CEA Comité Español de Automática
XLVII JORNADAS DE AUTOMÁTICA CÓRDOBA 2026
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Segmentación semántica basada en ResNet embarcada en vehículo robótico para búsqueda y rescate

Lucas Jurado-Martínez, Ricardo Vázquez-Martín*, Anthony Mandow

Grupo de Robótica y Mecatrónica, Instituto Universitario de Investigación en Ingeniería Mecatrónica y Sistemas Ciberfísicos (IMECH.UMA), Universidad de Málaga, Arquitecto Francisco Peñalosa, n° 6, 29071, Málaga, España.

To cite this article: Jurado-Martínez, L., Vázquez-Martín, R., Mandow, A. 2026. Embedded ResNet-Based Semantic Segmentation in a robotic vehicle for Search and Rescue. Jornadas de Automática, 47. https://doi.org/10.17979/ja-cea.2026.47.13810

Resumen

La percepción visual en robótica de búsqueda y rescate debe interpretar escenas no estructuradas, con iluminación variable y, además, mantener la ejecución en tiempo real en hardware embarcado. En este trabajo se presenta la implantación de SARNet en el robot móvil terrestre ARGO J8, una red de segmentación semántica basada en mecanismos de atención y orientada a escenarios de catástrofe. Para ello, se adapta la red incluyendo el reentrenamiento con sobremuestreo para equilibrar las clases críticas poco representadas y la integración del modelo en ROS 2 en una NVIDIA Jetson AGX Orin. El sistema captura imágenes RGB de una cámara ZED 2i, genera máscaras semánticas multiclase, estima la distancia y la orientación de las personas detectadas y muestra la información en una interfaz visual de apoyo al operador a 16,4 FPS, aunque puede superar los 30 FPS cuando se elimina la limitación del sensor de entrada y el postprocesado. El código está disponible en el repositorio `https://github.com/lucasjuradomartinez-cyber/TFGLucas_ws`.

Palabras clave: Percepción y detección, Tecnología robótica, Robótica de campo, Robots móviles autónomos, Aprendizaje automático, Sistemas robóticos autónomos

Embedded ResNet-Based Semantic Segmentation in a robotic vehicle for Search and Rescue

Abstract

Visual perception in search-and-rescue robotics must interpret unstructured scenes under varying lighting conditions while maintaining real-time performance on embedded hardware. This work presents the deployment of SARNet, a semantic segmentation network designed for disaster-response scenarios, on an ARGO J8 unmanned ground vehicle using an attention-based architecture. To achieve this, the network was adapted through retraining with oversampling techniques to balance underrepresented critical classes, and the model was integrated into a ROS 2 framework running on an NVIDIA Jetson AGX Orin. The system acquires RGB images from a ZED 2i camera, generates multi-class semantic masks, estimates the distance and orientation of detected individuals, and displays the information through a visual interface to assist the operator. The complete pipeline operates at 16.4 FPS, although it can exceed 30 FPS when the limitations imposed by the input sensor and post-processing stages are removed. The source code is available at: `https://github.com/lucasjuradomartinez-cyber/TFGLucas_ws`.

Keywords: Perception and sensing, Robotics technology, Field robotics, Autonomous mobile robots, Machine learning, Autonomous robotic systems

*Autor para correspondencia: rvmartin@uma.es

1. Introducción

Los escenarios de búsqueda y rescate (SAR, *Search and Rescue*) constituyen uno de los dominios más exigentes para la robótica móvil (Orbea et al., 2024; Bravo-Arrabal et al., 2026). Tras un terremoto, un derrumbe, una explosión o un accidente industrial, el entorno puede presentar obstáculos irregulares, zonas inestables, visibilidad reducida, polvo, humo y presencia simultánea de víctimas, rescatadores, vehículos y escombros. En estas condiciones, los robots terrestres permiten explorar zonas peligrosas y transportar sensores sin exponer directamente a los equipos de intervención, pero su utilidad depende en gran medida de la información que sean capaces de proporcionar al operador (Casper and Murphy, 2003; Murphy, 2014).

