Simposio CEA de Robótica, Bioingeniería, Visión por Computador y Automática Marina 2026

Sistema de visión para estimación de pose robusta en entornos de radiación

Andoni Benito González[a,1], Marta Centeno Tortajada[a,1], Endika Llano Matxin[a,1], Ainhoa Ruiz Delgado[a,1,*], Verónica González Ruiz[a]

[a] *SENER Aeroespacial, Getxo, España.*

Resumen

Los entornos con radiación ionizante presentan condiciones especialmente adversas para los sistemas de visión, generando ruido impulsivo, *hot pixels* y artefactos que degradan la fiabilidad de la percepción. En este trabajo presentamos un subsistema de estimación de pose de seis grados de libertad (6 GDL) basado exclusivamente en imágenes monocromas procedentes de una configuración de doble cámara. El objetivo es garantizar detección y localización robusta de componentes en instalaciones donde la radiación dificulta o imposibilita el uso de sensores convencionales o métodos basados en marcadores.

El modelo, entrenado con 20 000 muestras capturadas mediante un manipulador UR10e y reforzado con técnicas de aumento que reproducen degradaciones similares a las inducidas por radiación, mantiene bajos errores de traslación y orientación en el conjunto de prueba incluso bajo niveles significativos de ruido. Los resultados experimentales confirman la viabilidad de una solución ligera, *marker-free* y tolerante a perturbaciones para apoyar tareas autónomas de inspección y mantenimiento en entornos radiológicos.

Palabras clave: Estimación de pose, 6 GDL, Visión artificial, Aprendizaje profundo, Robótica autónoma, Entornos de radiación

Vision System for Robust Pose Estimation in Radiation Environments

Abstract

Ionizing-radiation environments are particularly adverse for vision systems, producing impulsive noise, hot pixels, and image artifacts that degrade perception reliability. This work presents a six-degree-of-freedom (6-DoF) pose-estimation subsystem based solely on monochrome images from a rigid dual-camera setup. The goal is robust component detection and localization in facilities where radiation complicates or prevents the use of conventional sensors or marker-based approaches.

Trained on 20,000 samples captured with a UR10e manipulator and strengthened with augmentations that emulate radiation-like degradations, the model achieves low translation and rotation errors on a held-out test set, even under significant noise conditions. Experimental validation indicates that a lightweight, marker-free, disturbance-tolerant solution can support autonomous inspection and maintenance tasks in radiological environments.

Keywords: Pose Estimation, 6-DoF, Computer Vision, Deep Learning, Autonomous Robotics, Radiation Environments

1. Introducción

La radiación ionizante presente en instalaciones nucleares, médicas o industriales degrada de forma notable el rendimiento de los sensores de imagen, generando *hot pixels*, ruido impulsivo e incremento de corriente oscura. Estos efectos han sido ampliamente observados en sensores CMOS irradiados bajo dosis acumuladas significativas, afectando directamente a tareas de inspección y mantenimiento remoto (Liu et al., 2023). De forma análoga, los sensores CCD pueden presentar incremento de corriente oscura, proliferación de *hot pixels* y degradación de la eficiencia de transferencia de carga (CTE) bajo radiación ionizante, efectos ampliamente documentados en instrumentación espacial (Sirianni and Mutchler, 2006).

En paralelo, muchos sistemas de percepción en robótica dependen de marcadores fiduciarios para asegurar detección y registro geométrico. Sin embargo, los métodos basados en marcadores pueden fallar ante oclusiones, suciedad sobre las superficies, iluminación no controlada o la aparición de artefactos visuales. En un entorno radiológico, donde la degradación puede

*Autor para correspondencia: ainhoa.ruiz@aeroespacial.sener

[1] Estos autores han contribuido por igual a este trabajo.

Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

variar con el tiempo y la dosis, depender de referencias externas puede comprometer la fiabilidad global del sistema (Townsend-Rose et al., 2024).

En este trabajo empleamos la *Standard Interface for Robotic Manipulation* (SIROM), una interfaz que proporciona acoplamiento mecánico, eléctrico y de datos entre módulos robóticos (Javier and Guerra, 2020). Su diseño modular y su orientación a operaciones robustas en entornos hostiles la hacen adecuada para escenarios radiológicos, donde la fiabilidad mecánica, eléctrica y de comunicación resulta crítica para garantizar un acoplamiento seguro entre módulos.

