Gilberto Ochoa Ruiz, Ricardo Espinosa Loera y Christian Daul

Un viaje fantástico: el papel de la visión computacional para el diagnóstico médico

La endoscopia digital es clave en diversas examinaciones médicas. Sin embargo, la adopción de técnicas de IA en endoscopia para facilitar las tareas de diagnóstico está aún en su infancia. El principal reto es mejorar la robustez de los métodos de visión computacional ante los cambios de iluminación. Dichas mejoras son esenciales para aplicaciones de cartografía 3D. En este artículo discutimos el desarrollo de nuevos métodos capaces de lidiar con este tipo de artefactos en imágenes endoscópicas.

Introducción y contexto médico

Las intervenciones endoscópicas son la referencia para explorar órganos internos como el estómago y el colon (véase la **Figura 1a**). Estas examinaciones son el único mecanismo para analizar características anatómicas (es decir, color, textura y forma) en las paredes epiteliales de dichos órganos ("Endoscopia gastrointestinal", 2018). La información extraída es esencial para diversas tareas médicas; por ejemplo, en la detección y caracterización de lesiones (inflamatorias, precancerosas o cancerosas), así como en el seguimiento de estas lesiones. En la endoscopia, la punta de la cámara está muy cerca del tejido, lo que permite adquirir imágenes de muy alta resolución. Sin embargo, esta forma de visualizar las zonas de interés tiene serias desventajas:

1. Debido al limitado campo de vista provisto por el endoscopio, las lesiones no son observadas en su totalidad en una sola toma (Sánchez-Montes y cols., 2020). Además, el médico debe observar la zona de interés a través de una pantalla, lo cual dificulta el procedimiento (**Figura 1b**).
2. El campo de vista restringido es un obstáculo para conducir la inspección endoscópica de forma cómoda para el médico, quien se ve obligado a regresar de forma constante a puntos de referencia anatómicos para reconstruir mentalmente la forma del órgano en 3D y localizar las lesiones. Además, el endoscopista no puede saber si ya inspeccionó la totalidad de las zonas de interés y, aún

