmica de cetáceos, validando técnicas novedosas de fusión multimodal basadas en deep learning y abordando desafíos técnicos como la alineación de imágenes y la escasez de datos.

Consideramos que la decisión de ampliar el alcance del problema hacia la detección de cuerpos no solo ha sido acertada, sino que también ha contribuido a mitigar la escasez de imágenes de soplos. Esta elección podría resultar crucial, ya que se han identificado dificultades en la captura térmica de estos eventos con el sistema óptico evaluado, especialmente en condiciones de larga distancia y visibilidad adversa. Estas limitaciones no solo subrayan la necesidad de mejorar la sensibilidad del sistema óptico, sino que también plantean interrogantes sobre la base misma de la detección térmica de cetáceos, dado que **no se ha podido validar la suposición del contraste térmico del soplo en aguas templadas.**

El uso de técnicas de deep learning mediante la fusión multimodal RGBT ha proporcionado una ventaja significativa en la detección de cuerpos, que generalmente no son considerados óptimos para la detección térmica. Por ello, creemos que la aproximación algorítmica adoptada ha sido adecuada y representa una dirección prometedora para futuros avances.

El conocimiento adquirido a través de esta prueba de concepto debería orientar futuras investigaciones y desarrollos en este campo.

Por otro lado, consideramos que **las limitaciones del sistema óptico han impactado en los resultados obtenidos**, pero se presentan propuestas concretas en la Sección 5 para su mejora. **También se destaca la necesidad de capturar un dataset más amplio y diverso** para impulsar la evolución de los algoritmos, y de forma similar se presentan una serie de mejoras en la sección 5.

En resumen, los resultados presentados reflejan objetivamente tanto los avances como las limitaciones del sistema evaluado. Mientras que la tasa de acierto general es alentadora, los desafíos asociados a la calidad de la imagen, las condiciones del entorno y la configuración de los algoritmos destacan como áreas clave para mejorar. Es evidente que factores como las condiciones del mar, la variabilidad de los datos y los ajustes subóptimos en los parámetros del sistema han influido en la precisión, especialmente en términos de falsos positivos y confirmaciones visuales. Sin embargo, las configuraciones optimizadas han demostrado un potencial prometedor para mejorar el rendimiento. De cara al futuro, será crucial abordar estos aspectos mediante estrategias adaptativas, un mayor control en la recolección de datos y enfoques más centrados en la detección de soplos frente a cuerpos, lo que permitirá mejorar la robustez del sistema y ampliar su aplicabilidad en condiciones más diversas.

5. MEDIDAS TÉCNICAS PARA SU APLICACION

En vista de los resultados y conclusiones presentados, consideramos que se ha logrado un avance significativo en la detección de cetáceos a bordo de ferris rápidos. La tecnología de detección mediante fusión multimodal y multifocal basada en deep learning ha obtenido resultados positivos, incluso con dificultades constatadas como la escasez de datos para entrenamiento de los modelos.

El desarrollo y la evaluación de la solución también han permitido identificar limitaciones y oportunidades de mejora, lo que sienta las bases para futuros trabajos. La demostración de la tecnología se ha llevado a cabo en un entorno controlado de laboratorio, procesando grabaciones en diferido y con una serie de suposiciones que no tienen por qué cumplirse en un entorno real en producción.

La transición hacia un sistema de detección automático o semi-automático operativo en un entorno real requiere abordar consideraciones prácticas adicionales, como el procesamiento de video en tiempo real y la integración con el sistema de sensores y otros instrumentos de navegación. Estos aspectos serán fundamentales para garantizar el desempeño y la aplicabilidad del sistema en condiciones reales.

A continuación, se resumen una serie de posibilidades de mejora identificadas, a nivel del sistema óptico, algoritmos y campañas de recolección de datos. Finalmente se presentan serie de medidas técnicas para el paso a producción y su estimación presupuestaria.

5.1. SENSORES Y SISTEMA ÓPTICO

A pesar de que el sistema de sensores ópticos utilizado cumple con las especificaciones requeridas y de que se considera la mejor opción disponible en el mercado en el momento de la realización del proyecto, se han identificado mejoras relevantes a nuestro caso de uso.