La percepción visual basada en aprendizaje profundo ha adquirido un papel relevante en este contexto porque permite transformar imágenes en información semántica de alto nivel (Ramon-Soria et al., 2019). En particular, la segmentación semántica asigna una clase a cada píxel de la imagen, de forma que el robot no solo detecta objetos aislados, sino que obtiene una interpretación densa del entorno. Este planteamiento resulta especialmente útil en escenas SAR, donde es necesario distinguir personas civiles, personal de intervención, vehículos, caminos, edificios, vegetación, escombros o puestos de mando. Sin embargo, llevar una red de segmentación a soluciones embarcadas, impone retos adicionales (Alonso et al., 2020). Esto se debe a que la implantación en un robot real exige resolver problemas que no aparecen en una evaluación puramente *offline*: latencia, sincronización de sensores, consumo de GPU, estabilidad del flujo de ROS 2 y limitaciones térmicas y energéticas de la plataforma embarcada.

Este trabajo presenta la implantación de un sistema de percepción en tiempo real para el robot móvil ARGO J8 (ver Figura 1) con una cámara ZED 2i del Grupo de Robótica y Mecatrónica de la Universidad de Málaga. El sistema propuesto emplea SARNet, una arquitectura orientada a la segmentación semántica en escenas SAR(Salas-Espinales et al., 2023) basada en un esquema tipo U-Net (Ronneberger et al., 2015), con un codificador profundo ResNet-152 (He et al., 2016) y mecanismos de atención. Esta red se adapta mediante reentrenamiento para mejorar la detección de la clase *civilian* y se despliega sobre una NVIDIA Jetson AGX Orin integrada en el robot.

Las principales aportaciones del trabajo son: (i) Adoptar técnicas de reentrenamiento para clases humanas poco representadas en los datos; (ii) la integración de SARNet en ROS 2 para trabajar tanto con imágenes procedentes de dataset como con imagen real de la cámara ZED 2i; (iii) la optimización de la inferencia mediante TensorRT para su ejecución en hardware embarcado; y (iv) la validación experimental con el sistema embarcado a bordo de un UGV en túneles y demostraciones realistas de intervención.

El resto del artículo se organiza como sigue: la Sección 2 presenta el sistema de percepción y entrenamiento; la Sección 3 detalla la integración en el robot y la localización de víctimas; la Sección 4 expone los resultados experimentales y de rendimiento; y la Sección 5 resume las conclusiones.

Figura 1: Robot móvil ARGO J8 empleado como plataforma de validación.

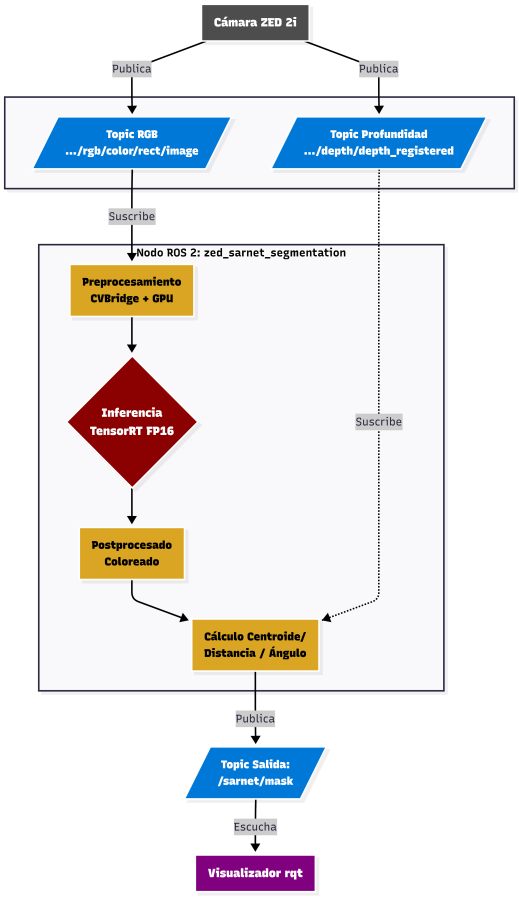
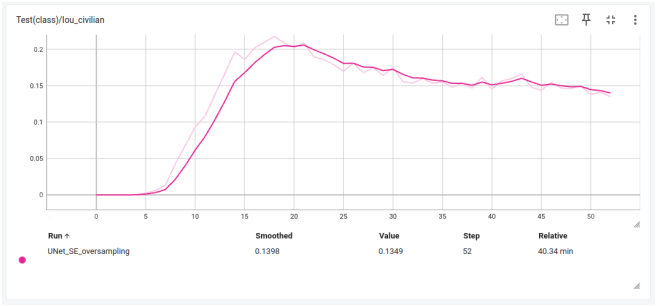


Figura 2: Arquitectura ROS 2 propuesta.