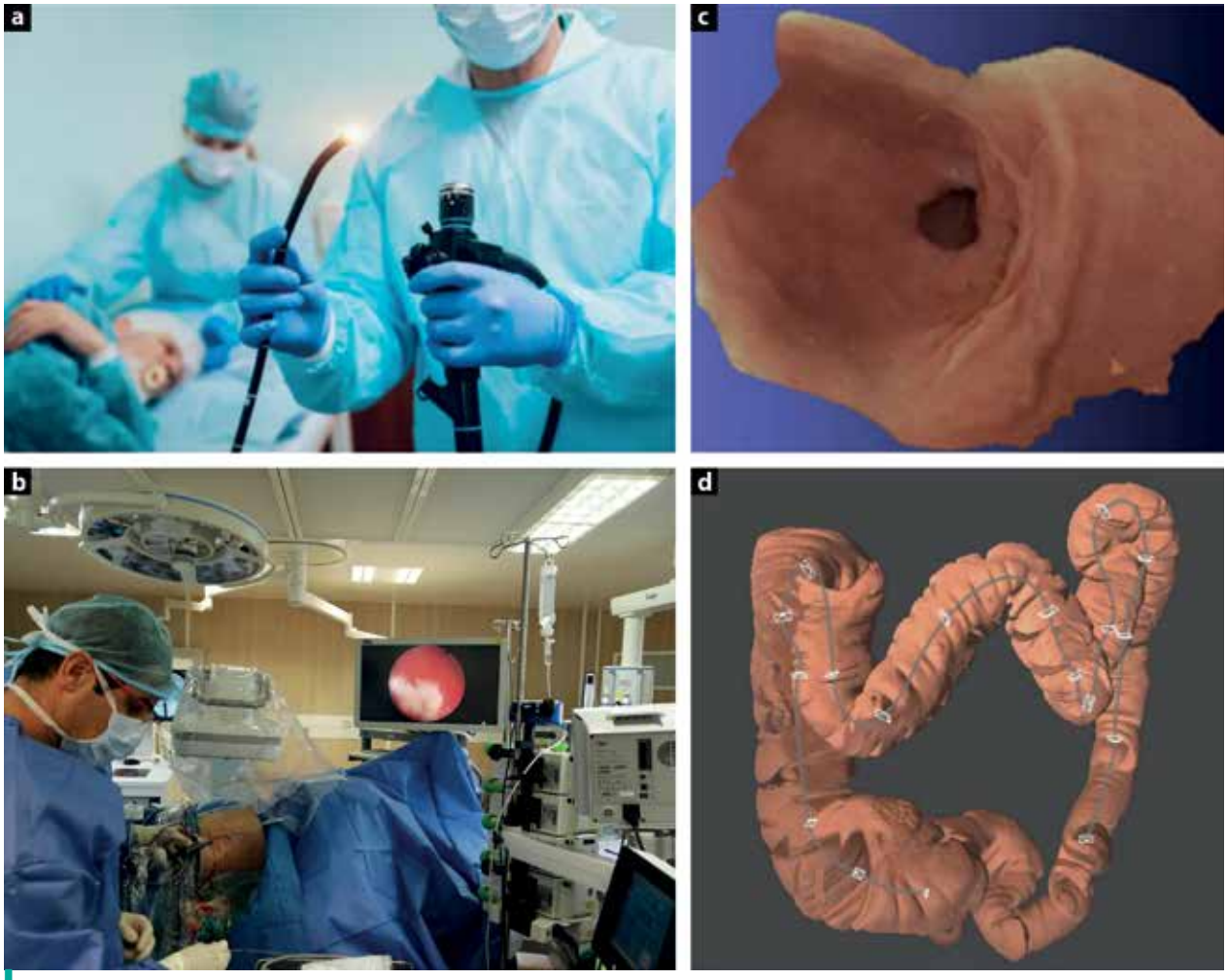

Figura 1. *a)* El endoscopio es un instrumento fundamental para realizar examinaciones de diversas cavidades huecas, de otra forma inaccesibles. *b)* Sin embargo, el campo de vista limitado y la interacción a través de una pantalla dificulta las inspecciones endoscópicas y el seguimiento de las posibles lesiones. Estos problemas pueden paliarse a través del uso de: *c)* panoramas anatómicos digitales, o *d)* la construcción de mapas 3D digitales e interactivos de las zonas de interés.

Métodos de cartografía

Se usan en visión computacional para producir mosaicos de imágenes. Un mosaico de fotografías es una imagen compuesta creada uniendo una serie de imágenes contiguas (aéreas en el caso de imágenes de drones, por ejemplo). Otras aplicaciones incluyen la creación de mapas en robótica y de imágenes de vista amplia en medicina.

peor, el hecho de simplemente desplegar el video en una pantalla no permite guiar las examinaciones de forma efectiva.

3. Adicionalmente, el video resultante es raramente grabado. Por ende, el médico no cuenta con ningún mecanismo para llevar a cabo un segundo diagnóstico o para que otros especialistas realicen una evaluación colegiada; además de que los diversos especialistas involucrados en tratamientos de cáncer de colon o estómago (endoscopistas, oncólogos y radiólogos, entre otros) no cuentan con ningún medio de discusión común para llevar a cabo sus actividades diarias e intercambiar información valiosa (es decir, soportes digitales).

En un proyecto financiado por un fondo ECOS Nord Conahcyt-Gobierno de Francia, llamado ML-inside, buscamos desarrollar nuevas herramientas de IA para mejorar las capacidades diagnósticas de la endoscopia. En particular, estamos interesados en el desarrollo de **métodos de cartografía** que permitan mejorar de forma significativa la extensión del campo de vista de escenas endoscópicas. Como se puede observar en la parte inferior de la **Figura 1**, la creación de mosaicos y mapas 3D digitales daría cabida a una visualización más completa y detallada de las lesiones, facilitando el diagnóstico durante examinaciones endoscópicas, así como tareas de diagnóstico posprocedimentales.

La comparación de estos mapas construidos a partir del mismo paciente durante examinaciones distintas permitiría determinar la evolución de una lesión o detectar la recurrencia de un cáncer (el riesgo de cáncer de uretra, colon o estómago es muy alto y requiere de examinaciones de control regulares). Adicionalmente, estos mapas anatómicos digitales pueden ser archivados, promoviendo la trazabilidad de lesiones sospechosas. Aún más, dichos mapas pueden servir como mecanismos de soporte e intercambio entre diferentes especialistas para establecer tratamientos mejor coordinados.