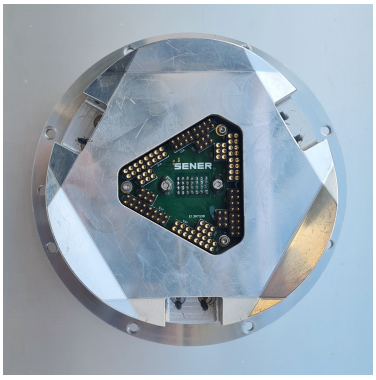

Figura 1: Ejemplo de la interfaz SIROM empleada como objetivo del sistema de estimación de pose.

En este contexto, se aborda el acoplamiento autónomo entre dos interfaces SIROM mediante la estimación en tiempo real de su pose relativa de 6 GDL. A partir de imágenes monocromas capturadas por una configuración rígida de doble cámara, el sistema infiere la pose relativa del SIROM pasivo respecto al SIROM activo en el marco del TCP, y la emplea para guiar las fases de aproximación, alineado y acoplamiento. La propuesta se basa en una arquitectura ligera de aprendizaje profundo reforzada con técnicas de aumento orientadas a mantener robustez frente a degradaciones compatibles con efectos radiológicos.

2. Estado del arte

La estimación de pose 6 GDL ha evolucionado desde métodos geométricos basados en correspondencias o plantillas, que funcionan en escenarios controlados pero pierden fiabilidad ante oclusiones, variaciones de iluminación o baja textura. Estas limitaciones han impulsado la transición hacia modelos de aprendizaje profundo capaces de inferir la pose directamente a partir de imágenes, mejorando la robustez y la generalización.

El aprendizaje profundo ha impulsado la estimación de pose basada únicamente en información RGB. Métodos de regresión directa como PoseCNN (Xiang et al., 2018) permiten predecir la pose a partir de una única imagen, aunque requieren entrenamiento específico por objeto. Otras aproximaciones detectan indicios 2D y resuelven posteriormente el problema geométrico (PnP), como SSD-6D (Kehl et al., 2017), BB8 (Rad and Lepetit, 2018), Pix2Pose (Park et al., 2019) o PVNet (Peng et al., 2018). Estos métodos alcanzan alta precisión, pero dependen de canalizaciones multietapa y del conocimiento explícito del modelo CAD del objeto. Modelos de una sola etapa como EfficientPose (Bukschat and Vetter, 2020) optimizan la eficiencia computacional, mientras que métodos RGB-D como DenseFusion (Wang et al., 2019) incrementan robustez a costa de sensores adicionales. Más recientemente, modelos con Transformers como FoundationPose (Wen et al., 2024) aportan generalización, aunque con mayor coste computacional y dependencia de datos sintéticos.

En aplicaciones autónomas, la mayoría de sistemas sigue apoyándose en marcadores fiduciarios o refinamientos geométricos para garantizar estabilidad (Vela et al., 2022; Han and Liu, 2025). Sin embargo, las soluciones *marker-free* basadas exclusivamente en imágenes monocromas y diseñadas para condiciones visuales adversas (incluyendo degradación por radiación) son menos frecuentes. En este trabajo proponemos un sistema libre de marcadores y basado en una arquitectura CNN–Transformer de doble cámara operando en escala de grises. La integración explícita de la simetría del objeto en la función de pérdida, junto con mecanismos de suavizado temporal, permite obtener estimaciones de pose estables incluso bajo perturbaciones de imagen.

3. Sistema propuesto

3.1. Adquisición y configuración sensorial

El montaje experimental emplea un manipulador UR10e y un conjunto de dos cámaras monocromas, cada una equipada con una lente de 8 mm. En esta configuración, el SIROM activo se integra en el efector final del UR10e, mientras que el SIROM pasivo se sitúa en el entorno dentro de un espacio de trabajo acotado.

Ambas cámaras se disponen de forma rígida para que sus campos de visión (FOV) presenten un solapamiento controlado, maximizando el FOV combinado y asegurando que la región central próxima al objetivo sea observada simultáneamente desde ambas perspectivas. Esta configuración proporciona redundancia visual y permite capturar variaciones locales sin pérdida de información crítica.

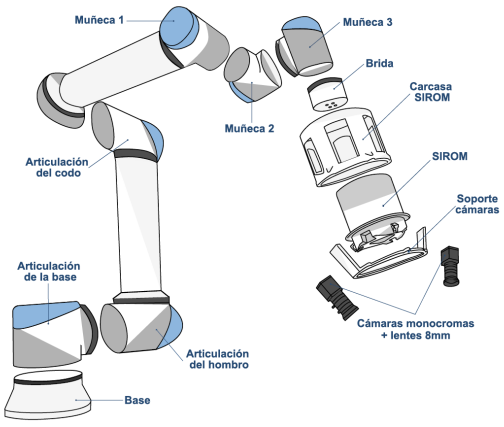

Figura 2: Montaje experimental: manipulador UR10e con SIROM activo y conjunto de sensorización de doble cámara (carcasa, soporte y cámaras) integrado en el extremo del robot.