Mejora	Descripción
Resolución radiométrica	Incrementar la resolución radiométrica de 8 bits a 16 bits (habitual en imágenes raw), especialmente en el sensor IR, permitiría detectar diferencias de temperatura menores. Esto es

	especialmente relevante para detección térmica de cetáceos en aguas templadas como ocurre en la zona de Canarias.
Rango dinámico	Actualmente el rango dinámico es adaptativo según los valores máximos y mínimos en la escena. Esto evita saturaciones, pero produce efectos no deseables (1) relación temperatura/intensidad de píxel es variable; (2) el rango dinámico puede ser muy amplio o no incluir las temperaturas de interés. Un rango dinámico más estrecho pero que incluya las temperaturas de interés puede mejorar el contraste de los soplos y cuerpos a detectar.
Sensibilidad del sensor IR	Se ha observado bajo rendimiento del sensor IR a largas distancias y en condiciones de baja visibilidad (niebla/calima), especialmente en comparación con el sensor RGB. Dentro de las limitaciones del espectro IR, una mayor sensibilidad del sensor IR o mejor ajuste del rango dinámico podría mitigar este problema.
Óptica con polarizadores	El uso de polarizadores ayudaría a filtrar reflexiones de la luz en la superficie del agua y crestas de olas, disminuyendo el ruido tanto en imágenes RGB como IR y permitiendo una visión más clara de los objetos en la superficie.
Estabilización mecánica	El sistema actual es capaz de corregir los ángulos de pitch (cabeceo) y yaw (giñada o deriva) gracias al movimiento PTZ de la cámara en azimut y elevación, pero no corrige mecánicamente el ángulo de roll (balanceo). El balanceo se corrige mediante software, aunque se ha evidenciado que esta corrección es insuficiente a veces, lo cual se manifiesta por una inclinación de la línea del horizonte. Una mejor estabilización (mecánica o software) mejoraría el rendimiento de los algoritmos de tracking al facilitar la asociación de detecciones en múltiples frames.
Reducción de vibraciones	Se ha evidenciado cierta inestabilidad en las ópticas TELE, que se manifiesta como vibraciones de alta frecuencia en la imagen y produce efectos 'motion blur' (desenfoque). Esto puede ser debido a las condiciones de instalación y del propio buque, pero también a características de fabricación. La solución puede ser una combinación de mejoras en la instalación, mejoras del equipamiento óptico y mejoras software para estabilización digital.

Tabla 12. Resumen de mejoras de los sensores

5.2. MÉTODOS Y ALGORITMOS

Durante el desarrollo del proyecto se ha trabajado al borde del estado del arte de la detección multimodal RGBT para detección de objetos en tiempo real, identificándose desafíos clave que en muchos casos continúan siendo problemas abiertos. Avanzar en la

solución de estos desafíos permitirá mejorar los métodos de detección y desarrollar medidas más confiables.

Mejora	Descripción
Mejora del co-registro multimodal de imagen	Se ha identificado como uno de los principales desafíos para la eficacia de la detección multimodal. Los resultados obtenidos han permitido mitigar el su efecto, pero continúa siendo un problema abierto. Se recomienda explorar soluciones del estado del arte identificadas como [15], [16]
Detección con tolerancia al desalineamiento.	La técnica de fusión middle-fusion CFT que se ha utilizado ha demostrado resultados aceptables cuando el alineamiento RGB-T no ha sido perfecto, posiblemente gracias al uso de capas de atención (Transformers), que permiten establecer relaciones espaciales multiescala y con diferente contexto. Para mejorar los resultados convendría explorar métodos de detección de objetos que abordan de forma explícita el problema de desalineamiento en la línea de publicaciones recientes [31], [44], [45].
Ampliación de conjuntos de datos	Uno de los aspectos clave para mejorar el rendimiento de los modelos de IA es disponer de un conjunto de datos amplio y diverso para entrenamiento y test. Deberán recogerse nuevos datos que cubran una mayor variedad de escenarios, condiciones ambientales y ángulos de captura. Asimismo, podrán utilizarse metodologías de etiquetado semiautomático utilizando los propios modelos de detección y revisión manual de las etiquetas.
Generación sintética de datos	Investigación de técnicas de aumentación de datos mediante generación sintética para pares de imágenes RGB-T de forma consistente, y coherente con la información térmica de la escena. Este tipo de técnicas pueden ser de gran ayuda para mitigar el problema de escasez de datos. La generación sintética de datos y la aumentación de datos son aspectos cruciales que deben seguir siendo estudiados y perfeccionados
Evaluación nocturna	Evaluación y adaptación de los métodos de detección en condiciones de baja luminosidad y oscuridad total.
Mejoras en cálculo de distancia	Actualmente basado en el cálculo de distancia mediante el método de triangulación esférica [43] utilizado por Zitterbart [3]. Las limitaciones de este método ya han sido expuestas y deberá trabajarse en mitigar estas limitaciones o explorar métodos alternativos basados en visión por computador como la proyección de vista de pájaro (Bird's-Eye-View, BEV). Estas mejoras pueden incluir la integración el software con el sistema de navegación GPS