Estos mapas 3D pueden servir además como un mecanismo de documentación, dado que las lesiones detectadas pueden ser resaltadas en los mapas mejorados, y métodos de explicabilidad pueden servir como un complemento diagnóstico. Nuestro proyecto busca desarrollar nuevos métodos de frontera en el área de visión computacional, automatizando tareas de análisis de video endoscópico. Sin embargo, dichas herramientas pueden ser usadas como soporte en aplicaciones como la asistencia procedimental (cirugía integrada por computadora, realidad aumentada, detección y seguimiento de instrumentos). En este artículo, nos enfocamos en la creación de mapas 3D a partir de imágenes monoculares, una activa área de investigación en diversos campos, como la robótica y la conducción autónoma, pero de reciente introducción en el área médica.

La reconstrucción 3D y los retos en endoscopia

La reconstrucción de estructuras 3D a partir de videos monoculares es un tema de amplio interés en diversas áreas de la visión artificial, la realidad aumentada y la robótica (Eddie Edwards y cols., 2022). La configuración tradicional de un sistema de reconstrucción 3D haciendo uso de imágenes monoculares se muestra en la **Figura 2a**. El uso de una sola cámara es indispensable en endoscopia, donde no es posible instrumentar el sistema de adquisición para obtener información 3D.

Los métodos de reconstrucción 3D hacen uso de mapas de profundidad como un paso necesario para obtener información geométrica de la escena

(véase la **Figura 2a**). Los métodos estéreo multivista tradicionales –como la estructura a partir del movimiento (Structure from Motion o SfM, por sus siglas en inglés) y la localización y mapeo simultáneos (Simultaneous Localization and Mapping o SLAM)– son capaces de reconstruir estructuras 3D en escenas regulares con iluminación constante entre fotogramas. Sin embargo, las superficies endoscópicas son sumamente complejas debido a la falta de textura (García-Vega y cols., 2022), lo que hace que los métodos clásicos resulten insuficientes. Además, los cambios repentinos de iluminación presentes en los exámenes endoscópicos (véase la **Figura 2b**) conllevan a la generación de mapas de profundidad poco fidedignos, produciendo mapas 3D también deficientes y con zonas huecas (**Figura 2c**).

Recientemente, métodos de estimación de profundidad monocular basados en aprendizaje supervisado (**Figura 2a**) han sido ampliamente investigados en el contexto de la endoscopia. Estos métodos heredan las ventajas clave de los métodos SfM y SLAM convencionales, evitando los problemas que afectan a los métodos tradicionales (es decir, susceptibilidad a regiones lisas o de poca textura). Por ejemplo, varios

