Universidad Miguel Hernández de Elche

MÁSTER UNIVERSITARIO EN ROBÓTICA

ANÁLISIS DE TÉCNICAS DE AUMENTO DE DATOS PARA LOCALIZACIÓN ROBUSTA DE ROBOTS MÓVILES MEDIANTE REDES NEURONALES CONVOLUCIONALES

Trabajo de Fin de Máster

Curso académico 2021-2022

Autor: Orlando José Céspedes Gómez

Tutores: Luis Payá Castelló

Sergio Cebollada López

Índice general

Índice de figuras

Índice de tablas

Capítulo 1
Introducción

La robótica se ha convertido a lo largo de décadas en una de las ramas más importantes de la ciencia, lo que ha supuesto un gran desarrollo e investigación en torno a este tema, hasta llegar a transformarse en un campo de estudio indispensable en la actualidad. Asimismo, la robótica autónoma son sistemas que sustituyen a los humanos para llevar a cabo tareas mecánicas, rutinarias o peligrosas que requieren un alto grado de autonomía para su realización.

Cuando el robot móvil tiene que desplazarse para llevar a cabo la tarea para la cual ha sido programado, se tendrán entonces que realizar tareas correspondientes al campo de la robótica móvil. En este sentido, el hecho de que este no sea fijo y pueda navegar por el entorno en el que se encuentra permite mejorar el trabajo en áreas grandes y heterogéneas sin tener que hacer ningún cambio en la estructura del robot. Es por ello por lo que se produce un aumento de los campos donde se puede utilizar la robótica. Un correcto ejemplo sería en entornos de difícil acceso para los humanos como lo son las tareas de reconocimiento en túneles [24], los trabajos de búsqueda y rescate [10] y los desplazamientos en ambientes no estructurados [1]. Actualmente, este tipo de robot también se utiliza en el hogar para tareas domésticas y de entretenimiento. Es por ello por lo que en la figura 1.1 se presentan algunos ejemplos de las distintas clases de robots desarrollados hoy en día.

Figura 1.1: Diferentes sistemas robóticos.

Durante los primeros años, el estudio de la robótica se centró en el rendimiento, la repetibilidad y la velocidad. Sin embargo, en la actualidad, se han creado campos más específicos como es el caso de la robótica móvil, el cual se centra en resolver tareas, entre otras muchas, de localización y navegación. Esta área se trata de un campo de conocimiento multidisciplinar que dota a los robots móviles de distintos sensores que les permiten extraer del medio la información necesaria para llevar a cabo de forma autónoma las tareas para las que fueron creados. El término autonomía se refiere al hecho de que el robot sea capaz de ejecutar dichas tareas sin necesidad de intervención del ser humano.

Para que un robot móvil realice una tarea específica de forma autónoma en un entorno desconocido es necesario crear un modelo de este medio. Dicho modelo debe poseer suficiente información para que el robot estime su posición y orientación con precisión y un coste computacional razonable; y luego planifique un camino libre de obstáculos para llegar a los puntos deseados (planificación de trayectorias). En estos aspectos radica la importancia de hallar soluciones óptimas a los problemas de construcción de mapas (*mapping*) y localización.

Asimismo, estos últimos se consideran dos áreas de estudio fundamentales en el desarrollo de robots móviles totalmente autónomos. Las soluciones a estos problemas han sido áreas de investigación muy activas en el campo de la robótica móvil durante años. Primero, el término *mapping* se refiere al problema de construir modelos espaciales en medios físicos utilizando la información capturada por los sistemas de percepción montados en el entorno y/o en uno o

más robots. Por otro lado, en cuanto al concepto de localización, este reside en estimar la pose del robot en el modelo, es decir, sus coordenadas de posición y orientación en todos los grados de libertad de movimiento que dispone el robot. De esta forma, se compara la información captada por sus sistemas de percepción con la información almacenada en el mapa creado con anterioridad. A veces, el problema de localización debe resolverse mientras se crea el mapa. Este es un problema muy común porque, cuando un robot comienza a llevar a cabo una tarea en un entorno desconocido, tiene que comenzar a construir un modelo del entorno desde cero mientras al mismo tiempo estima su posición. Este proceso se denomina *Simultaneous Localization and Mapping*, comúnmente conocido por sus siglas en inglés: SLAM.

Los tres conceptos anteriores (*mapping*, localización y planificación de trayectorias) están relacionados como se muestra en la figura 1.2. A continuación, se dará paso al resto de términos que aparecen en esta imagen, obviando la función de SLAM al haber sido ya explicada previamente. Por una parte, el uso simultáneo de algoritmos de creación de mapas con algoritmos de planificación de trayectorias se denomina exploración. El objetivo de este problema es determinar la trayectoria óptima que deben seguir los robots para crear un mapa del medio lo más rápido posible sin perder la precisión de la información que se almacena. Por otra parte, la combinación de los problemas de localización y planificación de trayectorias da lugar a la tarea de localización activa, donde el robot intenta seguir determinadas trayectorias que le ayudan a perfeccionar su localización en función de la información recopilada a lo largo de estas rutas. Finalmente, la ejecución paralela de problemas de *mapping*, localización y planificación de trayectorias conduce a lo que llamamos sistemas de navegación integrados, el último horizonte a considerar cuando se diseña un robot móvil autónomo [19].

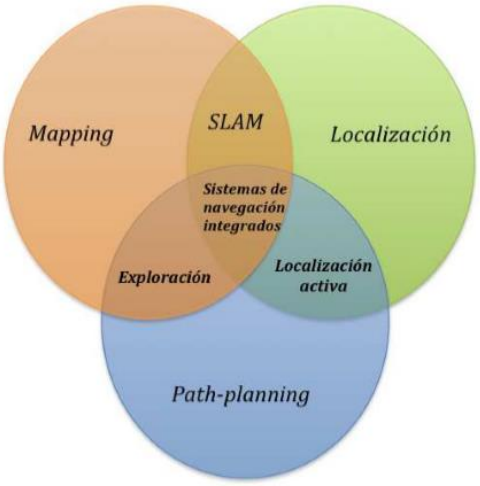


Figura 1.2: Interrelación entre los conceptos que engloba la navegación integrada en robótica móvil: *mapping*, localización y planificación de trayectorias.

Algunos de los sensores que se pueden utilizar para estos fines son el sonar, el ultrasónico, el láser o las cámaras. La utilización de éstos últimos se ha visto incrementada debido a las ventajas que presentan. De hecho, se aplica en distintos ámbitos de la robótica móvil, como la localización, la navegación visual, la odometría visual, la tarea de SLAM, etc. Los sistemas de visión ofrecen abundante información sobre la escena, siendo similar a la que obtendría el ojo humano. Entre otras de sus ventajas se encuentran su bajo coste y que son fáciles de usar. Además, son más eficientes en términos de consumo que otros sensores, por lo que son una buena opción en tareas cuya realización requiere un elevado tiempo [9]. De entre la alta variedad de tipos de cámaras que ofrece el mercado hoy en día, destaca el uso de las cámaras omnidireccionales, las cuales han presentado la ventaja de ofrecer un campo de visión más amplio. De esta manera, si se incorpora una cámara omnidireccional sobre un sistema robótico, con un solo sensor se puede obtener información en 360º del entorno que rodea al robot, incluyendo así toda la escena en una única imagen.

Generalmente, en visión por computador se extrae información relevante de las imágenes. Los métodos de extracción de información más utilizados son los basados en características locales y los basados en apariencia global (también conocidos como métodos de descripción holística). El primer modo consiste en obtener un conjunto de puntos pertenecientes a cada imagen con

unas determinadas características. Sin embargo, los métodos basados en la apariencia global utilizan las imágenes en su conjunto, no extraen ninguna información local. Por tanto, se obtendrá un único vector para cada imagen que contendrá información sobre su apariencia global. Este tipo de métodos presentan algunas ventajas en aquellos entornos dinámicos o mal estructurados [21]. Sin embargo, hay que tener en cuenta que esta información visual que se obtiene por medio de las imágenes puede verse afectada por los cambios de iluminación. Estos pueden hacer ver comprometida una correcta tarea de localización, haciendo creer al robot que se encuentra en un punto muy distinto al real debido a dichos efectos.

Además, una gran parte de la investigación y desarrollo de la robótica móvil actual se basa en técnicas que proporcionan a los robots un mayor grado de autonomía y versatilidad. Con el paso de los años se ha hecho imprescindible el uso de la tecnología de visión artificial para la tarea de navegación, ya que esta tecnología nos permite adquirir, procesar y analizar imágenes. De hecho, un ejemplo conocido puede ser el que se muestra en la figura 1.3, en el que se presenta a uno de los robots más importantes de la NASA pertenecientes al programa *Mars Exploration Rover* (MER). Hay que tener en cuenta que un rover es un dispositivo de exploración de superficie planetaria creado para moverse a través de la superficie de un planeta u otros cuerpos celestes. Debido a que los rovers no pueden ser controlados de forma remota en tiempo real, ya que la velocidad a la que viajan las señales de radio es demasiado lenta para tener una comunicación instantánea, estos deben operar de forma autónoma, con poca ayuda del control de tierra en lo que respecta a la navegación y adquisición de datos. Es por ello por lo que es tan importante la autonomía de estos robots móviles, ya que el conseguir que puedan tomar grandes decisiones por sí mismos hizo posible que el rover *Opportunity* superara múltiples tormentas y varios fallos mecánicos durante la misión.