	del buque o sensores inerciales IMU para medir la inclinación de la cámara con el movimiento del barco.
Mejoras en algoritmos de detección.	Mejoras en algoritmos de detección y métodos de fusión para aumentar la precisión y sensibilidad del detector. Parte de estas mejoras estarían enfocadas a realizar una optimización de hiperparámetros exhaustiva, al reentrenamiento con conjuntos de datos más amplios, o al refinamiento de los métodos de detección y fusión. Convendría estudiar nuevos métodos del estado del arte que han surgido en el transcurso de este proyecto, como nuevas versiones de YOLO o técnicas para mejorar la detección de objetos pequeños.
Robustez frente a funcionamiento degradado	Resulta conveniente estudiar la robustez de los métodos de detección cuando uno de los modos no contiene información útil. Esto puede producirse por ejemplo en condiciones de iluminación extremas, situaciones de baja sensibilidad del modo IR o un mal funcionamiento de alguno de los sensores. Igualmente será de utilidad desarrollar técnicas para estimar la confiabilidad o contribución de cada uno de los modos y poder ajustar el peso de cada uno dinámicamente.
Mejoras de los algoritmos de tracking	Actualmente, el tracking de detecciones se aplica de forma monomodal (RGB o IR). Se recomienda extender su funcionamiento a tracking multimodal para obtener mejores resultados. Además, se han identificado áreas de mejora en la asociación mediante descriptores de apariencia para el caso de cetáceos, ya que la implementación actual está optimizada para personas. El tracker se utiliza principalmente para confirmar detecciones, pero sería útil realizar un seguimiento continuo del objeto detectado para monitorizar y actualizar el nivel de riesgo según su movimiento. Estas mejoras permitirían desarrollar un sistema avanzado de alarmas basado en el seguimiento de detecciones.

Tabla 13. Resumen de mejoras de métodos y algoritmos

5.3. CAMPAÑAS Y CAPTURA DE DATOS

Se han identificado igualmente oportunidades de mejora en las campañas de avistamientos que se han llevado a cabo. Estas mejoras están enfocadas a maximizar la cantidad de datos útiles recogidos para cumplir con las demandas que requieren los modelos de IA para su entrenamiento y validación.

Mejora	Descripción
Campañas específicas para captura de datos	Se ha evidenciado que la captura de datos desde ferris rápidos en el transcurso de su ruta comercial no es óptima. Deben llevarse a

	cabo campañas específicas para captura de datos a bordo de embarcaciones que puedan desplazarse con libertad para maximizar el número de eventos de interés capturados, idealmente con el mismo tipo de sensores usados en producción.
Datos anotados durante las campañas	En dichas campañas debe ser una prioridad la captura y anotación de datos útiles para el entrenamiento de los modelos de detección frente a otros datos de interés biológico que puedan afectar a la capacidad de los observadores para realizar nuevas observaciones.
Campañas nocturnas	Evaluar el uso de prismáticos con visión nocturna para identificación de eventos de referencia que puedan ser utilizados para el entrenamiento y la evaluación de los algoritmos para detección nocturna.
Confirmación visual de observaciones en video	Este proceso es clave para la creación de un dataset de detección de objetos como se ha reflejado en el informe. En futuros trabajos, debería contarse con recursos de observadores profesionales para que también participen en este proceso directamente.

Tabla 14. Resumen de mejoras campañas y captura de datos

5.4. PASO A PRODUCCIÓN Y PILOTO

Se presentan medidas específicas para avanzar hacia la fase de producción, enfocadas en desarrollo e ingeniería para implementar y probar un piloto, diferenciándose de los esfuerzos previos de investigación (mejoras descritas en Tabla 12, Tabla 13, Tabla 14).

Se recomienda implementar estas medidas junto con las mejoras propuestas. De aplicarse únicamente estas acciones, se podría obtener un sistema operativo integrado en los ferris y sujeto a futuras mejoras. No obstante, este sistema estaría limitado a las capacidades de detección actuales de los algoritmos y el sistema óptico, además de posibles sesgos derivados de la limitada diversidad en los datasets empleados.

El objetivo de esta fase es desarrollar un prototipo de nivel de madurez TRL 7 [46], con prueba piloto en un entorno operacional, asumiendo su integración con el sistema óptico Sentry en los ferris rápidos de Fred Olsen. El avance hacia TRL 9 para un producto comercial certificado se abordaría en una fase posterior.