Para minimizar la sensibilidad frente a variaciones de iluminación, se habilitan el control automático de ganancia y la regulación automática de brillo, manteniendo condiciones de exposición consistentes a lo largo de la adquisición sin ajustes manuales.

3.2. *Generación del conjunto de datos*

La recopilación del conjunto de datos se realizó mediante el manipulador UR10e para capturar variabilidad espacial suficiente para el entrenamiento. En total, se obtuvieron 20 000 muestras empleando un objetivo pasivo, cuya posición se modificó aleatoriamente dentro de un espacio de trabajo acotado.

Durante cada ciclo de captura, el robot ejecutó trayectorias parabólico-helicoidales diseñadas para obtener vistas diversificadas desde múltiples alturas, orientaciones y distancias. Las poses relativas de referencia se obtuvieron directamente a partir de la cinemática del robot.

Para evitar sesgos de configuración y promover heterogeneidad, se introdujo aleatorización en distintos niveles: posición en la mesa, rotación en el plano, origen y altura de las trayectorias, número de vueltas helicoidales y puntos de muestreo por vuelta. El proceso fue completamente automático: el sistema reposicionaba el objetivo, ejecutaba el movimiento y activaba la captura sincronizada de ambas cámaras.

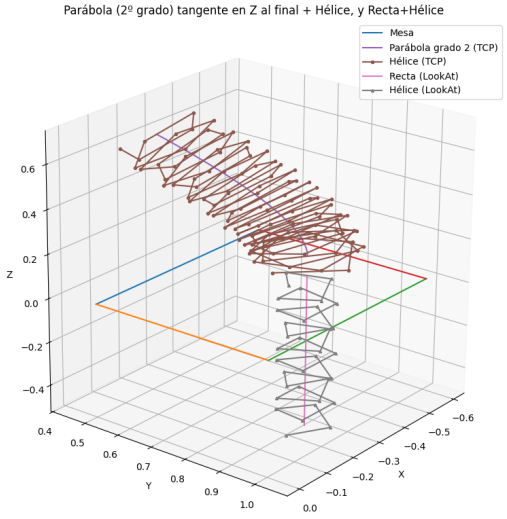

Figura 3: Ejemplo de trayectorias helicoidales y parabólicas empleadas para capturar múltiples puntos de vista durante la adquisición del conjunto de datos.

La diversidad visual se incrementó mediante la colocación de objetos aleatorios y patrones impresos en el entorno de trabajo. Asimismo, se añadieron muestras negativas en las que el objetivo estaba ausente o completamente ocluido, con el fin de robustecer el acoplamiento ante oclusiones o pérdida de visión.

3.3. *Preprocesado y aumento de datos*

Antes del entrenamiento, todas las imágenes se sometieron a un preprocesado uniforme. Cada par se redimensionó a una resolución fija de 512×612 píxeles y se normalizó, buscando un equilibrio entre coste computacional y preservación del detalle geométrico necesario para la estimación de pose.

En escenarios radiológicos, parte de la degradación puede mitigarse mediante un módulo de restauración (Sección 3.4) que estima el ruido presente y reconstruye la imagen por sustracción. En nuestra canalización, dicho módulo se considera un preprocesado opcional: puede aplicarse a las imágenes de entrada cuando se detectan artefactos compatibles con radiación o cuando se desea maximizar la calidad visual previa a la estimación de pose.

Adicionalmente, para mejorar la capacidad de generalización del estimador de pose frente a variabilidad realista (incluyendo degradación residual, cambios de iluminación u oclusiones), se aplicaron aumentos de apariencia durante el entrenamiento. Estos aumentos incluyen desenfoques, oclusiones sintéticas, *dropout* espacial, perturbaciones morfológicas y sombreado artificial. El objetivo no es "añadir radiación" de forma exacta, sino endurecer el modelo ante perturbaciones plausibles y evitar dependencia de condiciones ideales o de una restauración perfecta. No se emplearon transformaciones geométricas (rotaciones o escalados) para mantener la correspondencia exacta entre coordenadas de imagen y poses de referencia obtenidas por cinemática.