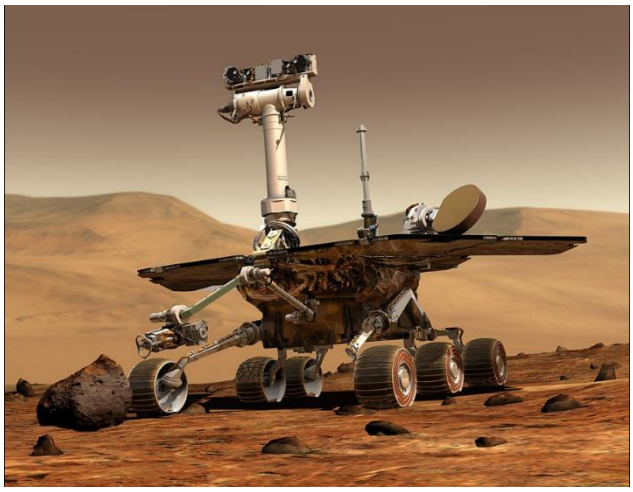


Figura 1.3: Robot rover *Opportunity*.

El hecho de que estos robots presenten navegación autónoma hace que tengan la capacidad de planificar y seguir una trayectoria de forma óptima evitando los obstáculos presentes en su entorno de trabajo [12].

Toda esta labor de procesamiento, en la tarea de saber dónde exactamente se encuentra nuestro robot, pasa a ser mucho más eficiente cuando trabajamos con técnicas de *machine learning*. Esta es una derivación de la inteligencia artificial que crea sistemas que aprenden de manera automatizada, es decir, identifican patrones complejos de entre millones de datos y predicen comportamientos mediante algoritmos. Dichos algoritmos son capaces, de manera independiente, de autoprogramarse según lo que aprenden y la experiencia que obtienen de la combinación de datos y el procesamiento de la información. Esta técnica es muy importante a la hora de la creación de mapas que representen la actividad y la navegabilidad del entorno en el que se encuentra nuestro robot móvil [29]. Asimismo, dentro del área de *machine learning*, podemos trabajar con herramientas de *deep learning*, ya que con ello conseguimos un aprendizaje profundo empleando modelos informativos y creando redes neuronales artificiales para la transmisión y análisis de datos. Esta técnica más reciente ha mostrado resultados sobresalientes a la hora de resolver una amplia variedad de tareas en robótica móvil, como puede ser en las áreas de percepción, planificación, localización y control [4].

Respecto a la tarea de localización desde un punto de vista jerárquico, trabajos previos han demostrado que el uso de estos modelos con descriptores holísticos e imágenes omnidireccionales lleva a una solución eficiente y robusta para abordar la tarea de localización [22]. Asimismo, la localización jerárquica se lleva a cabo en dos pasos y en ambos se emplea la arquitectura de una red neuronal convolucional (CNN) con diferentes objetivos. En primer lugar, se lleva a cabo una localización gruesa que consiste en identificar la estancia en la que se encuentra el robot por medio de la CNN. Seguidamente, se realiza una localización fina en dicha estancia, en la cual la red neuronal convolucional es empleada para la obtención de descriptores holísticos a partir de las capas intermedias de la red. Estos descriptores globales permiten hallar la posición donde se encuentra el robot de manera más precisa por medio de una búsqueda del vecino más cercano, por lo que se compara el descriptor correspondiente de la imagen de test con cada uno de los descriptores de las imágenes capturadas en la estancia seleccionada en el primer paso [3].

Con el fin de mejorar el desempeño de la red se recurre a un aumento de datos (*data augmentation*). Esta técnica es utilizada para aumentar la cantidad de datos agregando copias ligeramente modificadas de datos ya existentes. Este proceso ayuda a evitar que la red se sobreajuste y memorice los detalles exactos de las imágenes de entrenamiento. Dicha técnica demuestra ser una solución eficiente y robusta para afrontar el problema de localización tal y como se muestra en el capítulo de resultados.

Con todo ello podemos ver las grandes oportunidades que nos ofrece el tratamiento de imágenes mediante estas técnicas, debido a que estaríamos trabajando con mayor precisión, anticipándonos a los problemas que fuesen apareciendo en el camino a recorrer y aprovechando mejor nuestros recursos en un menor tiempo de actuación.

Por tanto, este trabajo fin de máster tiene como principal objetivo abordar la localización de un robot móvil mediante la readaptación y entrenamiento de una CNN (*Convolutional Neural Networks*). Para ello, se analizan los efectos utilizados para llevar a cabo el aumento de datos del modelo visual para, de esta

forma, realizar una localización jerárquica tanto con cambios de iluminación en el entorno como sin ellos. Dado que se trata de una técnica menos madura, es interesante estudiar con profundidad su comportamiento en tareas de localización y la robustez de cada uno de estos algoritmos de clasificación ante determinados fenómenos que hacen cambiar el entorno real de trabajo. La manera de llevar a cabo dicha tarea es por medio del método conocido como *room retrieval* o recuperación de la estancia, el cual se explicará con más detalle en capítulos posteriores.

Asimismo, durante el proceso de aumento de datos se ha realizado un estudio acerca de cuán robusta es la CNN empleada, por lo que para ello se han obtenido sus resultados de clasificación de estancias comprobando así su eficiencia ante el cambio sustancial de la apariencia visual del entorno. No hay que olvidar que de estar en un ambiente de trabajo real el robot podría enfrentarse a diversos fenómenos perjudiciales, los cuales debe saber solventar con éxito para poder llevar a cabo una correcta tarea. Estos, por ejemplo, podrían ser una variación en el entorno, ya que puede verse parcialmente ocluido debido a la actividad humana en el espacio de trabajo, o una nueva organización del medio que haya hecho cambiar el mobiliario de las estancias.

Por último, el presente trabajo se encuentra estructurado de la siguiente manera:

- En el capítulo 2 se presenta un breve estado del arte acerca, por un lado, de las cámaras omnidireccionales y el beneficio que estas aportan a la hora de ser usadas en la tarea de localización de robots móviles en entornos de interior. Y, por otro lado, del uso de herramientas de *deep learning* para llevar a cabo tareas de localización jerárquica, tanto para la capa de alto nivel (localización gruesa) como para la capa de bajo nivel (localización fina).

- A continuación, el capítulo 3 está centrado en la descripción de la base de datos utilizada, de la cual hemos creado nuestro modelo visual. Se detalla, también, cómo han sido obtenidas las imágenes,

así como las estancias que el robot móvil ha recorrido. En este mismo capítulo, posteriormente, se especifica la red neuronal convolucional empleada (CNN) para el tratamiento de la información. Asimismo, se describe el funcionamiento de la misma y se trata desde su configuración estructural hasta el tipo de imagen que esta acepta. Para finalizar, se desarrolla con mayor énfasis la técnica de *data augmentation* y se presentan los efectos que son aplicados a nuestro modelo visual.

- En el capítulo 4 se explican los seis experimentos que se realizan para conocer cuán robusta es nuestra CNN ante cada uno de los diferentes efectos que en esta investigación se quieren llevar a cabo.

- Finalmente, las principales conclusiones se exponen en el capítulo 5, en el cual se detallan asimismo las aportaciones más relevantes obtenidas tras haber realizado lo expuesto en la sección anterior y las futuras líneas de investigación.

Capítulo 2
Estado del arte

Para poder realizar las tareas de *mapping*, localización y navegación de los robots móviles es imprescindible utilizar la información percibida por los sistemas sensoriales del robot. Desde los primeros trabajos sobre robots móviles [13], los sistemas de control diseñados han hecho un mayor o menor uso de la información recopilada del entorno por el que se mueve el robot. La manera de trabajar con esta información ha ido evolucionando a medida que surgían nuevos algoritmos y aumentaba la capacidad de los sistemas de percepción y de computación.

El hecho de que los robots integren la función de visión les da la capacidad de percibir e interactuar con el entorno que les rodea. Además, esta se utiliza en diversas técnicas debido a la gran cantidad de información que suministran sobre la escena. La robótica visual utiliza métodos de visión por computador para realizar las tareas que el robot debe llevar a cabo. El objetivo de estos métodos es tratar de comprender la escena y los objetos que se encuentran en ella, de forma similar a la función que realizan los humanos.