Medida	Descripción
Procesamiento de video en tiempo real	Actualmente el demostrador trabaja de manera offline, donde los videos son procesados previamente por los algoritmos de detección y tracking. Posteriormente los resultados se visualizan en la GUI de forma diferida. Para llevar este sistema a producción y adaptarlo a un entorno en tiempo real, sería necesario transformar el proceso en un flujo online, en el cual las detecciones se realicen de manera simultánea a la lectura del video y se muestren en pantalla con un retraso mínimo. Esto implicaría un esfuerzo

	considerable en términos de reimplementación del código y optimización, incluyendo la optimización y compresión de los modelos de IA para reducir el tiempo de inferencia.
Lectura de video desde el sistema óptico	Actualmente, el demostrador utiliza videos que han sido previamente grabados, sincronizados y convertidos a un formato compatible para su procesamiento. Un sistema en producción debe ser capaz de leer las fuentes de video en streaming directamente desde el sistema óptico. Esto puede ser un proceso complejo ya que debe asegurar la sincronización temporal de los modos RGB e IR para la fusión de datos. La eficacia de esta sincronización puede depender de las capacidades del sistema óptico, por ejemplo, del envío de metadatos con marca de tiempo en cada frame.
Lectura de metadatos y posición de la cámara	En el sistema actual, el algoritmo del cálculo del rumbo de las detecciones asume que la cámara está orientada hacia la proa. Un sistema en producción debe ser capaz de obtener información del sistema óptico para conocer el ángulo de rotación de la cámara. Esto permitiría además utilizar el sistema con otros modos de operación, como el barrido en un sector. Puede ser útil leer otros metadatos, como la inclinación de la cámara por el movimiento del barco gracias a los sensores IMU integrados. Previamente deberá analizarse el protocolo de envío de metadatos por la cámara Sentry y su sincronización con los frames de video.
Alertas con seguimiento de posición	El demostrador implementa un sistema de alertas básico, con una estimación del riesgo de colisión basado en la distancia y en el rumbo de la observación en el momento de ser detectada. Es conveniente que, en producción, se haga un seguimiento de las detecciones y actualización de su posición de forma que los objetos que permanezcan en el rumbo del buque generen alertas de forma repetida.
Ajuste de la sensibilidad de detección	Un sistema en producción debe disponer de un mecanismo para el ajuste de la sensibilidad de los algoritmos de detección y tracking para poder adaptarse a diferentes condiciones ambientales y de funcionamiento. Deberán analizarse diferentes opciones, incluido el ajuste manual simplificado mediante una sola variable de control, ajuste manual con configuraciones predefinidas, ajuste semi-automático mediante introducción de parámetros ambientales, y el ajuste automático mediante análisis de la imagen o realimentación con otros sensores disponibles.
Integración con sistemas de navegación	La integración con el GPS del barco y el uso de cartografía es fundamental para mejorar la precisión en la estimación de

	distancia y contextualización de las observaciones. Gracias al registro de las coordenadas GPS, se podría calcular de manera exacta la distancia a las observaciones cuando no se disponga de la línea del horizonte como referencia en posiciones cercanas a la costa. En su defecto deberán estudiarse otras alternativas a la estimación de distancia que no tengan esta dependencia.
Diseño y desarrollo de interfaz operacional	La interfaz o GUI del demostrador es una versión simplificada, diseñada con una funcionalidad mínima orientada a la validación de la tecnología. Para su implementación en un sistema en producción, será necesario diseñar y desarrollar una nueva interfaz desde cero, adaptada a los requerimientos específicos de operación y usabilidad del sistema final. Esta nueva interfaz deberá implementar de forma efectiva el concepto HITL (Human-In-The-Loop), permitiendo una interacción fluida entre el operador humano y el sistema automatizado. Esto incluye la capacidad de supervisar y corregir decisiones automáticas, introducir ajustes en tiempo real y validar los resultados generados por el sistema. Además, la interfaz deberá priorizar la accesibilidad, la presentación clara de la información crítica y la minimización de la carga cognitiva del usuario para garantizar una operación eficiente y segura en condiciones reales.
Mantenimiento y mejora continua de los modelos.	El mantenimiento y mejora continua de los modelos de IA para detección es crucial para garantizar que estos sigan siendo efectivos y confiables en entornos reales durante todo el ciclo de vida, donde las condiciones pueden variar significativamente respecto a las vistas durante el entrenamiento. Este proceso no solo incluye la actualización periódica de los modelos, sino también la implementación de un enfoque sistemático para su desarrollo, despliegue, monitoreo y mejora, un marco conocido como MLOps (Machine Learning Operations). La mejora continua requiere diseñar e implementar procesos para automatizar la recolección y etiquetado de nuevos datos, mantenimiento de los datasets, reentrenamiento y evaluación de los modelos, así como su despliegue. A menudo es más eficiente que este mantenimiento se ofrezca como servicio con un coste añadido. En cualquier caso, el software deberá contar con la funcionalidad que permita la implementación de estos procesos.
Integración y pruebas	Integrar los distintos componentes técnicos y de software del prototipo en un entorno controlado. Incluye la realización de pruebas funcionales y de compatibilidad para garantizar que el sistema opera como se espera antes de su implementación en un entorno real.
Desarrollo de prototipo piloto	Desarrollo del prototipo final para su validación mediante prueba piloto, incluyendo el equipamiento necesario. Despliegue del