2. Sistema de segmentación semántica

2.1. *Arquitectura de segmentación semántica basada en SARNet*

En la Figura 2 se describe la arquitectura del sistema propuesto, para la que se ha desarrollado un nodo ROS 2. El proceso comienza con la captura de una imagen RGB-D mediante la cámara *ZED 2i*, cuyos tópicos son procesados por un nodo de ROS 2. Este realiza un preprocesamiento en GPU para alimentar el motor *TensorRT*, que ejecuta la inferencia de la red *SARNet* en precisión FP16. Tras el postprocesamiento de la máscara y el cálculo de parámetros espaciales (centroides, distancia y ángulo), el resultado se publica en el tópico `/sarnet/mask` para su visualización en *rqt* o su uso en otros nodos del control del vehículo.



(a) Entrenamiento con sobremuestreo.

(b) Entrenamiento con aumento de datos a partir del punto de control conseguido con sobremuestreo.

Figura 3: Evolución del *IoU* de la clase *civilian* en en el visualizador TensorBoard. Las líneas solidas indican la tendencia suavizada.

2.2. *Modelo de datos*

La robótica de rescate adolece de una gran escasez de datasets públicos etiquetados a nivel de píxel. Para el entrenamiento se ha utilizado el conjunto de datos etiquetado del UMA-SAR dataset (Morales et al., 2021), empleado en el entrenamiento previo de SARNet (Salas-Espinales et al., 2023). El conjunto de clases considerado incluye *background*, *first-responder*, *civilian*, *vegetation*, *building*, *dirt-road*, *road*, *sky*, *civilian-car*, *response-vehicle*, *debris* y *command-post*. Entre ellas, las clases humanas tienen una importancia especial: *first-responder* permite reconocer al personal de emergencia y *civilian* representa a posibles víctimas o personas que requieren asistencia. El dataset empleado consta de 655 imágenes etiquetadas distribuidas en una relación 85:10:5 para entrenamiento, test y validación, respectivamente.

En nuestro conjunto de datos, y como es habitual en datasets tomados del mundo real (Buda et al., 2018), el desequilibrio de clases es un problema crítico: los civiles representan solo el 0,15 % de los píxeles. Consecuentemente, la implementación original de SARNet confunde la clase *civilian* con *first-responder*, obteniendo un *IoU* del 0 % como medida de aciertos en la segmentación de civiles Salas-Espinales et al. (2023).

2.3. *Entrenamiento en escenarios de rescate*

El proceso de entrenamiento se ejecutó usando Pytorch en una estación NVIDIA DGX con 4 GPUs Tesla V100, mientras que el despliegue final se realizó en la Jetson AGX Orin. Esta separación permitió aprovechar una plataforma de alto rendimiento en la fase de aprendizaje.

Para solucionar el bajo rendimiento en la clase *civilian*, se propone una metodología para recalcular los pesos del modelo. En ésta se combina *Dice Loss* (que maximiza el solapamiento espacial, mejorando la detección de objetos pequeños) con *CrossEntropy Loss* ponderada, lo que penaliza severamente los errores en clases minoritarias. Esta fusión obligó a la red a no ignorar la clase *civilian*. El proceso se refuerza con sobremuestreo de imágenes con civiles (*oversampling*) y un aumento de datos (*data augmentation*) mediante cambios de brillo, volteos y recortes forzados sobre los civiles (*Random CropAroundClass*). El objetivo no es maximizar únicamente una métrica global, sino mejorar la respuesta de la red ante la clase minoritaria sin degradar significativamente el rendimiento en el resto de las clases.