2.1. Cámaras omnidireccionales

Una cámara es un instrumento de formación de imágenes como resultado de la luz proyectada sobre una superficie sensible a esta. Se pueden lograr diferentes cámaras cambiando tres elementos: (1) la geometría de la superficie, (2) la distribución geométrica y las propiedades ópticas de los fotorreceptores, y (3) la forma en que la luz se recolecta y se proyecta sobre la superficie (simples o múltiples lentes). Los sistemas de visión procesan estas imágenes para reconocer, navegar y, en general, interactuar con el medio [18]. Las cámaras

presentan diversas ventajas frente a otro tipo de sensores, por lo que un ejemplo de estas puede ser su bajo coste, peso o consumo energético. Esta última es una propiedad muy útil para los robots autónomos, junto con el alto nivel de información que proporcionan sobre el entorno.

En cuanto al sensor, podemos hallar sistemas basados en configuraciones monoculares [25] y binoculares (pares estéreo) [26], aunque algunos autores también presentan soluciones basadas en configuraciones trinoculares [2]. Estas dos últimas permiten cuantificar la profundidad de las imágenes, pero necesitan de varias de estas para adquirir información completa del entorno. Además, los sistemas de visión omnidireccional pueden constar de múltiples cámaras apuntando en diferentes direcciones, o de una sola cámara y una superficie reflectante [21].

En las últimas décadas, los sistemas de visión omnidireccional [28] han ganado popularidad gracias a las siguientes ventajas: la gran cuantía de información que proporcionan debido a que tienen un campo de visión de 360º alrededor del robot; la estabilidad de las características que se muestran en la imagen, ya que permanecen en el campo de visión durante más tiempo mientras el robot se desplaza por el entorno; su costo relativamente bajo en comparación con otros sensores como pueden ser los sensores láser; y su bajo consumo de energía, que es útil en el diseño de robots autónomos que necesitan funcionar con baterías durante prolongados períodos de tiempo. Estos sistemas de visión omnidireccional se basan mayoritariamente en la combinación de una cámara convencional con un espejo convexo cónico, esférico, parabólico o hiperbólico (sistemas catadióptricos). La información visual de estas cámaras se puede presentar en varios formatos: omnidireccional, panorámica o vista en planta [20], [11]. En este trabajo utilizaremos la representación panorámica, ya que, como mostraremos en capítulos posteriores, esta contiene suficiente información para estimar la posición y la orientación del robot siempre que el movimiento de este esté restringido al plano del suelo.

Asimismo, el sensor usado en este trabajo fin de máster ha sido un sistema de visión catadióptrico montado sobre el propio robot, tal como podemos ver en

la figura 2.1, que proporciona imágenes con un campo de visión de 360° alrededor del mismo. Es con este robot móvil con el que se ha creado el modelo visual del entorno a partir de las imágenes omnidireccionales capturadas.

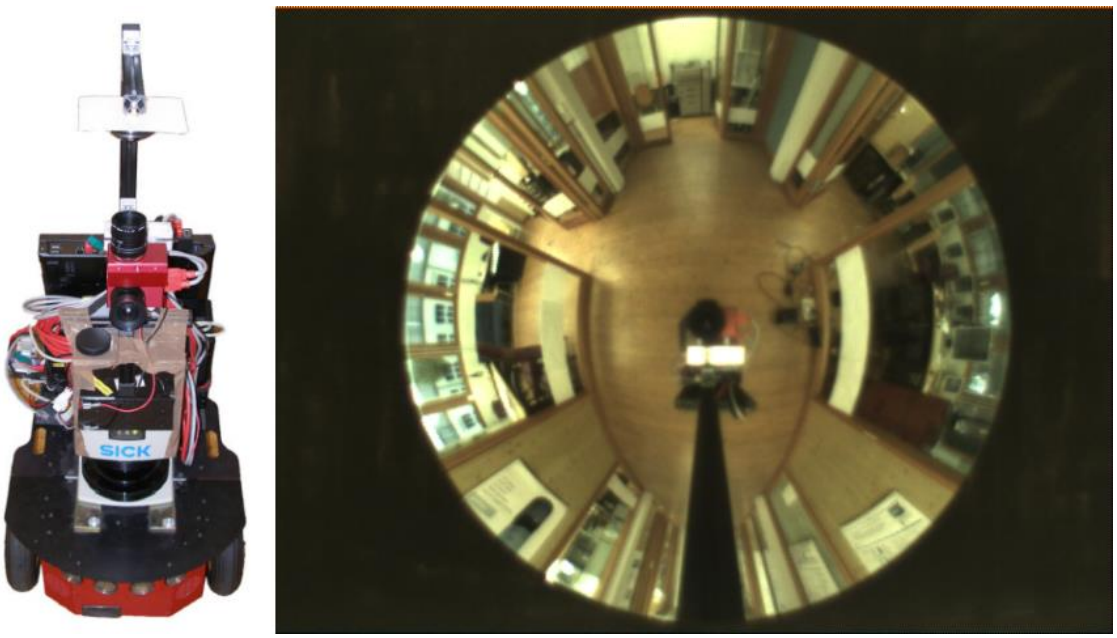


Figura 2.1: Plataforma móvil empleada para la adquisición de las imágenes por el laboratorio de Friburgo.

2.2. Localización jerárquica

El uso de información visual se lleva a cabo comúnmente mediante descriptores, es decir, métodos que extraen la información más relevante de la imagen.

Dentro de esta técnica hay dos grandes vertientes: por un lado, la utilización de descriptores de apariencia local, que se basan en la extracción de puntos, objetos o regiones características de la escena y obtienen un vector de información por cada punto seleccionado dentro de esta [17]; y, por otro lado, la utilización de descriptores de apariencia global o descriptores holísticos, que se basan en la creación de un único vector por imagen que contiene información de toda la escena [8].

El hecho de tener un único descriptor hace que los métodos de *mapping* y localización sean sencillos, basados muchas veces en la comparación de pares de imágenes. Asimismo, la aplicación a entornos estructurados es más directa y su coste computacional suele ser menor. A su vez, los descriptores de apariencia global se pueden subdividir en dos ramas en función de la forma en que pueden ser calculados. En primer lugar, tenemos los descriptores holísticos basados en métodos analíticos, que se fundamentan principalmente en cálculos de gradientes de orientación y/o color de los diferentes píxeles que componen la imagen [19]. Y, en segundo lugar, obtenemos los descriptores holísticos basados en *deep learning*, que son obtenidos de las capas estructurales que forman las redes neuronales convolucionales y cuyo uso es bastante frecuente en la actualidad para resolver problemas de robótica móvil [22].

Asimismo, este tipo de localización consiste en crear un mapa compuesto por varias capas con una estructura jerárquica. Aquellas de alto nivel presentan una cantidad de información relativamente compacta, que permite una localización aproximada, pero rápida. Sin embargo, las capas de bajo nivel suelen tener más información y se utilizan para afinar la posición [7]. A continuación, se presentan algunas investigaciones que han llevado a cabo esta tarea.

Por un lado, podemos encontrar en el trabajo de Sarlin et al. [23] un enfoque de localización jerárquica basado en una CNN monolítica que predice simultáneamente características locales y descriptores globales para una localización precisa de 6 grados de libertad. Este enfoque jerárquico supone un importante ahorro de tiempo de ejecución y hace que el sistema implementado sea apto para funcionar en tiempo real. Además, el método consigue una notable solidez en la localización aun cuando se producen grandes variaciones en la apariencia del entorno.

Seguidamente, hallamos en la investigación de Jiang et al. [15] una propuesta de un método sencillo, pero eficaz, denominado LayerCAM. Este puede producir mapas fiables de activación de clase para diferentes capas de la CNN. Esta propiedad permite recoger información de la localización de objetos

desde niveles gruesos (localización espacial aproximada) hasta niveles finos (detalles precisos a nivel de píxel). Además, son integrados en un mapa de activación de clase de alta calidad, en el que los píxeles relacionados con el objeto pueden destacar mejor sobre el entorno. Asimismo, para evaluar la calidad de este tipo de mapas producidos por LayerCAM, estos son aplicados a la localización de objetos débilmente supervisada y a la segmentación semántica. Como resultado de esta experimentación, se demuestra que los mapas de activación de clase generados por este método son más eficaces y fiables que los implementados por los métodos de atención ya existentes.

También encontramos en el trabajo de Xu et al. [30] la propuesta de dos métodos que adaptan las técnicas de recuperación de la imagen (*image retrieval*) utilizadas para el reconocimiento visual de lugares a la formulación de estimación de estado bayesiana para la tarea de localización. Con ello obtienen mejoras significativas en la precisión de la pose correspondiente a la etapa de localización gruesa. Además, se mantiene la precisión obtenida en investigaciones previas aun cuando el cambio de la apariencia del entorno es sustancialmente notable. Por tanto, esta mejora de la estimación inicial de la pose para localizar secuencias de imágenes abre la posibilidad de mejorar el rendimiento general de dicha tarea.