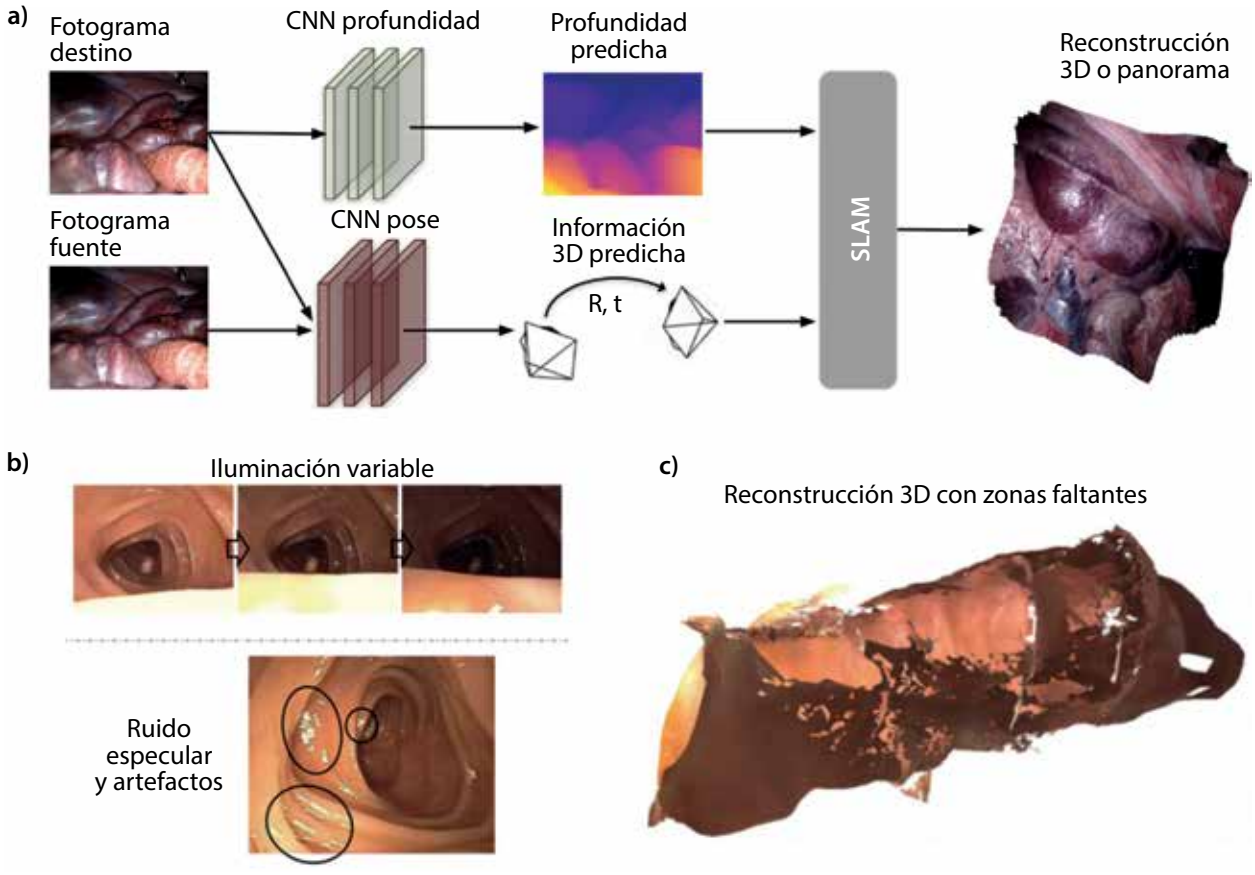

Figura 2. *a)* Panoramas o mapas 3D pueden ser creados entrenando modelos de IA que aprenden un mapeo entre fotogramas continuos, mapas de profundidad y la correspondiente información geométrica. *b)* Sin embargo, dichos modelos tienen serios problemas con cambios de iluminación súbitos y artefactos presentes en imágenes endoscópicas. *c)* Esto produce mapas de profundidad subóptimos y, por ende, reconstrucciones 3D incorrectas.

trabajos han aplicado redes neuronales y han logrado resultados sobresalientes en imágenes naturales. Sin embargo, estos métodos dependen de la disponibilidad de datos reales adquiridos con sensores avanzados para capturar datos en entornos reales (**verdad fundamental**, o *ground truth* en inglés). En marcado contraste, en endoscopia, este enfoque es inviable debido a los desafiantes escenarios reflectantes dentro de las cavidades humanas y dada la dificultad de instrumentar el endoscopio con sensores adicionales en condiciones realistas (Aguilera-Chuchuca y cols., 2022).

Verdad fundamental
Término usado en aprendizaje automático y visión artificial, entre otras áreas; es información que se sabe que es real o verdadera, proporcionada por la observación y medición directa, a diferencia de la información proporcionada por inferencia.

En nuestra investigación hacemos uso de técnicas de aprendizaje no supervisadas para abordar la falta de verdad de terreno necesaria para entrenar a las redes neuronales usadas en reconstrucción 3D. En nuestro método, la profundidad y la pose son predichas por una red neuronal y se utilizan para inferir el fotograma siguiente (destino) a partir del fotograma fuente, deformando el fotograma de origen mediante el uso de una homografía. La diferencia perceptual entre el cuadro proyectado y el cuadro fuente es, por lo tanto, el principal objetivo de aprendizaje.

El principio fundamental para la autosupervisión en estos métodos es el supuesto de constancia del brillo, que supone que la intensidad del brillo entre fotogramas es constante o sin cambios repentinos. Sin embargo, en endoscopia, los métodos descritos anteriormente enfrentan desafíos específicos, pues la iluminación de la escena depende en gran medida de la orientación de la cámara en relación con la superficie del tejido, como se discutió antes y se muestra en la **Figura 2b**. Las imágenes recopiladas suelen estar subexpuestas o sobreexpuestas, según la forma de la superficie, o afectadas por reflejos especulares u otros artefactos, lo que impacta en la visibilidad de las lesiones, así como en el desempeño de los métodos

de reconstrucción 3D (véase la Figura 2c). En este sentido, nuestro trabajo busca desarrollar métodos menos susceptibles a dichos cambios fotométricos, como se detalla a continuación.