3.4. *Mitigación de artefactos inducidos por radiación*

Para abordar degradaciones severas compatibles con radiación, se dispone de un módulo de restauración basado en una arquitectura tipo U-Net (Figura 4). La red sigue la estructura clásica *encoder–bottleneck–decoder*: el *encoder* está compuesto por cuatro bloques convolucionales que extraen características mientras reducen resolución; el *bottleneck* emplea un Transformer para modelar relaciones globales entre características; y el *decoder* reconstruye una estimación del ruido apoyándose en *skip connections* para recuperar detalle espacial. La capa de salida del *decoder* utiliza activación tangencial para restringir el rango a $[-1, 1]$.

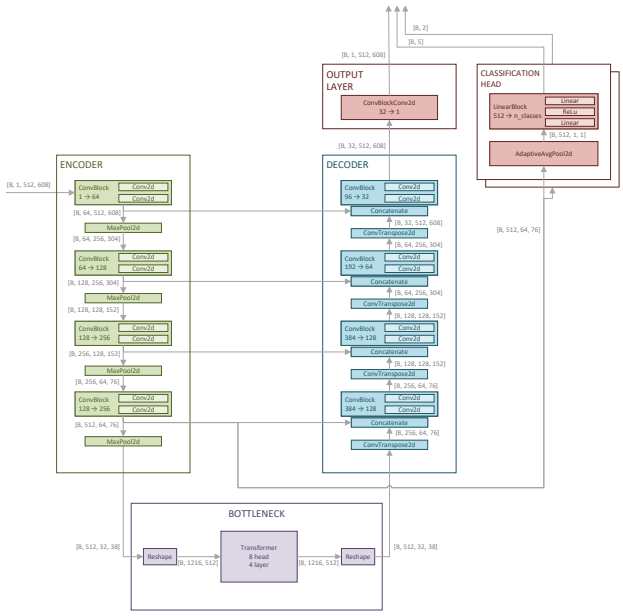

Figura 4: Diagrama del módulo de restauración de imagen frente a artefactos compatibles con radiación: *encoder* convolucional, *bottleneck* con Transformer, *decoder* con *skip connections* y cabezas auxiliares de clasificación.

Además de la salida densa de ruido, se incorporan dos cabezas auxiliares de clasificación: (i) identificación del tipo de degradación estructurada predominante y (ii) detección de *hot clusters* asociados a eventos de tipo SET. La reconstrucción final se obtiene restando el ruido estimado a la imagen degradada. La Figura 5 muestra ejemplos representativos del proceso (entrada degradada, ruido estimado y reconstrucción).

La efectividad del módulo depende de que la simulación de degradaciones sea suficientemente realista y de que el conjun-

to de entrenamiento esté equilibrado. En la generación de datos se consideran efectos habituales en sensores CMOS/CCD: corriente oscura (*dark current*), ineficiencia de transferencia de carga (CTI), *single event upset* (SEU) por filas/columnas y *single event transient* (SET) en forma de *hot clusters*. Dado que SET afecta a un número reducido de píxeles y tiene contribución pequeña a la pérdida de reconstrucción, se trata de forma adicional durante el entrenamiento: se incrementa su presencia y se combina con otros efectos sin interferir significativamente en su estructura. Entre el resto de efectos, se puede sobrerepresentar CTI en fases tempranas para evitar que la red lo ignore por baja contribución relativa, y posteriormente reequilibrar para eliminar sesgos en clasificación. El *random telegraph noise* (RTN) no se incluye como degradación independiente, al presentar una manifestación visual cercana a patrones tipo *hot pixels* y ruido impulsivo, ya contemplados en la simulación y el aumento de datos.

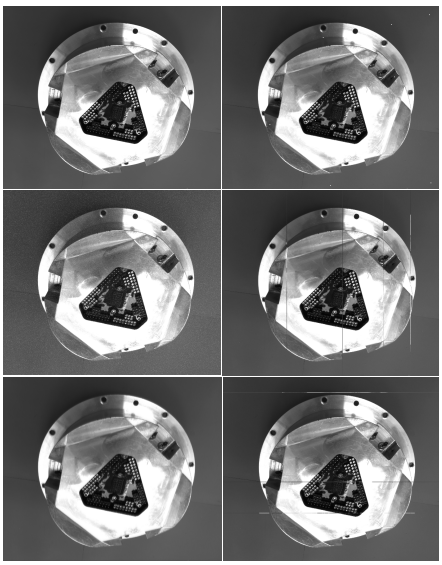

Figura 5: Ejemplo de cada tipo de efecto inducido por radiación aplicado a una imagen. De izquierda a derecha: fila superior: original, SET y *dark current*; fila inferior: CTI, SEU (columnas) y SEU (filas).