Por otro lado, se utiliza también la CNN como modelo jerárquico con los siguientes objetivos: (a) abordar la localización aproximada mediante el método conocido como *room retrieval* o recuperación de la estancia (capa de alto nivel), en el que se parte de la imagen de test, y (b) obtener descriptores holísticos de las imágenes de entrada. Los descriptores de las imágenes de entrenamiento formarán la capa de bajo nivel junto con los descriptores holísticos de las imágenes de test (también obtenidos mediante la CNN), por lo que se permite así resolver una localización fina mediante el método conocido como *image retrieval* o recuperación de la imagen.

Tal como podemos ver en la investigación de Cabrera et al. [3], se resolvió la tarea de localización jerárquica de la siguiente forma: en primer lugar (etapa de localización gruesa), se introduce una imagen de test en la red neuronal

convolucional y se estima, a partir de la información de las capas de salida, la estancia más probable en la que se capturó dicha imagen. Posteriormente, tras identificar la habitación, se lleva a cabo una localización más precisa (etapa de localización fina). Al ser la CNN capaz de proporcionar descriptores holísticos a partir de sus capas intermedias, en este paso se selecciona el descriptor de la imagen de test y se coteja con los descriptores del conjunto de datos de entrenamiento que pertenecen a la estancia seleccionada previamente. Por último, se almacena el descriptor con mayor similitud para así estimar la localización de la imagen de test en las coordenadas en las que se capturó dicha imagen de entrenamiento.

Sin embargo, hay que tener en cuenta que la necesidad de disponer de un *dataset* con un elevado número de imágenes y que, al mismo tiempo, presente todos los efectos visuales que se pueden producir durante el proceso de localización –cambios de iluminación, oclusiones debidas a objetos o personas, cambios en la orientación del robot, etc.– es uno de los factores más importantes para poder llevar a cabo de manera correcta el entrenamiento de la CNN. Si no se dispone de un *dataset* con dichas características, una de las técnicas propuestas es realizar un aumento de datos (*data augmentation*).

Esta técnica consiste en aplicar efectos visuales sobre las imágenes originales del conjunto de datos de entrenamiento. Asimismo, en la bibliografía podemos encontrar que las técnicas tradicionales de aumento de datos consideran algunas alteraciones en las imágenes, como pueden ser las siguientes: giros, traslaciones a lo largo de los ejes horizontal y vertical, rotaciones puras de los píxeles de la imagen, escalados o recortes [14], [6]. De la misma manera, cabe destacar que el aumento de datos ayuda a evitar que la red neuronal convolucional se sobreajuste y memorice los detalles exactos de las imágenes de entrenamiento.

Por último, antes de terminar el capítulo, se quiere dejar claro que como disponer de un gran conjunto de datos de entrenamiento es crucial para el rendimiento del modelo, y nuestro *dataset* disponible es más pequeño de lo necesario, hemos tenido que implementar la técnica de aumento de datos para

que el modelo pudiese ser entrenado adecuadamente y que así alcanzara la solución deseada. Además, el presente trabajo únicamente se centrará en la localización gruesa del entorno, por lo que esta tarea corresponderá solamente a la capa de alto nivel, y no se presentarán resultados pertinentes acerca de la localización fina en la estancia seleccionada.

Capítulo 3
Material y métodos

En este capítulo se expondrá de forma detallada, por una parte, la información más relevante acerca de la base de datos empleada junto a los materiales que han sido utilizados, así como también la estructura de la red neuronal convolucional que ha sido usada para la realización de este trabajo fin de máster. Y, por otra parte, se desarrollará con más detalle la técnica conocida como *data augmentation* o aumento de datos.

3.1. Base de datos empleada

Por un lado, como ya vimos en la figura 2.1, tenemos la plataforma robótica móvil empleada para la adquisición de imágenes en el laboratorio de Friburgo, espacio en el cual se basa nuestro estudio. Esta estaba equipada con escáneres láser tipo SICK, que proporcionan una medida de distancia escalar del sensor al objeto con una precisión y velocidad de escaneo óptima; y encoders en las ruedas también de tipo SICK, que miden el giro de las mismas y permiten calcular la posición y velocidad del robot móvil, lo cual hace posible una correcta odometría. Estos sensores posibilitan la obtención del *ground truth* (posición real) del robot en cada instante; sin embargo, aunque la información de estos pueda ser tomada a efectos comparativos, en el presente trabajo el problema de localización se resolverá considerando únicamente información visual. Además, durante su estancia en el laboratorio, el robot estaba controlado manualmente mediante un joystick.

La configuración de la cámara se creó mediante dos cámaras digitales *Videre Design* MDCS2; una para imágenes perspectiva y otra para imágenes omnidireccionales. Los parámetros detallados y las configuraciones de las

cámaras se pueden observar en la tabla 3.1. El sistema de visión omnidireccional catadióptrico fue construido usando un espejo hiperbólico. Las dos cámaras y el espejo se montaron juntos en un soporte portátil tal como se mostró en la figura 2.1 [27].

Robot móvil modelo ActivMedia Pioneer-3 Friburgo		
Tipo de cámara	Perspectiva	Omnidireccional
Velocidad de fotogramas	5 imágenes / segundo	
Resolución	640 × 480 píxeles, Patrón de Bayer	
Exposición	Automática	
Campo de visión	68.9º × 54.4º	----
Altura de la cámara	66 cm	91 cm

Tabla 3.1: Parámetros y configuración de las cámaras.

Por otro lado, la base de datos con la que se ha decidido trabajar es la denominada como *The COLD Database* [27]. Esta base de datos ofrece tres entornos de trabajo diferentes: el *Autonomus Intelligent Systems Laboratory* en la universidad de Friburgo, Alemania; el *Visual Cognitive Systems Laboratory* en la universidad de Ljubljana, Eslovenia; y el *Language Technology Laboratory* en el Centro Alemán de Investigación en Inteligencia Artificial de Saarbrücken, Alemania. Sin embargo, de una forma más conocida, estos tres *dataset* que componen nuestra base de datos son denominados por el nombre de la ciudad donde la adquisición de las imágenes fue realizada (COLD – Saarbrücken, COLD – Friburgo y COLD – Ljubljana). La base de datos COLD es un banco de pruebas ideal, compuesto por imágenes omnidireccionales, para evaluar la robustez de los algoritmos de localización y *mapping*. Esta ofrece no solo cambios de iluminación, ya que la apariencia visual de las distintas habitaciones, en especial de las regiones que se encuentran cercanas a las ventanas, son fuertemente afectadas por las condiciones lumínicas y climáticas; sino también cambios dinámicos como la actividad humana, la variación en la distribución y aspecto del mobiliario, etc.

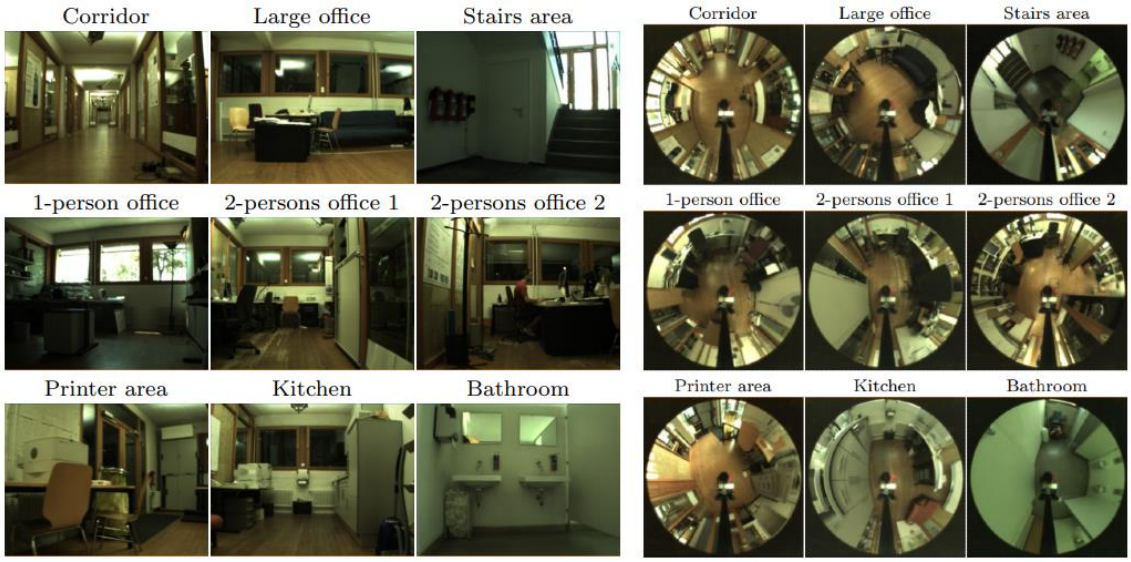
La elección del conjunto de imágenes a utilizar para la realización de este estudio ha sido el denominado como COLD – Friburgo. Estas se adquirieron en diferentes condiciones de iluminación, en un lapso de tiempo que varía en torno a dos/tres días. Las distintas estancias que recorrió el robot móvil para la adquisición de las imágenes se muestran en la tabla 3.2. Asimismo, un ejemplo de las imágenes capturadas durante este proceso puede ser visualizado en la figura 3.1, tanto en formato perspectiva como en omnidireccional. Como se puede observar, las habitaciones tienen diferentes propósitos y por ello estarán organizadas de una forma distinta, por lo que no serán afectadas por la actividad humana de la misma manera. A su vez, este *dataset* contiene también imágenes difuminadas debido al hecho de capturar las instantáneas en movimiento (efecto *blur*) y que por ende proporcionan menos información con respecto a la posición de adquisición en la que fueron tomadas. Todas estas desventajas hacen que este conjunto de imágenes sea adecuado para llevar a cabo experimentos en condiciones reales de funcionamiento, característica ideal para la realización de este estudio.