Método propuesto

La estimación de la profundidad monocular autosupervisada en imágenes endoscópicas es un desafío debido a las superficies de baja textura y las difíciles condiciones de iluminación. Estos factores hacen complejo el entrenamiento de los modelos, pues la iluminación irregular en el video de entrada genera problemas de convergencia o produce predicciones de profundidad inexactas. Estas variaciones afectan también los resultados durante su uso en tiempo real.

Diversos trabajos han empleado mecanismos de ajuste de iluminación lineales y no lineales. Aunque estos métodos han tenido resultados prometedores, estas estrategias se basan en suposiciones específicas sobre las variaciones de apariencia subyacentes al modelo, así como el movimiento que se puede encontrar entre fotogramas. Sin embargo, las imágenes endoscópicas presentan características complejas que son difíciles de capturar completamente en un modelo. Con el fin de mitigar este problema, en nuestro método introducimos una nueva función de pérdida, haciendo uso de descriptores invariantes a la iluminación, sin modificar el esquema prototípico de estimación de profundidad. Nuestro enfoque está resumido de forma conceptual en la Figura 3.

Esta novedosa formulación está inspirada en los descriptores de vecindario (ND), que han demostrado su eficacia para el cálculo del flujo óptico bajo grandes cambios de iluminación y en escenarios de baja textura. Con el fin de robustecer la señal de autosupervisión contra cambios de iluminación, extendemos la ampliamente usada pérdida de similitud estructural (Structural Similarity Index Measure o SSIM) agregando la diferencia entre las características invariantes de iluminación extraídas de la imagen de origen y el objetivo como señal secundaria. En segunda instancia, hacemos uso de una transformación neural de intensidad a nivel pixel para lidiar mejor con zonas de poca textura, que producen predicciones borrosas en zonas con cambios abruptos de gradiente.

Nuestro método rectifica la pérdida de calidad en la predicción de profundidad, lo que le permite guiar eficazmente el entrenamiento en las difíciles condiciones de los videos endoscópicos, como puede observarse en la Figura 4, y así producir mejores resultados en la predicción de profundidad, a pesar de múltiples artefactos (fila 1, cambios por movimiento

Artefactos
Se dice de errores o alteraciones engañosas o confusas en los datos o la observación; en el contexto de la visión computacional para procesamiento de imágenes endoscópicas, alude a alteraciones en la imagen digital producidas por anomalías debidas al sistema de adquisición, cambios de iluminación y otros procesos físicos.