La Figura 3 muestra cómo, tras introducir el sobremuestreo de la clase *civilian*, esta logró superar el 0 % del porcentaje de *IoU* en el que se encontraba estancada. Así, se eleva el *IoU* para esta clase, que posteriormente desciende debido al sobreentrenamiento (ver Figura 3(a)). El entrenamiento a partir de ese punto con aumento de datos también logra una mejora notable (Figura Figura 3(b)). En total, el modelo alcanzó un *IoU* medio de 70,33 %, obteniendo un 34,449 % para la clase *civilian*. Aunque este IoU evidencia que la detección de víctimas sigue siendo un reto, representa una mejora significativa respecto al comportamiento inicial.

3. Implementación a bordo

3.1. *Sistema robótico y de visión*

La plataforma utilizada para la validación es el robot móvil ARGO J8 de la Universidad de Málaga, mostrado en la Figura 1. Se trata de un vehículo anfibio de tracción 8x8 diseñado para desplazarse por terrenos extremos. Sus dimensiones aproximadas son 1,40 x 2,94 x 1,00 m, con una masa cercana a 750 kg, una velocidad máxima de 30 km/h, una capacidad de carga de hasta 600 kg y estabilidad en pendientes de hasta 40° (De Nóbrega-Buyón et al., 2025). Estas características lo convierten en una plataforma adecuada para la evaluación de sistemas de percepción en escenarios de catástrofes y de robótica de campo.

El cómputo embarcado se realiza en una NVIDIA Jetson AGX Orin, que permite ejecutar redes neuronales convolucionales en GPU mediante CUDA y TensorRT. Al tratarse de una plataforma embarcada, no basta con que el modelo produzca una segmentación correcta: la inferencia debe convivir con el resto de los procesos del robot, mantener una frecuencia suficiente y evitar bloqueos por el uso excesivo de recursos.

El sistema de visión se basa en una cámara estéreo ZED 2i. En el flujo final, la imagen RGB rectificada alimenta la red de segmentación, mientras que el mapa de profundidad registrado se utiliza para estimar la posición relativa de las personas detectadas.

3.2. *Integración en ROS 2 e inferencia en tiempo real*

La implementación se diseñó de forma modular sobre ROS 2. Se desarrollaron nodos capaces de operar en dos modos principales: un modo de prueba con imágenes del dataset y un modo de despliegue con una imagen RGB en tiempo real proveniente de la ZED 2i. En ambos casos, la salida principal es una máscara semántica publicada en un tópico de ROS 2, acompañada de una imagen coloreada para su visualización.

(a) Imagen rectificada real (b) Salida del nodo

Figura 4: Segmentación semántica en tiempo real tras la aceleración con TensorRT.

(a) Imagen rectificada real (b) Salida del nodo

Figura 5: Resultados del procesamiento: imagen real y salida del nodo.

El modelo se exportó a ONNX para lograr una inferencia en tiempo real sin necesidad de recurrir a procesos innecesarios, como los que ejecuta PyTorch con librerías para el entrenamiento, la monitorización y el análisis de datos. Además, de esta forma se puede emplear el motor TensorRT, que permite fusionar operaciones, seleccionar implementaciones eficientes de capas y reducir la precisión cuando es posible, acelerando la inferencia sin modificar la lógica de alto nivel del sistema (Guo et al., 2025). La Figura 4 muestra una salida representativa del sistema tras la aceleración. En la Figura 4(a) se muestra la imagen original de la cámara ZED 2i rectificada con una resolución de 1280×720 píxeles, mientras que la imagen 4(b) muestra la imagen segmentada por la red con una resolución de 640×480.

Además de convertir el modelo, se optimizó el flujo de datos: se redujo la cola de ROS 2, se configuró el perfil de potencia de la Jetson, se estabilizó la GPU y se ajustó la publicación de la cámara. Estas medidas evitaron el uso de imágenes anteriores al tiempo actual y permitieron obtener una salida útil en tiempo real, prioritaria en robótica de rescate frente a una segmentación más precisa pero más retrasada.

3.3. *Localización espacial de víctimas*

Para el cálculo de parámetros espaciales de las instancias de las clases *civilian* y *first-responder*, se realiza primero el análisis de componentes con 8-conectividad separando las distintas instancias que puedan aparecer en la imagen. Posteriormente, se calcula la profundidad como la mediana de la región. Con esta profundidad y gracias a los parámetros intrínsecos de la cámara, se consigue calcular por geometría 3D la distancia y ángulo respecto a la cámara, y publicar ambas cosas junto con la máscara coloreada (Hartley and Zisserman, 2004).