Estancias COLD – Friburgo	
Printer area	2 - persons office 1
Corridor	2 - persons office 2
Kitchen	1 - person office
Large office	Bathroom
Stairs area	

Tabla 3.2: Estancias recorridas por el robot móvil, en el laboratorio de Friburgo, mientras se realizaba la adquisición de las imágenes.

En la figura 3.2 se muestran los mapas generales de los entornos de interior de las dos partes del laboratorio que fueron consideradas por separado. Estas pueden verse como dos entornos individuales similares, los cuales están denotados como ‘Parte A’ y ‘Parte B’. Como se puede comprobar, cada una de las partes está compuesta por dos rutas distintas; una de color rojo, en la que el recorrido del robot móvil es el más extenso, obteniendo así un *dataset* mucho

más amplio, y otra de color azul, que aporta el recorrido más corto al no incluir todas las estancias del laboratorio.

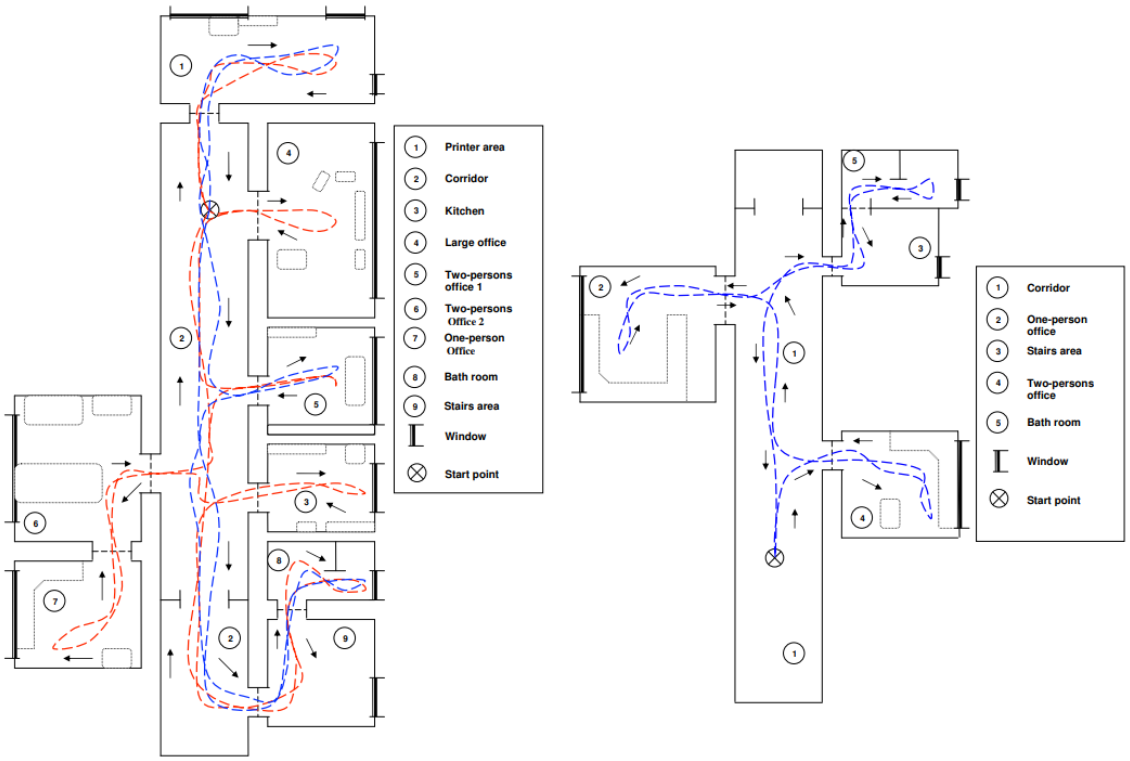


(a) (b)

Figura 3.1: Imágenes adquiridas en el laboratorio de Friburgo. La figura (a) muestra las imágenes tomadas mediante la cámara perspectiva, mientras que la figura (b) lo hace mediante la cámara omnidireccional. Imágenes obtenidas de COLD *database* [27].

Se ha tomado la decisión de trabajar con la 'Parte A' del mapa del laboratorio de Friburgo, ya que era el único de los dos que contenía información acerca de los tres tipos de iluminación que necesitábamos: nublado, soleado y noche. Asimismo, se ha tomado la trayectoria de color rojo, la cual nos proporciona información acerca de todas las estancias del entorno. Por otro lado, cabe señalar que la velocidad media de este robot móvil es de 0.3 m/s y la distancia que obtenemos entre imágenes, teniendo en cuenta que este está continuamente adquiriendo información a una velocidad de 5 fotogramas por segundo, es de aproximadamente 6 cm.

La base de datos COLD también nos ofrece la estimación de la pose del robot mediante un sensor láser, tanto la posición 'x - y' como la orientación, para cada imagen adquirida durante el recorrido.



(a) Map of Freiburg Portion A (b) Map of Freiburg Portion B

Figura 3.2: Recorrido del robot móvil en el laboratorio de Friburgo. Imágenes obtenidas de COLD *database* [27].

El primer paso para poder trabajar con nuestra base de datos es tratar las imágenes de COLD – Friburgo, para ello se hará una conversión de formato, pasando de imagen omnidireccional a panorámica, pero siempre respetando los canales de color de la misma. Esta conversión se realiza tomando la imagen omnidireccional en color (RGB) que obtenemos directamente del *dataset* de Friburgo, la cual se presenta en la figura 3.3, para posteriormente recortarla y eliminar toda la información posible de los bordes de la imagen que no sea de interés, quedando esta tal como podemos observar en la figura 3.4. Seguidamente, se ha implementado una función que transforma la imagen omnidireccional recortada en color en una imagen panorámica RGB, únicamente introduciendo como parámetros de entrada la imagen omnidireccional recortada en color y el centro de esta, el cual será el mismo para todas las imágenes a convertir en este trabajo. El resultado final de la conversión de la imagen queda reflejado tal como vemos en la figura 3.5.

Figura 3.3: Imagen omnidireccional RGB.

Figura 3.4: Imagen omnidireccional recortada RGB.

Figura 3.5: Imagen panorámica RGB.

Una vez se ha realizado esta conversión para todas las imágenes de nuestra base de datos –nublado, soleado y noche (imágenes test)–, obtendremos un cuarto *dataset* de entrenamiento, el cual será el utilizado para llevar a cabo la tarea de *mapping*, y con el que compararemos las imágenes test en nuestra tarea de localización. Para la creación de este modelo se decidió coger una de cada cinco imágenes del *dataset* adquirido durante un día nublado, ya que este es el que resulta menos afectado por los cambios de iluminación del entorno, de este modo nuestro modelo está creado por la adquisición de una imagen cada 30 cm, aproximadamente. La tabla 3.3 muestra la información acerca de los detalles de los cuatro *datasets* con los que realizaremos nuestros experimentos en el capítulo 4.

Dataset	**Número de imágenes**
Test (nublado)	2778
Test (soleado)	2807
Test (noche)	2707
Entrenamiento (nublado)	556

Tabla 3.3: Número de imágenes de cada *dataset*, tanto de entrenamiento como de test, creado a partir del entorno COLD – Friburgo.

3.2. Red neuronal utilizada

Las redes neuronales convolucionales, comúnmente conocidas como CNN, son actualmente la herramienta más popular entre las técnicas de aprendizaje profundo, ya que han dado resultados exitosos en muchas aplicaciones prácticas. Estas son un tipo especializado de red neuronal para procesar datos que presentan una topología ya conocida, siendo comúnmente diseñadas para recibir imágenes como parámetros de entrada y teniendo diferentes aplicaciones como la clasificación o la detección de objetos.