	prototipo integrado en el entorno operativo, asegurando su correcta instalación y configuración. Esta fase también contempla la recolección de datos adicionales y la optimización exhaustiva de los modelos de detección mediante ajustes de hiperparámetros para mejorar su rendimiento y precisión.
Prueba piloto	Fase enfocada en la ejecución del prototipo en un entorno real, con el objetivo de validar su funcionamiento, recopilar datos de desempeño y realizar ajustes necesarios. Este proceso permite refinar el sistema, evaluar su impacto operativo y preparar la transición hacia un despliegue completo.
Documentación técnica y formación.	Como en cualquier sistema en producción, la generación de documentación técnica detallada y la formación adecuada del personal encargado de su operación son aspectos fundamentales para garantizar el correcto funcionamiento y mantenimiento del sistema.

Tabla 15. Resumen de medidas técnicas para paso a producción y piloto

5.5. VALORACIÓN PRESUPUESTARIA

Para avanzar en los desafíos de investigación descritos en las tablas (Tabla 13, Tabla 14) se estima que es necesario un proyecto de investigación aplicada con una dedicación de 4 investigadores durante 18 meses y campañas de observación específicas para recolección de datos (Tabla 16). La viabilidad de las mejoras referentes a los equipos ópticos y sensores (Tabla 12) deberá estudiarse con el fabricante de los mismos y estimarse, en su caso, los costes que estas implicarían.

Concepto	Recursos	Estimación aproximada (€)
Proyecto I+D para avance del estado del arte en detección RGBT multimodal de cetáceos.	4 investigadores, 18 meses. Costes operacionales. Costes de computación y almacenamiento de datos.	609.960
Campañas de observación para recolección de datos y validación. Presupuesto para 90 salidas en barco para oceanografía.	2 observadores MMO. Equipos ópticos. Costes operacionales.	280.000

Tabla 16. Valoración presupuestaria para avance del estado del arte

Para el paso a producción y prueba piloto con un TRL 7, se estima que es necesario un proyecto de desarrollo e innovación con un esfuerzo humano expresado en PM (Personas-Mes) según la Tabla 17. Y con una valoración presupuestaria detallada por paquetes de trabajo correspondientes a las medidas descritas en la Tabla 15.

ID	Paquete de trabajo	Dedicación (PM)	Coste estimado (€)
WP1	Procesamiento de video en tiempo real	10	68.050
WP2	Lectura de video desde el sistema óptico	5	34.025
WP3	Lectura de metadatos y posición de la cámara	3	20.415
WP4	Alertas con seguimiento de posición	5	34.025
WP5	Ajuste de la sensibilidad de detección	5	34.025
WP6	Integración con sistemas de navegación	4	27.220
WP7	Diseño y desarrollo de interfaz operacional	10	68.050
WP8	Mantenimiento y mejora continua (MLOPS)	6	40.830
WP9	Integración y pruebas	4	27.220
WP10	Prototipo piloto	15	102.075
WP11	Prueba piloto	4	27.220
WP12	Documentación técnica y formación.	3	20.415
WP13	Coordinación de proyecto	4	27.220
	TOTAL	78	530.790

Tabla 17. Valoración presupuestaria para para paso a producción y piloto

Adicionalmente deberán tenerse en cuenta los costes adicionales de viajes, computación y almacenamiento de datos, y equipamiento hardware para el prototipo (Tabla 18). En caso de ejecutarse previamente la fase de I+D, estos gastos podrían reducirse al compartirse en parte.