La Figura 4 ofrece un ejemplo de localización espacial de una persona (clasificada como rescatista) donde se indica el centroide. En esta imagen, con la persona a aproximadamente 2 m de la cámara, el error máximo es de +5 cm. La Figura 10 muestra distintas instancias de *rescatadores* (naranja) y *civiles* (morado), donde la segmentación es correcta. Esta figura también ilustra alguna limitación de la 8-conectividad, que unifica instancias que se solapan a distintas profundidades.

(a) Imagen rectificada real (b) Salida del nodo

Figura 6: Detección fallida de persona tumbada en pendiente.

(a) Imagen rectificada real (b) Salida del nodo

Figura 7: Detección correcta de persona tumbada en pendiente.

4. Experimentación y resultados

4.1. *Procedimiento experimental*

Además de las pruebas con imágenes del dataset y del flujo completo en ROS 2, la experimentación se planteó desde el inicio en tres partes. En primer lugar, se definieron dos experimentos específicos para analizar el comportamiento de SARNet ante situaciones que se preveían desfavorables: la detección de víctimas en posturas no convencionales y la segmentación en condiciones de baja iluminación. Posteriormente, se contempló una validación en un entorno de intervención realista, que pudo llevarse a cabo mediante la participación en las XX Jornadas Internacionales de la Universidad de Málaga sobre Seguridad, Emergencias y Catástrofes 2026 (JEMERG26). De este modo, la evaluación permitió estudiar tanto limitaciones concretas de la red como su funcionamiento integrado en un escenario próximo a una operación real de búsqueda y rescate.

4.1.1. *Influencia de la postura de la víctima*

Este experimento evaluó la detección de personas tumbadas, una situación habitual en escenarios SAR pero poco representada en el conjunto de entrenamiento, donde predominan personas en posición erguida. Las Figuras 6 y 7 muestran que la respuesta de la red depende de la postura, la inclinación del terreno y el punto de vista de la cámara. La detección varía entre las pruebas cuesta arriba y cuesta abajo, ya que la perspectiva modifica la silueta de la víctima y puede facilitar o dificultar su segmentación.

4.1.2. *Influencia de la iluminación*

Se analizó también el comportamiento del sistema en túneles y zonas con iluminación reducida. Estas condiciones disminuyen el contraste entre la víctima y el entorno, dificultando la segmentación basada en imagen RGB. Estas pruebas, representadas en la Figura 8, demostraron que solo con la imagen RGB en escenas completamente oscuras, la segmentación puede fallar. Sin embargo con una ligera iluminación, que podría llevar el propio robot, se consigue una correcta segmentación.

(a) Imagen rectificada real

Figura 8: Segmentación de una per
ligeramente la iluminación de la esc

Cabe señalar que estas validac
se presentan como análisis cualita
máscara de referencia (ground truth
en tiempo real en campo impide u
exhaustiva, limitando el análisis a l
tamiento operativo y la respuesta de
críticas.

4.1.3. *Validación de los resultado*

La validación en un entorno de
bo durante las XX Jornadas Interna
de Málaga sobre Seguridad, Emerg
(JEMERG26), en las que el sistema
durante ejercicios con profesionales

El primer escenario consistió e
túnel tras un incendio y un derrum
bot móvil ARGO J8 realizó una exploración del interior del túnel tras la extinción del incendio por parte de los bomberos, enviando imágenes al personal de la Unidad Militar de Emergencias (UME) ubicado en la entrada del túnel antes de su intervención para el rescate de las víctimas. De esta forma, la exploración realizada por el vehículo permitió a los rescatadores obtener información visual remota del interior del túnel, a partir de las imágenes captadas por la cámara ZED y de la segmentación semántica de la escena en tiempo real. La salida del sistema se mostró en la interfaz diseñada y se explicó al personal de rescate, lo que permitió comprobar su utilidad como apoyo para interpretar la imagen RGB en zonas degradadas o de baja visibilidad.