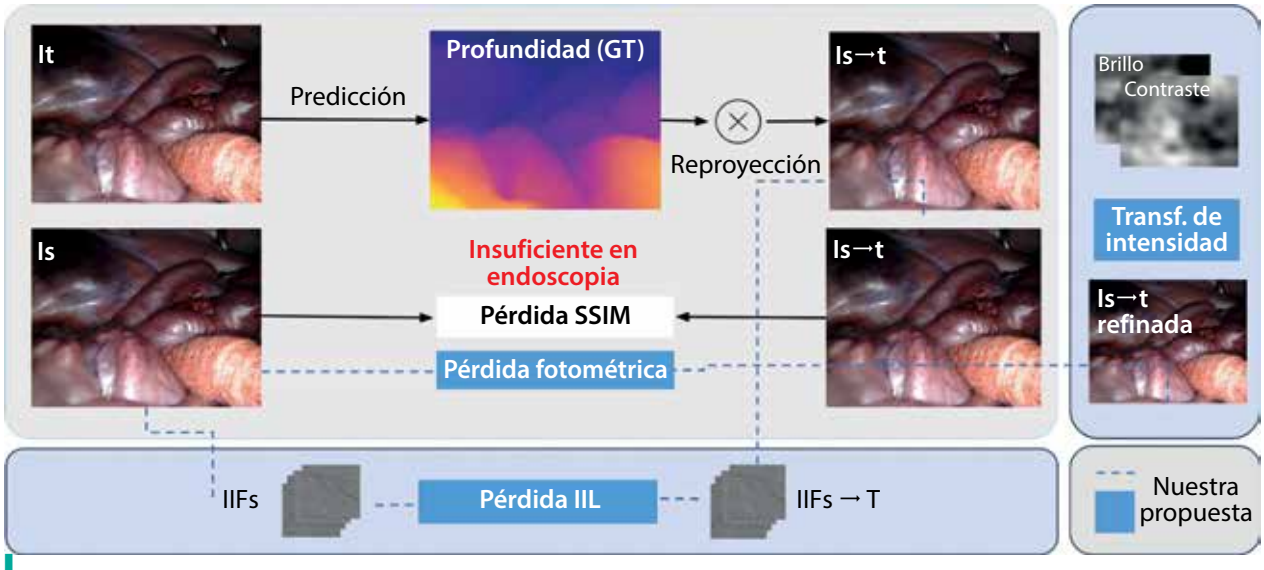

Figura 3. El método propuesto está basado en un esquema de aprendizaje profundo autosupervisado, en el que señales de entrenamiento auxiliares a la pérdida clásica (SSIM) guían a la red a predecir la profundidad de forma robusta, invariante a la iluminación.

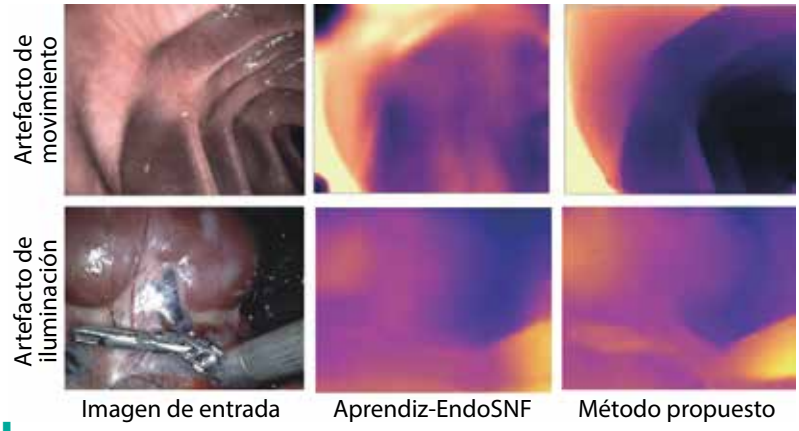

Figura 4. El método propuesto es capaz de producir mapas de profundidad fidedignos (última columna), aun en la presencia de artefactos por movimiento en la imagen de entrada (primera fila), o en el caso de presencia de cambios de iluminación o reflejos especulares, en comparación con los resultados obtenidos por un método del estado del arte (columna central).

súbitos, fila 2, artefactos de iluminación) comparados con otros métodos del estado del arte.

Nuestro modelo consta de módulos de estimación de profundidad, de movimiento y de calibración de cambio de iluminación. Para mejorar el módulo de predicción de profundidad, empleamos una arquitectura que combina información local y global mediante la combinación de bloques CNN y Transformer. El módulo de calibración de cambio de iluminación ajusta la intensidad del brillo desde el cuadro fuente compensando los cambios de iluminación a nivel pixel.

Resultados

Nuestros experimentos en diferentes conjuntos de datos de referencia muestran que las funciones de pérdida propuestas son capaces de lidiar con cambios de iluminación y producir mapas de profundidad fidedignos (véase la **Figura 4**). Para evaluar la calidad de los mapas de profundidad utilizamos métricas de fidelidad de señales e imágenes (el SSIM, o el radio de señal a ruido: PSNR) que miden la similitud entre el mapa de profundidad predicho contra una imagen de verdad de terreno, buscando minimizar distorsiones perceptuales.

Como se puede observar en las **Figuras 5a** y **5b,** el mejoramiento del mapa de profundidad conlleva a