Este módulo se integra como preprocesado opcional del sistema: puede utilizarse para mejorar la calidad de imagen antes de la estimación de pose, mientras que el estimador principal se entrena con aumentos de apariencia para mantener robustez incluso ante degradación residual o discrepancias entre simulación y condiciones reales.

3.5. Arquitectura del estimador de pose

El estimador recibe como entrada un par de imágenes monocromas sincronizadas adquiridas por dos cámaras rígidamente montadas, y predice la pose completa del objetivo: traslación cartesiana (x, y, z), orientación como cuaternión unitario, y una medida escalar de visibilidad $s \in [0, 1]$, que indica si el SIROM pasivo es observable. Las imágenes de entrada pueden ser las capturas directas o, si se habilita, las imágenes restauradas por el módulo de mitigación de radiación (Sección 3.4).

Cada imagen se procesa mediante un *backbone* convolucional ConvNeXtV2-Atto para extraer representaciones jerárquicas. Dichas representaciones se aplanan y reformulan como secuencias de tokens enriquecidas con codificaciones posicionales bidimensionales y *embeddings* que distinguen el origen de cada token. La Figura 6 ilustra la etapa de extracción convolucional y la conversión a tokens.

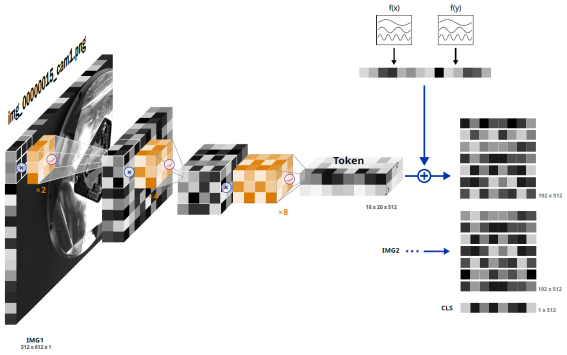

Figura 6: Extracción de características por vista mediante el *backbone* convolucional (ConvNeXtV2-Atto) y formación de tokens con codificación posicional.

La fusión multivista se realiza mediante un codificador Transformer ligero que modela dependencias de largo alcance dentro y entre vistas. La salida se condensa en un token de clasificación (CLS) que actúa como representación global. La Figura 7 muestra el bloque de fusión basado en atención.

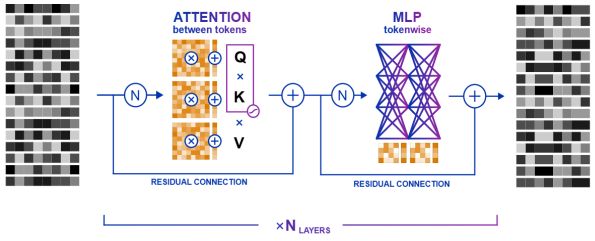

Figura 7: Fusión multivista mediante un codificador Transformer ligero con *self-attention* entre vistas.

A partir de la representación global, tres cabezas MLP independientes generan las salidas de traslación, orientación y visibilidad. La orientación se expresa mediante un cuaternión unitario, evitando singularidades de ángulos de Euler y discontinuidades de otros esquemas de rotación. Las tres cabezas se ilustran en la Figura 8.

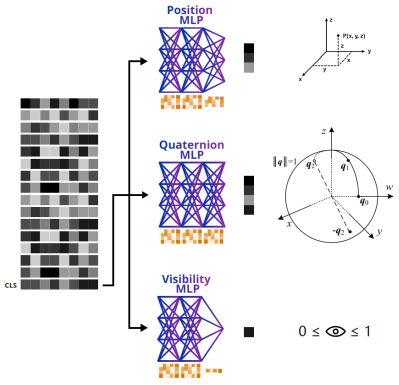

Figura 8: Cabezas de predicción del estimador: regresión de traslación, regresión de orientación (cuaternión) y estimación de visibilidad.

3.6. *Entrenamiento*

El modelo se entrena mediante una función de pérdida compuesta que integra términos de traslación, orientación y visibilidad. El error traslacional se calcula como distancia euclídea. Para la orientación se emplea una métrica en espacio de cuaterniones con tratamiento explícito de la simetría, tomando la rotación equivalente más cercana. La visibilidad se optimiza mediante entropía cruzada binaria y los pesos relativos se ajustan para evitar dominancia de un término.