El coordinador de la UME destacó que la visualización combinada facilita una identificación más rápida de los elementos de la escena. En condiciones de baja visibilidad o presencia de humo, la máscara semántica les permite interpretar el entorno con mayor claridad que mediante la imagen RGB convencional, reduciendo así los tiempos de toma de decisiones y agilizando la ejecución de la operación.

(a) Intervención en el primer escenario.

Figura 10: Robot J8 durante la intervención, con la ubicación de la cámara.

Tabla 1: Evolución aproximada de rendimiento durante el desarrollo.

Versión de implementación	**Velocidad** (FPS)
Sistema completo con profundidad, interfaz y postprocesado	16,4
Segmentación mínima con entrada ZED a 60 FPS	22,5
Entrada desde dataset, sin dependencia de cámara	32

El segundo escenario se desarrolló en una zona exterior de rescate, con mejores condiciones de iluminación y una geometría más abierta. En este caso, el robot J8 actuó como plataforma de apoyo para aproximarse a la zona de intervención y ofrecer una vista cercana del entorno, mientras los bomberos y el personal sanitario trabajaban con las víctimas. La Figura **??** muestra el despliegue inicial del sistema y la situación del robot durante la intervención. La realización de estas pruebas en los dos escenarios de las jornadas permitió validar el sistema en condiciones realmente realistas de intervención, comprobando su comportamiento fuera del laboratorio y con el robot

integrado en ejercicios de emergencia, donde aparecen limitaciones propias del despliegue en campo.

4.2. Análisis de resultados

La Tabla 1 resume la velocidad de inferencia que se puede alcanzar en distintos casos y cómo se reduce en condiciones reales al integrar la segmentación semántica con el resto de procesos como la sincronización con las cámaras, el procesamiento estéreo y la localización espacial.

La versión funcional de demostración, que incluye cámara, segmentación, profundidad, distancia, ángulo y visualización, alcanzó 16,4 FPS (61 ms). Este valor es suficiente para proporcionar percepción de apoyo al operador. Las pruebas de máximo rendimiento en las que se eliminó el cálculo tridimensional de la víctima mostraron que la inferencia sola, alcanza 22,5 FPS (44 ms) con entrada real de cámara, y 32 FPS (31 ms) al alimentar la red desde el dataset, confirmando que el postprocesado y la localización espacial añaden una sobrecarga controlada de 17 ms en el ciclo completo.

Desde el punto de vista cualitativo, el sistema fue capaz de segmentar elementos del entorno y detectar personas tanto en escenarios exteriores como en túneles. Sin embargo, también se observaron limitaciones. La red detecta peor posturas poco representadas en el dataset de entrenamiento, como personas tumbadas en determinadas configuraciones; la segmentación se degrada ante oclusiones severas; y la ausencia casi total de iluminación puede impedir la detección, aunque una mejora ligera de la visibilidad puede ser suficiente para recuperar la respuesta del sistema. Estos resultados indican que la solución es funcional y prometedora, pero que la robustez final depende de ampliar el dataset con escenas reales capturadas por el propio robot y con una mayor diversidad de condiciones.

5. Conclusiones

En este trabajo se ha presentado un sistema de percepción en tiempo real para la robótica de búsqueda y rescate basado en la segmentación semántica. La solución parte de SARNet, una red orientada a escenas SAR, y aborda su adaptación completa a una plataforma robótica real: anotación y comprensión del dataset, reentrenamiento de clases críticas, integración con ROS 2, optimización con TensorRT y despliegue en una NVIDIA Jetson AGX Orin instalada en el robot ARGO J8.

Los resultados muestran que el sistema completo puede ejecutar segmentación, profundidad, localización aproximada de personas y visualización a 16,4 FPS, mientras que las pruebas de benchmark demuestran que la red optimizada puede superar los 30 FPS cuando no está limitada por el sensor de entrada. Además, la integración de profundidad permite convertir una detección semántica en una referencia espacial útil para el operador, proporcionando distancia y orientación aproximadas de posibles víctimas o intervinientes.