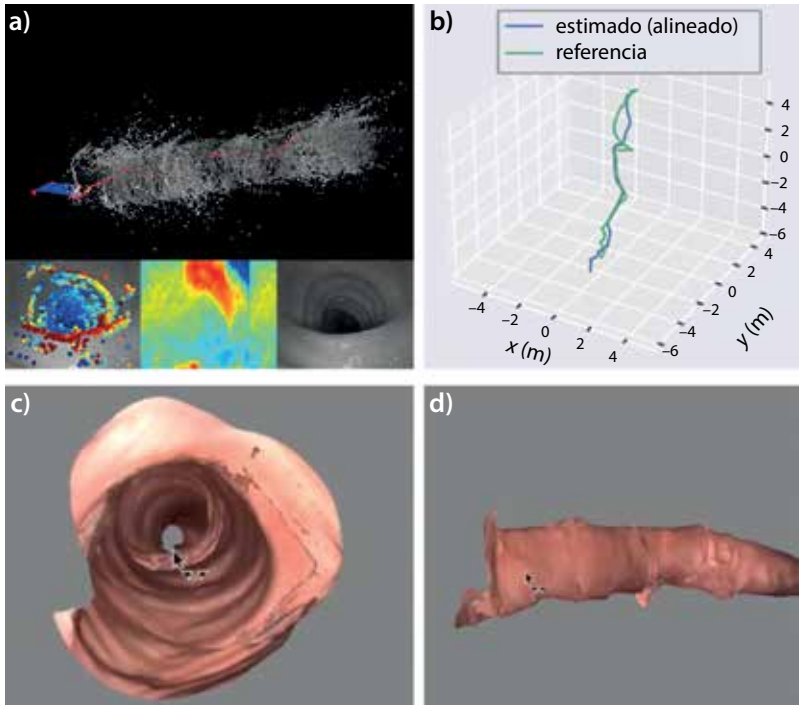

Figura 5. El método propuesto fue integrado en un esquema de reconstrucción 3D usando SLAM para endoscopia. Los resultados de este ejemplo de un segmento de colon muestran que el método es capaz de producir: *a)* una mejor estimación de la nube de puntos, y *b)* un mejor desempeño en el proceso de rastreo de puntos de interés, así como *c)* reconstrucciones 3D en tiempo real, y *d)* mapas 3D con menos valores atípicos y zonas faltantes debido a los cambios de iluminación.

una menor dispersión en las nubes de puntos estimados por los métodos de SLAM (producidos durante el proceso de reconstrucción 3D), así como a un mejor rastreo de la trayectoria con respecto a una referencia base, obtenido usando SFM (**Figura 5b**), lo cual permite generar mejores mapas tridimensionales en tiempo real (**Figura 5c**) que contienen menos huecos (**Figura 5d**).

Gilberto Ochoa Ruiz
Tecnológico de Monterrey, Guadalajara, México.
gilberto.ochoa@tec.mx

Ricardo Espinosa Loera
Universidad Panamericana, México; Université de Lorraine, Francia.
respinosa@up.edu.mx

Christian Daul
Université de Lorraine, Centre de Recherche en Automatique de Nancy, Francia.
christian.daul@u-lorraine.fr

Lecturas recomendadas

Aguilera-Chuchuca, M. J., S. A. Sánchez-Luna, B. González Suárez, K. Ernest-Suárez, A. Gelrud y T. M. Berzin (2022), "El papel emergente de la inteligencia artificial en la endoscopia gastrointestinal: una revisión de la literatura", *Gastroenterología y Hepatología*, 45(6):492-497.

Eddie Edwards, P. J., D. Psychogyios, S. Speidel, L. Maier-Hein y D. Stoyanov (2022), "SERV-CT: A Disparity Dataset from Cone-Beam CT for Validation of Endoscopic 3D Reconstruction", *Medical Image Analysis*, 76. Disponible en: ‹https://doi.org/10.1016/j.media.2021.102302›, consultado el 18 de enero de 2025.

"Endoscopia gastrointestinal" (2018), *Revista de Gastroenterología de México*, 83:90-92. Disponible en: ‹https://www.revistagastroenterologiamexico.org/es-endoscopia-gastrointestinal-articulo-X037509061863283X›, consultado el 18 de enero de 2025.

García-Vega, A., R. Espinosa, G. Ochoa-Ruiz, T. Bazin, L. Falcón-Morales, D. Lamarque y C. Daul (2022), "A Novel Hybrid Endoscopic Dataset for Evaluating Machine Learning-Based Photometric Image Enhancement Models", *Advances in Computational Intelligence*, pp. 267-281.

Sánchez-Montes, C., A. García-Rodríguez, H. Córdova, M. Pellisé y G. Fernández-Esparrach (2020), "Tecnologías de endoscopia avanzada para mejorar la detección y caracterización de los pólipos colorrectales", *Gastroenterología y Hepatología*, 43(1):46-56.