Las CNN consisten en conexiones locales entre neuronas y transformaciones jerárquicamente organizadas de los datos. Básicamente, estas redes están compuestas principalmente por tres tipos de capas: convolucionales

(conv), de agrupación (*pooling*) y completamente conectadas –*fully connected* (fc)–. Cada una de estas transforma la entrada y genera una salida de acuerdo con los parámetros establecidos. Este proceso de conexión se aborda mediante la transferencia de información a través de todas sus capas y finaliza en la última de ellas, la cual es una capa completamente conectada que genera un vector de características 1D, que proporciona la predicción más probable.

Sin embargo, la construcción y el entrenamiento de una red desde cero requiere tanto experiencia con arquitecturas de redes como una gran cantidad de datos para el entrenamiento y, por tanto, un tiempo de computación importante. Además, este trabajo continua la propuesta realizada en investigaciones anteriores [5]: adaptar y entrenar redes preexistentes con un objetivo distinto a aquel para el que inicialmente se diseñaron. En este sentido, se propone partir de la CNN *Places* [31], ya que presenta una arquitectura sencilla y ha sido utilizada con éxito en trabajos anteriores, tal como podemos ver en las investigaciones de Céspedes [8]. Por tanto, a continuación, se hará una breve explicación del origen de dicha red neuronal convolucional para que así pueda ser entendida tanto la procedencia de la misma como la finalidad actual que posee.

Por un lado, la CNN *AlexNet* fue presentada por Krizhevsky et al. [16]. Esta red fue entrenada con alrededor de 1.2 millones de imágenes para clasificar 1000 posibles tipos de objetos –asimismo, de esta red neuronal convolucional fue creada la conocida CNN *Caffe*–. *AlexNet* consta de 25 capas diferentes, entre ellas una capa de entrada, la cual toma una imagen RGB de 227 × 227 × 3, por lo tanto, cada imagen se debe normalizar antes del entrenamiento y/o clasificación; ocho capas intermedias, cinco de ellas convolucionales y tres completamente conectadas; tres capas de agrupación; un *softmax* final de 1000 categorías posibles de objetos como, por ejemplo, un teclado, un bolígrafo, una mesa o una variedad de animales; y una capa de salida, que tiene la función de indicar qué clase de objeto tenemos como imagen.

Por otro lado, la red neuronal convolucional *Places*, que presenta una arquitectura similar a la ya vista en la CNN *AlexNet*, fue entrenada con alrededor

de 2.5 millones de imágenes para clasificar 205 posibles tipos de escenas. En la figura 3.6 hemos tratado la arquitectura de *Places* de una forma detallada para que con esta información se pueda entender de forma correcta cómo está organizada internamente. Asimismo, esta red nace de la CNN *Caffe* –y de ahí la importancia de haber comenzado la explicación con *AlexNet*–, que fue entrenada para clasificar diferentes tipos de objetos, por lo que ambas comparten la misma arquitectura de red y tienen fortalezas complementarias en tareas centradas tanto en escenas como en objetos. La red neuronal convolucional *Places* está compuesta por una capa de entrada, la cual toma una imagen RGB de tamaño 227 × 227 × 3, por lo tanto, cada imagen se debe normalizar antes del entrenamiento y/o clasificación; cinco capas convolucionales, las cuales tienen una función de aprendizaje o caracterización; tres capas totalmente conectadas, que tienen una labor de clasificación; tres capas de agrupación; una capa *softmax*, que indica la probabilidad de que la imagen de entrada corresponda a cada una de las 205 estancias posibles; y una capa de salida, la cual otorgará el lugar probable en el que se haya tomado la imagen.

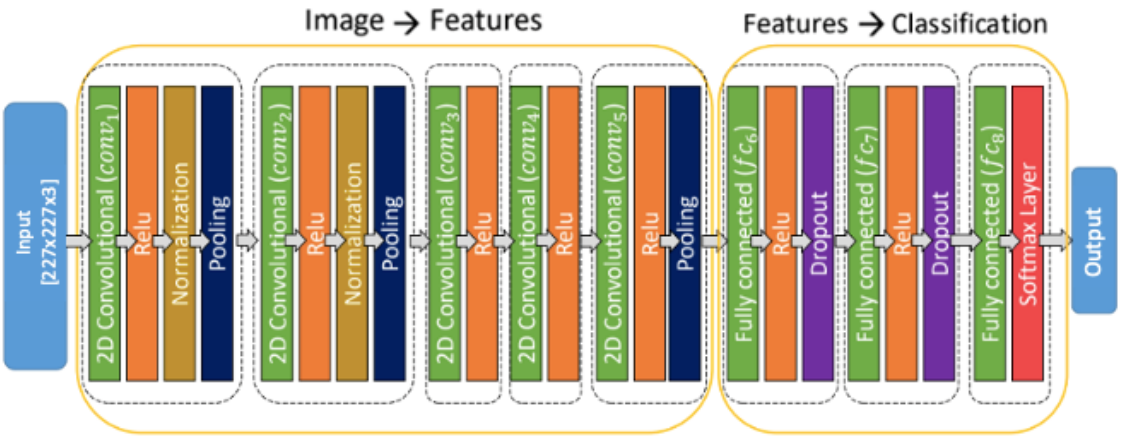


Figura 3.6: Diseño de la arquitectura de la red neuronal convolucional *Places*.

Sin embargo, para llevar a cabo las investigaciones del presente trabajo, debemos realizar tanto la modificación de algunas capas de la red *Places* como un entrenamiento completo desde cero para adaptar así la arquitectura de esta red a la tarea de clasificación de estancias propuesta. Para comenzar, se redimensiona la capa de entrada, la cual inicialmente es de 227 x 227 x 3, ya que el tamaño correspondiente que poseen las imágenes panorámicas con las

que vamos a trabajar está establecido en 128 x 512 x 3. Seguidamente, se sustituyen las tres últimas capas para adaptar la red a una tarea de clasificación de habitaciones. Estas capas son las siguientes: la capa totalmente conectada (fc8), la capa *softmax* y la capa de clasificación. En primer lugar, la capa fc8 se readapta para dar salida a un vector de nueve componentes (las posibles 9 estancias del entorno). En segundo lugar, las capas *softmax* y de clasificación se reajustan para determinar, respectivamente, las probabilidades entre nueve categorías de clases (clasificación en una de las 9 estancias que contiene el entorno de destino).

No obstante, cabe remarcar las siguientes cosas a tener en cuenta: redimensionar las imágenes panorámicas de entrada a un tamaño de 227 × 227 evitaría empezar el entrenamiento desde cero, ya que se podrían aprovechar los pesos iniciales de *Places*, pero redimensionar las imágenes panorámicas cambiaría bruscamente su aspecto y afectaría significativamente al rendimiento de la red. De la misma forma, a la hora de readaptar la red neuronal convolucional, se debe saber que la capa totalmente conectada (fc6) debe ser, a menudo, reestablecida debido a que al cambiar el tamaño de la capa inicial de la CNN esta necesita volver a ser llamada durante el proceso.

Tras estos cambios de capas, la red está lista para ser entrenada con el conjunto de imágenes panorámicas de entrenamiento para así resolver la tarea de localización mediante el problema de recuperación de la estancia (*room retrieval*), es decir, recuperar la habitación donde se obtuvo la imagen de entrada.

3.3. Data augmentation

Se debe tener en cuenta que disponer de un gran conjunto de datos de entrenamiento es crucial para el rendimiento del modelo. Sin embargo, a veces, el conjunto de datos disponible es más pequeño de lo necesario y, entonces, el modelo no puede ser entrenado adecuadamente para alcanzar la solución deseada [3]. Por tanto, para resolver este problema y mejorar así el rendimiento de la red neuronal convolucional, se ha propuesto aumentar el número de

instancias de entrenamiento. Para ello, se crearán nuevas muestras de información (imágenes) aplicando diferentes efectos sobre las imágenes originales del entorno, ya que, al considerar dichos efectos, el modelo de aprendizaje profundo alcanza una mayor robustez frente a ellos.

Es por ello por lo que, en el presente trabajo, el aumento de datos se ha diseñado específicamente para obtener una CNN robusta para la tarea de localización. Por lo tanto, para obtener nuevas muestras, consideramos una variedad de efectos visuales para cada imagen de entrenamiento, que pueden ocurrir realmente cuando el robot opera en condiciones reales de funcionamiento. Por lo tanto, a través de este *data augmentation*, se espera que la red neuronal convolucional sea capaz de hacer frente a las difíciles condiciones que pueden darse en el entorno en el que se mueve el robot.

Teniendo en cuenta esto, los efectos que han sido considerados para realizar el aumento de datos son los mostrados a continuación. No obstante, cabe recalcar que los dos primeros corresponden a efectos locales, por lo que únicamente se han hecho variaciones de la imagen en determinadas zonas de la misma, mientras que los cuatro últimos atañen a efectos globales, es decir, se han producido cambios en la totalidad de los píxeles de la escena.