Como trabajo futuro, se plantea ampliar el conjunto de datos con imágenes reales capturadas desde el ARGO J8 en condiciones de despliegue, incorporar nuevas clases relevantes en intervenciones reales, como perros de rescate, camillas o herramientas, y evaluar arquitecturas más ligeras para aumentar la frecuencia de inferencia sin reducir la precisión en clases críticas. También resulta necesario estudiar cuantitativamente el error de profundidad en función de la iluminación, la distancia, el ángulo de orientación y el tipo de superficie. Por último, la incorporación de imágenes térmicas para la fusión de información puede resolver problemas en condiciones de baja visibilidad.

Agradecimientos

Este trabajo ha sido realizado parcialmente gracias al apoyo del Ministerio de Ciencia, Innovación y Universidades, Gobierno de España, proyecto PID2021-122944OB-I00. El trabajo ha recibido apoyo técnico del Instituto Universitario de Investigación en Ingeniería Mecatrónica y Sistemas Ciberfísicos de la Universidad de Málaga a través del proyecto PPRO-IUI-2023-02. Los autores quieren agradecer la colaboración de la Cátedra de Seguridad, Emergencias y Catástrofes de la Universidad de Málaga, de la Unidad Militar de Emergencias y del Real Cuerpo de Bomberos del Ayuntamiento de Málaga.

Referencias

Alonso, I., Riazuelo, L., Murillo, A.C., 2020. MiniNet: an efficient semantic segmentation ConvNet for real-time robotic applications. IEEE Transactions on Robotics 36, 1340 – 1347. doi:10.1109/TRO.2020.2974099.

Bravo-Arrabal, J., Vázquez-Martín, R., Fernández-Lozano, J.J., García-Cerezo, A., 2026. Strengthening multi-robot systems for search and rescue: Co-designing robotics and communications toward 6g. IEEE Communications Standards Magazine 10, 305–312. doi:10.1109/MCOMSTD.2026.3678471.

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks 106, 249–259. doi:10.1016/j.neunet.2018.07.011.

Casper, J., Murphy, R.R., 2003. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 33, 367–385. doi:10.1109/TSMCB.2003.811794.

De Nóbrega-Buyón, K., Fernández-Lozano, J.J., Vázquez-Martín, R., 2025. Modelado y simulación de un robot móvil en entornos naturales mediante ROS2 y Unity3D. Jornadas de Automática 46. doi:10.17979/ja-cea.2025.46.12187.

Guo, X., Liu, T., Li, Y., Lin, Z., Deng, Z., 2025. TUNI: Real-time rgb-t semantic segmentation with unified multi-modal feature extraction and cross-modal feature fusion. arXiv preprint arXiv:2509.10005 doi:10.48550/arXiv.2509.10005.

Hartley, R., Zisserman, A., 2004. Multiple View Geometry in Computer Vision. 2 ed., Cambridge University Press.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. doi:10.1109/CVPR.2016.90.

Morales, J., Vázquez-Martín, R., Mandow, A., Morilla-Cabello, D., García-Cerezo, A., 2021. The UMA-SAR Dataset: Multimodal data collection from a ground vehicle during outdoor disaster response training exercises. The International Journal of Robotics Research 40, 835–847. doi:10.1177/02783649211004959.

Murphy, R.R., 2014. Disaster Robotics. The MIT Press. doi:10.7551/mitpress/9407.001.0001.

Orbea, D., Ulloa, C.C., Cerro, J.d., Barrientos, A., 2024. Mobile victim signs monitoring through non-invasive robotic system. Lecture Notes in Networks and Systems 1114 LNNS, 141 – 153. doi:10.1007/978-3-031-70722-3_15.

Ramon-Soria, P., Gomez-Tamm, A., Garcia-Rubiales, F., Arrue, B., Ollero, A., 2019. Autonomous landing on pipes using soft gripper for inspection and maintenance in outdoor environments, in: IEEE International Conference on Intelligent Robots and Systems, pp. 5832 – 5839. doi:10.1109/IROS40897.2019.8967850.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer. pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

Salas-Espinales, A., Vázquez-Martín, R., García-Cerezo, A., Mandow, A., 2023. SAR Nets: An Evaluation of Semantic Segmentation Networks with Attention Mechanisms for Search and Rescue Scenes, in: IEEE International Symposium on Safety, Security, and Rescue Robotics, IEEE. pp. 139–144. doi:10.1109/SSRR56537.2023.10400741.