Concepto	Descripción	Estimación aproximada (€)
Viajes y desplazamientos	Viajes y desplazamientos durante el desarrollo del proyecto.	5.000
Computación y almacenamiento de datos	Infraestructura de computación y almacenamiento masivo de datos. Gastos de computación en la nube. Hardware de computación para el prototipo piloto.	20.000
Equipos ópticos	Cámara Sentry para calibración, integración y pruebas. Incluida junto al prototipo piloto.	50.000

Tabla 18. Costes adicionales para paso a producción y piloto

Por tanto, se calcula un **presupuesto de casi 1,5M€** para avanzar en la generación de una aplicación en producción con un **nivel de madurez TRL 7**, en un **plazo aproximado de entre 2 y 3 años**. La estimación considera un proyecto previo de investigación de 18 meses y un esfuerzo de 78 personas-mes para el paso a producción con un equipo de 4-5 personas (estimado en 15-19 meses). Además, se ha tenido en cuenta la posibilidad de solapar parcialmente ambas fases, especialmente en la parte final del proyecto de investigación.

6. REFERENCIAS

- [1] «CVAT». Accedido: 20 de abril de 2023. [En línea]. Disponible en: <https://www.cvat.ai/>
- [2] T.-Y. Lin *et al.*, «Microsoft COCO: Common Objects in Context», 20 de febrero de 2015, *arXiv*: arXiv:1405.0312. doi: 10.48550/arXiv.1405.0312.
- [3] D. P. Zitterbart, L. Kindermann, E. Burkhardt, y O. Boebel, «Automatic Round-the-Clock Detection of Whales for Mitigation from Underwater Noise Impacts», *PLOS ONE*, vol. 8, n.º 8, p. e71217, ago. 2013, doi: 10.1371/journal.pone.0071217.
- [4] D. P. Zitterbart *et al.*, «Scaling the Laws of Thermal Imaging–Based Whale Detection», *Journal of Atmospheric and Oceanic Technology*, vol. 37, n.º 5, pp. 807-824, may 2020, doi: 10.1175/JTECH-D-19-0054.1.
- [5] «ObsPy Documentation (1.4.1) — ObsPy 1.4.1 documentation». Accedido: 16 de octubre de 2024. [En línea]. Disponible en: <https://docs.obspy.org/>
- [6] E. Rublee, V. Rabaud, K. Konolige, y G. Bradski, «ORB: An efficient alternative to SIFT or SURF», en *2011 International Conference on Computer Vision*, nov. 2011, pp. 2564-2571. doi: 10.1109/ICCV.2011.6126544.
- [7] T. Lindeberg, «Scale Invariant Feature Transform», en *Scholarpedia*, vol. 7, 2012. doi: 10.4249/scholarpedia.10491.
- [8] X. Jiang, J. Ma, G. Xiao, Z. Shao, y X. Guo, «A review of multimodal image matching: Methods and applications», *Information Fusion*, vol. 73, pp. 22-71, sep. 2021, doi: 10.1016/j.inffus.2021.02.012.
- [9] J. Canny, «A Computational Approach to Edge Detection», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, n.º 6, pp. 679-698, nov. 1986, doi: 10.1109/TPAMI.1986.4767851.
- [10] S. Baker y I. Matthews, «Lucas-Kanade 20 Years On: A Unifying Framework», *International Journal of Computer Vision*, vol. 56, n.º 3, pp. 221-255, feb. 2004, doi: 10.1023/B:VISI.0000011205.11775.fd.
- [11] J. Shi y Tomasi, «Good features to track», en *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, jun. 1994, pp. 593-600. doi: 10.1109/CVPR.1994.323794.
- [12] Y.-H. Tseng, «Image-to-Image Registration by Matching Area Features Using Fourier Descriptors and Neural Networks», 1997.
- [13] K. Liu, Z. Ren, X. Wu, J. Di, y J. Zhao, «SSG-Net: A robust network for adaptive multi-source image registration based on SuperGlue», *Digital Signal Processing*, vol. 140, p. 104128, ago. 2023, doi: 10.1016/j.dsp.2023.104128.
- [14] «SuperGlue CVPR 2020». Accedido: 23 de mayo de 2024. [En línea]. Disponible en: <https://psarlin.com/superglue/>
- [15] D. Wang, J. Liu, L. Ma, R. Liu, y X. Fan, «Improving Misaligned Multi-modality Image Fusion with One-stage Progressive Dense Registration», 22 de agosto de 2023, *arXiv*: arXiv:2308.11165. Accedido: 18 de octubre de 2024. [En línea]. Disponible en: <http://arxiv.org/abs/2308.11165>
- [16] D. Wang, J. Liu, X. Fan, y R. Liu, «Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration», 24 de mayo de 2022, *arXiv*: arXiv:2205.11876. Accedido: 27 de mayo de 2024. [En línea]. Disponible en: <http://arxiv.org/abs/2205.11876>
- [17] K. Cui, J. Huang, Z. Luo, G. Zhang, F. Zhan, y S. Lu, *GenCo: Generative Co-training on Data-Limited Image Generation*. 2021. doi: 10.48550/arXiv.2110.01254.
- [18] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, y T. Aila, «Training Generative Adversarial Networks with Limited Data», 7 de octubre de 2020, *arXiv*: arXiv:2006.06676. doi: 10.48550/arXiv.2006.06676.