Se utiliza AdamW con un calendario de *cosine annealing* con reinicios cálidos. Los pesos iniciales de la pérdida se determinan estimando la escala esperada de cada tarea a partir de predicciones triviales, garantizando una contribución equilibrada desde el inicio.

3.7. *Inferencia y control en lazo cerrado*

Una vez entrenado, el sistema de acoplamiento opera en un esquema en lazo cerrado, incorporando las estimaciones de pose y visibilidad en cada ciclo de control. En lugar de promediar directamente poses instantáneas, se emplea estabilización basada en incrementos de movimiento, ya que el SIROM pasivo puede haber cambiado de pose y las muestras recientes pueden corresponder a estados distintos.

Las poses se transforman a un marco global y, entre instantes consecutivos, se calculan variaciones de traslación y rotación mediante diferencias de cuaternión compatibles con la simetría del modelo. Estos incrementos se suavizan con una ventana temporal corta, reduciendo ruido y preservando cambios coherentes.

Se asignan pesos temporales dependientes de la magnitud del movimiento suavizado y del nivel de visibilidad, de modo que predicciones recientes y coherentes tienen mayor influencia que las afectadas por inestabilidad o baja confianza. Con estos pesos se obtiene la pose final mediante promedios ponderados de posición y orientación. Los mismos pesos permiten calcular indicadores de estabilidad, como desviación estándar ponderada y una longitud de memoria efectiva, que gobiernan las transiciones entre fases de operación (*search*, *approach*, *align* y *dock*) mediante umbrales específicos.

3.8. *Rendimiento computacional*

Desde el punto de vista computacional, el modelo se ejecuta a aproximadamente 30 Hz en una GPU de clase escritorio, permitiendo mantener la operación en tiempo real del lazo cerrado.

4. Resultados

4.1. *Evaluación cuantitativa*

Se reservó un 10 % del conjunto de datos como partición de prueba. El error de traslación se midió con la distancia euclídea entre posición estimada y referencia, y el error de orientación mediante distancia geodésica entre cuaterniones, expresada en grados. La visibilidad se evaluó mediante exactitud de clasificación. Los resultados muestran errores bajos y estables en todo el espacio operativo, sin degradación notable ante variaciones de orientación o configuraciones no centrales, manteniéndose típicamente en el rango 5–7 mm y 3–5° bajo degradaciones de imagen significativas (Figura 10).

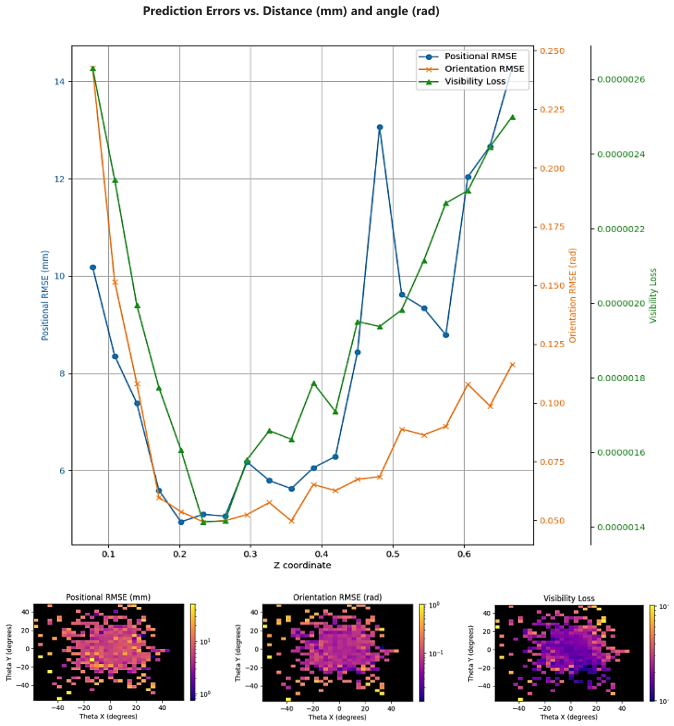

Figura 10: Análisis de error en el conjunto de prueba. **Arriba:** RMSE posicional y rotacional y pérdida de visibilidad frente a la coordenada z de referencia. **Abajo:** mapas de calor del error en función de los ángulos θ_x y θ_y respecto del eje z.