- **Foco de luz / Foco de sombra**. Las formas circulares, como las bombillas, son habituales en la adquisición de datos. Por tanto, para simular esta fuente de luz, se edita la intensidad de las distintas regiones siguiendo dicha forma. Para ello, se aumenta el valor del píxel para reproducir una intensidad más brillante (foco de luz) o se disminuye para simular un resultado de sombra (foco de sombra). Para reproducir un efecto de desvanecimiento realista, la intensidad se reduce gradualmente desde el centro del foco hacia el borde del mismo como consecuencia de la atenuación de la luz. El tamaño de la forma y la posición del foco se seleccionan aleatoriamente para simular el efecto de diferentes maneras, así como también el valor máximo para realizar la variación de la intensidad. Por último, no olvidar que estos dos efectos se aplican por separado a las

imágenes, sin embargo, para la explicación se han tomado como uno solo para así facilitar el proceso de entendimiento de los mismos.

- **Brillo**. Los valores de baja intensidad son incrementados para generar datos con más brillo y, por otro lado, los valores altos de intensidad son reducidos para crear información con menos brillo. Este efecto trata de imitar los cambios que pueden experimentar las condiciones de iluminación del entorno. Sin embargo, estas dos posibilidades no se aplican al mismo tiempo en la misma imagen.

- **Contraste**. El contraste de la imagen juega un papel muy importante a la hora de resaltar los diferentes objetos en la escena. Además, las imágenes de bajo contraste suelen tener un aspecto más suave, por lo que presentan menos sombras y reflejos.

- **Saturación**. La saturación del color de la imagen se refiere a la intensidad del color que ofrecen los diferentes píxeles. Cuanto menor sea la saturación, menos colorida será la imagen –incluso puede llegar a parecerse a una imagen en escala de grises si la saturación es muy baja–; por el contrario, se obtienen colores más vivos cuando el resultado de la saturación es alto. Asimismo, este efecto puede aparentar situaciones en las que la iluminación del entorno cambia significativamente.

- **Rotación**. Este efecto simula la situación en la que el robot está en la misma posición, pero la orientación en la que fue capturada la imagen es diferente. Además, la rotación no está relacionada con los efectos de iluminación que se han expuesto previamente, pero mejora mucho la base de datos al permitir obtener posibilidades reales de observar la misma escena desde una orientación distinta. Finalmente, recalcar que se ha aplicado a la imagen inicial rotaciones comprendidas entre 0 y 359 grados.

La imagen original de nuestro conjunto de datos de entrenamiento ha sido vista anteriormente tal como se presentó en la figura 3.5. Por tanto, a continuación, en la figura 3.7 se muestran los citados efectos que son aplicados a nuestro modelo visual para llevar a cabo la tarea de localización del robot móvil.

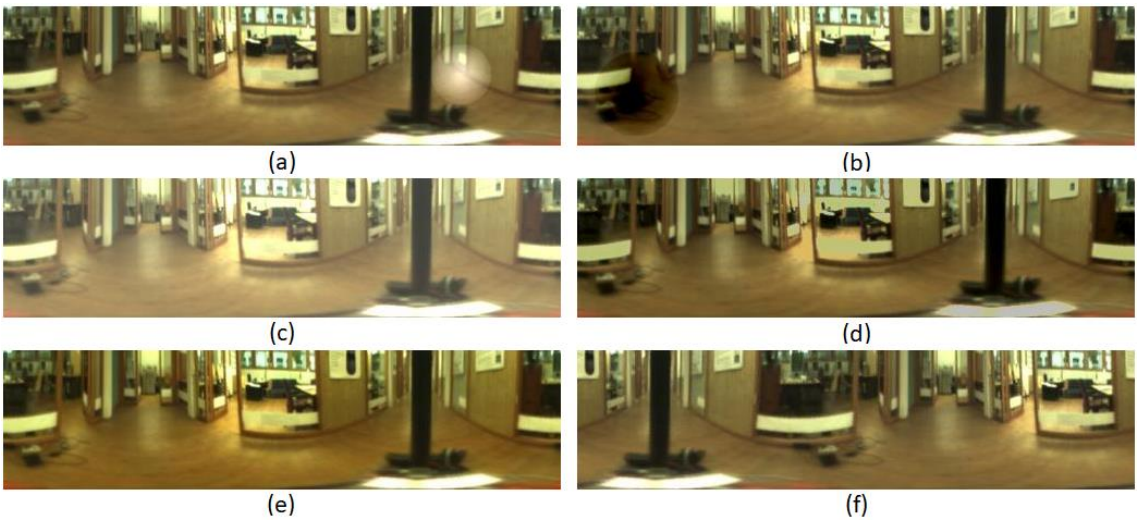


Figura 3.7: Ejemplo de aumento de datos. Los efectos estudiados son los siguientes: (a) foco de luz, (b) foco de sombra, (c) brillo, (d) contraste, (e) saturación y (f) rotación.

De la misma forma, la tabla 3.4 muestra la información acerca de los detalles de los siete *datasets* con los que realizaremos nuestros experimentos en el capítulo 4.

Dataset	**Número de imágenes**
Entrenamiento (nublado)	556
Aumento de datos (foco de luz)	3336
Aumento de datos (foco de sombra)	3336
Aumento de datos (brillo)	3892
Aumento de datos (contraste)	3336
Aumento de datos (saturación)	3336
Aumento de datos (rotación)	17236

Tabla 3.4: Número de imágenes de cada *dataset*, tanto de entrenamiento como de aumento de datos, creado a partir del entorno COLD – Friburgo.

Capítulo 4
Resultados

En este capítulo se estudia la eficacia de nuestra red neuronal convolucional –readaptada y reentrenada a partir de la CNN *Places*– en la tarea de localización de un robot móvil mediante la comparativa entre las múltiples estancias del *dataset* de Friburgo seleccionado para este trabajo. Por tanto, para la realización de este estudio, se han hecho seis experimentos para analizar la conveniencia de utilizar diferentes efectos visuales para llevar a cabo un aumento de datos y un posterior entrenamiento de la red neuronal convolucional. Por un lado, tenemos los efectos correspondientes a cambios de variación en la iluminación del entorno, como son el foco de luz, el foco de sombra, el brillo, el contraste y la saturación. Y, por otro lado, el efecto correspondiente a una posible rotación de la imagen debido a la orientación con la que el robot hubiese tomado la instantánea.

El problema de localización se puede plantear de manera absoluta mediante la comparación de la imagen capturada por el robot móvil con las almacenadas en el modelo visual. En el presente trabajo estudiaremos el proceso de localización jerárquica como un problema de *room retrieval,* en el cual se obtiene la estancia que representa una mayor similitud en relación con la nueva imagen capturada. Asimismo, esto es llevado a cabo sin hacer uso ni de la información métrica relativa a la posición en el plano de trabajo en que se capturaron las imágenes del modelo ni de la información de la odometría interna del robot [19].

Además, para dicho propósito, previamente se ha tenido que realizar un proceso de *mapping* en el que el robot ha obtenido información visual del medio,

es decir, imágenes de entrenamiento capturadas desde diferentes estancias del entorno.

Una vez que se ha construido la representación del medio, es necesario dotar de funcionalidad a dicho modelo visual, comprobando si permite al robot móvil realizar correctamente el proceso de localización mediante la estimación de su posición a través de la comparación de la información visual que captura en un instante determinado con la información almacenada en el modelo. Sin embargo, en los entornos de interior que estudiaremos en este trabajo fin de máster, se plantea el problema fundamental de similitud visual de zonas que se encuentran geométricamente alejadas, pero que el robot móvil estima en una misma región; este problema se conoce como *visual aliasing*.

Una vez creado el modelo visual, la tarea de localización se resuelve mediante los siguientes pasos: (1) el robot captura una nueva imagen omnidireccional desde una posición desconocida, la cual denominaremos de test. (2) Seguidamente, esta imagen se transforma en panorámica. (3) A continuación, una vez que la conversión está disponible, se introduce la imagen de test a la CNN para estimar la habitación en la que esta fue capturada. Cabe indicar que se consideran tres condiciones de iluminación: nublado, noche y soleado.

El proceso de localización consiste en los siguientes puntos: en primer lugar, se lleva a cabo el entrenamiento de nuestra red neuronal convolucional, la cual fue readaptada tal como se explicó en la sección 3.2, con imágenes correspondientes al modelo visual –a este no se le había aplicado todavía ningún aumento de datos–. De esta manera, una vez terminado el entrenamiento, se realiza el testeo de la CNN reentrenada para clasificar las diferentes estancias con los tres tipos de iluminación propuestos. Es por ello por lo que podemos observar en la figura 4.1 los porcentajes de precisión obtenidos. A través de esta se puede ver que para ningún tipo de iluminación la red neuronal convolucional reentrenada alcanza porcentajes mayores al 68% de los valores, ya que se obtiene una precisión del 67.28%, 62.65% y 56.32% para los *datasets* de nublado, noche y soleado, respectivamente.