- [19] «Adobe Firefly - IA generativa gratuita para el personal creativo». Accedido: 16 de octubre de 2024. [En línea]. Disponible en: <https://www.adobe.com/es/products/firefly.html>
- [20] H. Zhang, S. Zhang, y R. Zou, «Select-Mosaic: Data Augmentation Method for Dense Small Object Scenes», 8 de junio de 2024, *arXiv*: arXiv:2406.05412. doi: 10.48550/arXiv.2406.05412.
- [21] «Albumentations: Efficient Image Augmentation Library for Machine Learning». Accedido: 16 de octubre de 2024. [En línea]. Disponible en: <https://albumentations.ai>
- [22] G. Ghiasi *et al.*, «Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation», 23 de junio de 2021, *arXiv*: arXiv:2012.07177. doi: 10.48550/arXiv.2012.07177.
- [23] A. Mukherjee, *Opletts/Copy-Paste-Augmenter*. (4 de julio de 2024). Python. Accedido: 16 de octubre de 2024. [En línea]. Disponible en: <https://github.com/Opletts/Copy-Paste-Augmenter>
- [24] P. Pérez, M. Gangnet, y A. Blake, «Poisson image editing», *ACM Trans. Graph.*, vol. 22, n.º 3, pp. 313-318, jul. 2003, doi: 10.1145/882262.882269.
- [25] C.-Y. Wang, A. Bochkovskiy, y H.-Y. M. Liao, «YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors», 6 de julio de 2022, *arXiv*: arXiv:2207.02696. doi: 10.48550/arXiv.2207.02696.
- [26] S. Hwang, J. Park, N. Kim, Y. Choi, y I. S. Kweon, «Multispectral pedestrian detection: Benchmark dataset and baseline», en *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, jun. 2015, pp. 1037-1045. doi: 10.1109/CVPR.2015.7298706.
- [27] K. Roszyk, M. R. Nowicki, y P. Skrzypczyński, «Adopting the YOLOv4 Architecture for Low-Latency Multispectral Pedestrian Detection in Autonomous Driving», *Sensors*, vol. 22, n.º 3, Art. n.º 3, ene. 2022, doi: 10.3390/s22031082.
- [28] S. Yoon y J. Cho, «Deep Multimodal Detection in Reduced Visibility Using Thermal Depth Estimation for Autonomous Driving», *Sensors*, vol. 22, n.º 14, Art. n.º 14, ene. 2022, doi: 10.3390/s22145084.
- [29] F. Qingyun, H. Dapeng, y W. Zhaokui, «Cross-Modality Fusion Transformer for Multispectral Object Detection», 4 de octubre de 2022, *arXiv*: arXiv:2111.00273. doi: 10.48550/arXiv.2111.00273.
- [30] X. Jia, C. Zhu, M. Li, W. Tang, S. Liu, y W. Zhou, «LLVIP: A Visible-infrared Paired Dataset for Low-light Vision», 14 de junio de 2023, *arXiv*: arXiv:2108.10831. doi: 10.48550/arXiv.2108.10831.
- [31] L. Zhang *et al.*, «Weakly Aligned Feature Fusion for Multimodal Object Detection», 21 de abril de 2022, *arXiv*: arXiv:2204.09848. doi: 10.48550/arXiv.2204.09848.
- [32] F. Zhuang *et al.*, «A Comprehensive Survey on Transfer Learning», 23 de junio de 2020, *arXiv*: arXiv:1911.02685. Accedido: 21 de octubre de 2024. [En línea]. Disponible en: <http://arxiv.org/abs/1911.02685>
- [33] C. Liu, C. Gao, X. Xia, D. Lo, J. Grundy, y X. Yang, «On the Replicability and Reproducibility of Deep Learning in Software Engineering», 25 de junio de 2020, *arXiv*: arXiv:2006.14244. doi: 10.48550/arXiv.2006.14244.
- [34] C.-Y. Wang, A. Bochkovskiy, y H.-Y. M. Liao, «YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors», 6 de julio de 2022. Accedido: 19 de diciembre de 2022. [En línea]. Disponible en: <http://arxiv.org/abs/2207.02696>
- [35] K. Roszyk, M. R. Nowicki, y P. Skrzypczyński, «Adopting the YOLOv4 Architecture for Low-Latency Multispectral Pedestrian Detection in Autonomous Driving», *Sensors*, vol. 22, n.º 3, Art. n.º 3, ene. 2022, doi: 10.3390/s22031082.
- [36] *onnx/onnx*. (20 de mayo de 2024). Python. Open Neural Network Exchange. Accedido: 20 de mayo de 2024. [En línea]. Disponible en: <https://github.com/onnx/onnx>
- [37] «ONNX Runtime | Home». Accedido: 20 de mayo de 2024. [En línea]. Disponible en: <https://onnxruntime.ai>