4.2. *Acoplamiento autónomo en el sistema físico*

Para validar el rendimiento del sistema completo en condiciones reales, se realizaron intentos de acoplamiento autónomo con el manipulador UR10e y el subsistema de visión integrado. El SIROM pasivo se situó aleatoriamente en diferentes posiciones dentro del espacio alcanzable y se expuso a variaciones de iluminación, oclusiones parciales y perturbaciones angulares y traslacionales durante la ejecución. En las configuraciones ensayadas, el sistema mantuvo una estimación de pose estable y permitió completar el acoplamiento incluso en presencia de reflexiones intensas, iluminación reducida o modificaciones imprevistas del movimiento, condiciones donde los métodos basados en marcadores tienden a perder fiabilidad.

4.3. *Acoplamiento en escenarios complejos*

Se realizaron demostraciones adicionales de acoplamiento en contextos más exigentes, incluyendo escenas con elevado desorden visual por objetos cercanos, movimientos externos no planificados y configuraciones elevadas o con inclinaciones pronunciadas. La Figura 9 muestra tres demostraciones representativas bajo condiciones progresivamente más exigentes. A pesar de estas condiciones, la estimación de pose se mantuvo coherente y con degradación mínima, evidenciando capacidad de operación bajo variabilidad significativa sin pérdida de estabilidad temporal.

4.4. *Rendimiento en tiempo real*

Durante las pruebas físicas, el sistema operó en tiempo real sin latencias perceptibles y con una frecuencia compatible con el lazo de control.

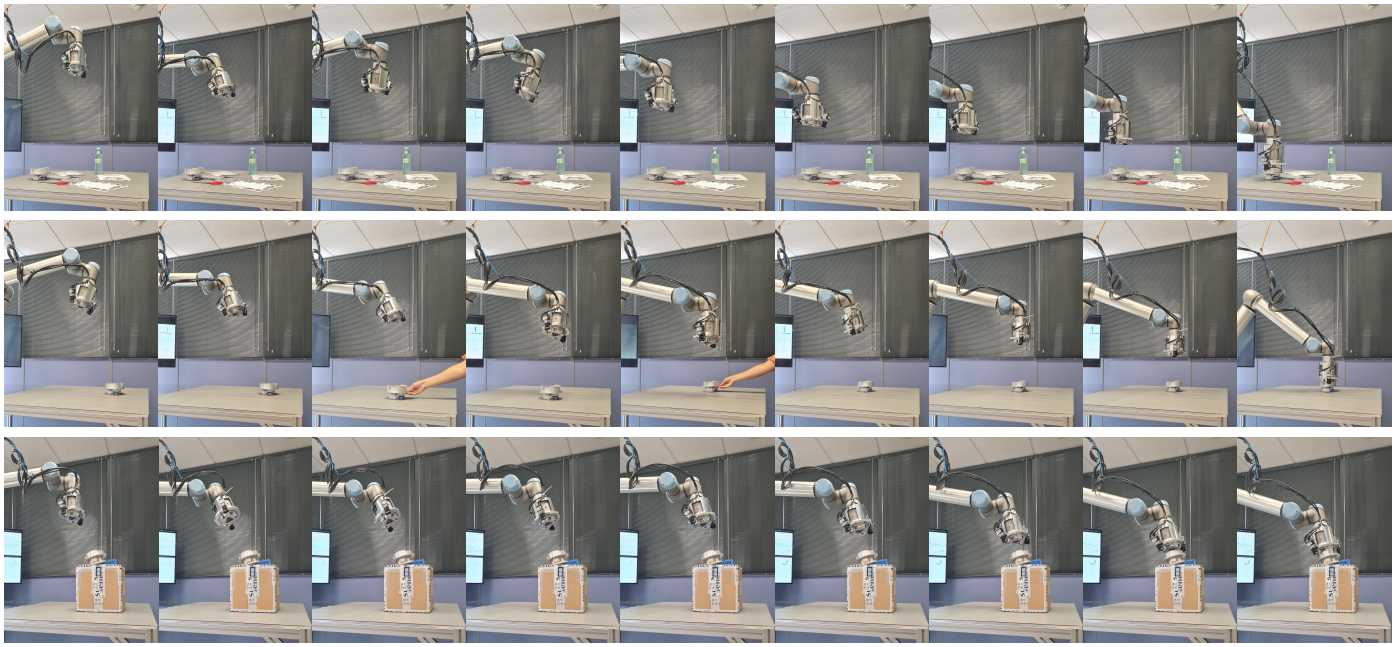

Figura 9: Demostraciones de acoplamiento autónomo. **Arriba:** SIROM pasivo colocado aleatoriamente entre objetos sobre la mesa. **Centro:** SIROM pasivo movido dinámicamente por una persona, con el robot reaccionando en tiempo real. **Abajo:** SIROM pasivo elevado e inclinado, demostrando acoplamiento bajo orientaciones exigentes. En los tres escenarios se completa el acoplamiento con éxito.