- [38] Z. Soleimanitaleb y M. A. Keyvanrad, «Single Object Tracking: A Survey of Methods, Datasets, and Evaluation Metrics», 31 de enero de 2022, *arXiv*: arXiv:2201.13066. doi: 10.48550/arXiv.2201.13066.
- [39] N. Wojke, A. Bewley, y D. Paulus, «Simple Online and Realtime Tracking with a Deep Association Metric», 21 de marzo de 2017, *arXiv*: arXiv:1703.07402. doi: 10.48550/arXiv.1703.07402.
- [40] G. Welch y G. Bishop, «An Introduction to the Kalman Filter», 2002.
- [41] *opencv/opencv*. (16 de octubre de 2024). C++. OpenCV. Accedido: 16 de octubre de 2024. [En línea]. Disponible en: <https://github.com/opencv/opencv>
- [42] N. Otsu, «A Threshold Selection Method from Gray-Level Histograms», *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, n.º 1, pp. 62-66, ene. 1979, doi: 10.1109/TSMC.1979.4310076.
- [43] J. Lerczak y R. Hobbs, «Calculating sighting distances from angular readings during shipboard, aerial, and shore-based marine mammal surveys», *Marine Mammal Science*, vol. 14, pp. 590-598, ago. 2006, doi: 10.1111/j.1748-7692.1998.tb00745.x.
- [44] Y. Xing, S. Wang, S. Zhang, G. Liang, X. Zhang, y Y. Zhang, «MS-DETR: Multispectral Pedestrian Detection Transformer with Loosely Coupled Fusion and Modality-Balanced Optimization», 11 de noviembre de 2023, *arXiv*: arXiv:2302.00290. Accedido: 25 de marzo de 2024. [En línea]. Disponible en: <http://arxiv.org/abs/2302.00290>
- [45] C. Chen *et al.*, «Weakly Misalignment-Free Adaptive Feature Alignment for UAVs-Based Multimodal Object Detection», en *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun. 2024, pp. 26826-26835. doi: 10.1109/CVPR52733.2024.02534.
- [46] Juan Miguel Ibáñez de Aldecoa Quintana, «NIVELES DE MADUREZ DE LA TECNOLOGÍA. TECHNOLOGY READINESS LEVELS. TRLS». [En línea]. Disponible en: <https://www.mintur.gob.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/393/NOTAS.pdf>

ANEXO 1 – LISTADO DE ENTREGABLES

A continuación, se indica el listado de entregables que se incluyen junto al informe final E3. Por cuestiones prácticas la entrega se realizará en soporte físico (memoria USB), incluyendo el formato de archivo más apropiado en cada caso.

ID	Descripción	Formato
E3.1	Diario de embarques correspondientes a las campañas de avistamientos requeridas por la licitación. Incluye el listado de observaciones con información detallada. Datos anotados por observadores MMO profesionales a cargo de la ULL.	XLS
E3.2	Observaciones sistema Vigía proporcionadas por Fred.Olsen como datos adicionales a los requeridos en la licitación. Datos han anotados por personal de la tripulación de Fred.Olsen.	XLS
E3.3	Volcado de datos SIG correspondientes a las observaciones de las campañas de avistamientos (E3.1). Estos datos han sido incluidos en la base de datos de carácter científico BIOTA, como alternativa a la base de datos REDPROMAR como mejora respecto a los requerimientos de la licitación	SIG
E3.4	Demostrador tecnológico para la evaluación de la tecnología y las medidas de detección. Incluye la aplicación instalable, modelos de detección, y manual de uso e instalación.	Aplicación instalable
E3.5	Videos originales utilizados para la validación cualitativa	MP4
E3.6	Resultados del procesamiento para validación cualitativa. Incluye los resultados generados al procesar los videos para evaluación cualitativa (E3.5) mediante la ejecución del demostrador tecnológico (E3.4). En particular: configuración del demostrador (.yaml), listado de detecciones (.json), subclips con detecciones (.mp4) y video completo con detecciones resaltadas (.mp4).	Varios
E3.7	Resultados completos de la validación cualitativa realizada por la ULL para cada uno de los videos procesados.	XLS