5. Conclusiones

El sistema presentado constituye un subsistema ligero y libre de marcadores para la estimación de pose 6 GDL en condiciones visuales adversas. La combinación de CNN y Transformer, junto con una función de pérdida formulada sobre cuaterniones con tratamiento explícito de simetría, permite fusión multivista robusta y resuelve ambigüedades rotacionales inherentes a geometrías simétricas. Adicionalmente, los mecanismos de suavizado temporal y ponderación adaptativa proporcionan comportamiento consistente en tiempo real incluso bajo oclusiones parciales, variaciones de iluminación y degradaciones compatibles con radiación.

Los resultados muestran que es posible obtener precisión milimétrica en traslación y errores rotacionales del orden de grados sin necesidad de marcadores, sensores de profundidad o modelos de gran tamaño. En escenarios radiológicos, donde la degradación del sensor puede ser determinante, este enfoque reduce dependencia de instrumentación externa y mejora la robustez operativa del sistema de percepción.

Agradecimientos

Los autores agradecen el apoyo de SENER Aeroespacial en el desarrollo experimental del sistema y en la preparación de este trabajo.

Referencias

Bukschat, Y., Vetter, M., 2020. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach.
URL: https://arxiv.org/abs/2011.04307

Han, Z., Liu, L., 2025. A 6d object pose estimation algorithm for autonomous docking with improved maximal cliques.
URL: https://www.mdpi.com/1424-8220/25/1/283
DOI: 10.3390/s25010283

Javier, V., Guerra, G., 10 2020. Standard interface for robotic manipulation (sirom): Src h2020 og5 final results – future upgrades and applications.

Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N., 2017. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again.
URL: https://arxiv.org/abs/1711.10006

Liu, B., Li, Y., Wen, L., Zhang, X., Guo, Q., 2023. Effects of hot pixels on pixel performance on backside illuminated complementary metal oxide semiconductor (cmos) image sensors. Sensors 23 (13).
URL: https://www.mdpi.com/1424-8220/23/13/6159
DOI: 10.3390/s23136159

Park, K., Patten, T., Vincze, M., Oct. 2019. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, p. 7667–7676.
URL: http://dx.doi.org/10.1109/ICCV.2019.00776
DOI: 10.1109/iccv.2019.00776

Peng, S., Liu, Y., Huang, Q., Bao, H., Zhou, X., 2018. Pvnet: Pixel-wise voting network for 6dof pose estimation.
URL: https://arxiv.org/abs/1812.11788

Rad, M., Lepetit, V., 2018. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth.
URL: https://arxiv.org/abs/1703.10896

Sirianni, M., Mutchler, M., 2006. Radiation damage in hst detectors. In: Beletic, J. E., Beletic, J. W., Amico, P. (Eds.), Scientific detectors for astronomy 2005. Springer Netherlands, Dordrecht, pp. 171–178.

Townsend-Rose, C., Buggey, T. W., Ivory, J. M., Dazzazi, I., Stefanov, K. D., Hall, D. J., 2024. Non-ionizing radiation effects in a soft X-ray CMOS image sensor. Journal of Astronomical Telescopes, Instruments, and Systems 10 (3), 036002.
URL: https://doi.org/10.1117/1.JATIS.10.3.036002
DOI: 10.1117/1.JATIS.10.3.036002

Vela, C., Fasano, G., Opromolla, R., 2022. Pose determination of passively cooperative spacecraft in close proximity using a monocular camera and aruco markers.
URL: https://www.sciencedirect.com/science/article/pii/S0094576522004283
DOI: https://doi.org/10.1016/j.actaastro.2022.08.024

Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S., 2019. Densefusion: 6d object pose estimation by iterative dense fusion.
URL: https://arxiv.org/abs/1901.04780

Wen, B., Yang, W., Kautz, J., Birchfield, S., 2024. Foundationpose: Unified 6d pose estimation and tracking of novel objects.
URL: https://arxiv.org/abs/2312.08344

Xiang, Y., Schmidt, T., Narayanan, V., Fox, D., 2018. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes.
URL: https://arxiv.org/abs/1711.00199