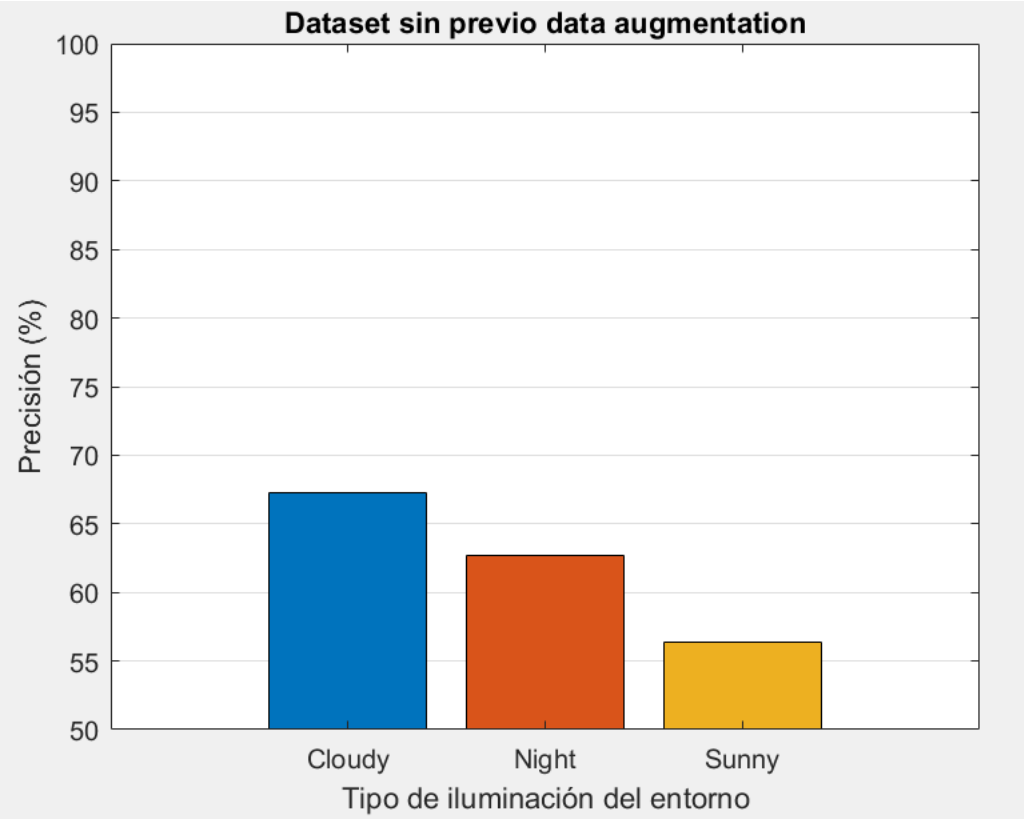


Figura 4.1: Precisión media del proceso de clasificación de estancias de la CNN reentrenada para llevar a cabo la tarea de localización gruesa sin aumento de datos.

Asimismo, se muestra en la figura 4.2 –para la condición de nublado–, en la figura 4.3 –para la categoría de noche– y en la figura 4.4 –para la condición de soleado– la matriz de confusión de los resultados de clasificación de estancias obtenidos tras finalizar el proceso de testeo. Tal como podemos ver, se aprecia claramente que la CNN reentrenada presenta grandes dificultades para realizar la clasificación correspondiente a las estancias de *Large office*, *Bathroom* y *Stairs area*.

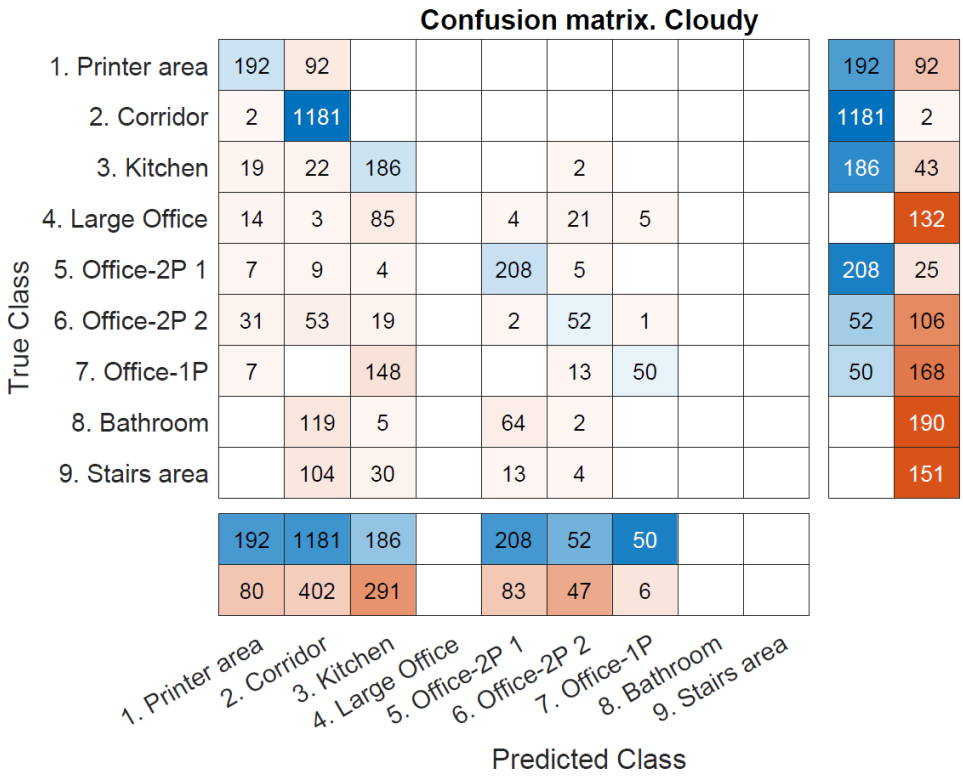


Figura 4.2: Matriz de confusión –clasificación de las estancias– para la condición lumínica de nublado de la CNN reentrenada sin aumento de datos.

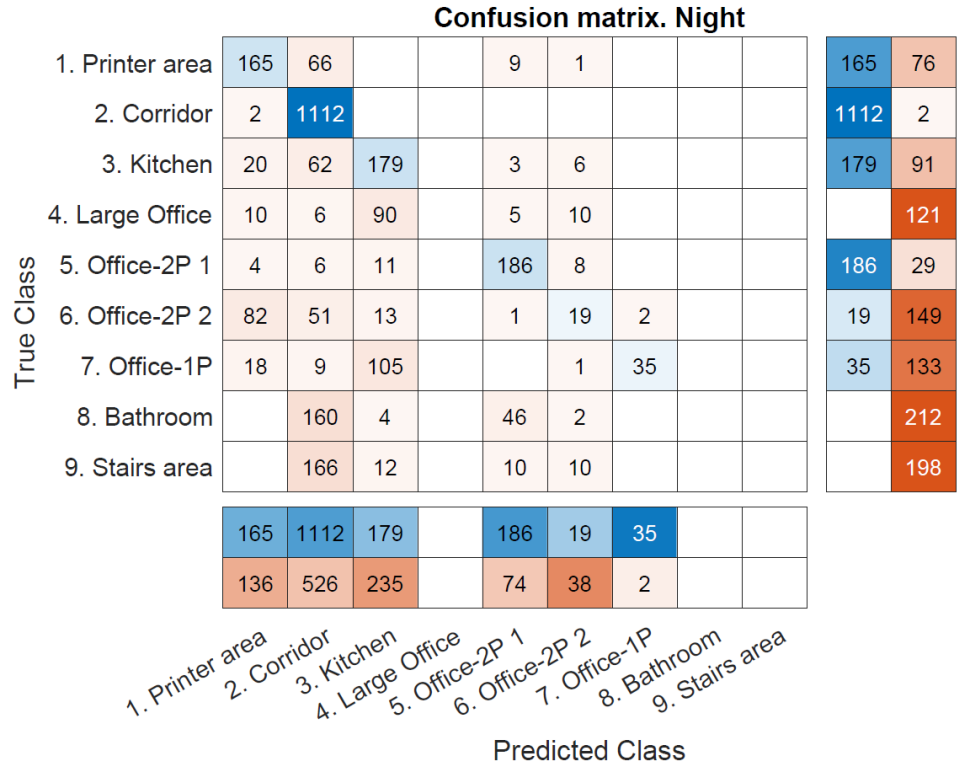


Figura 4.3: Matriz de confusión –clasificación de las estancias– para la condición lumínica de noche de la CNN reentrenada sin aumento de datos.

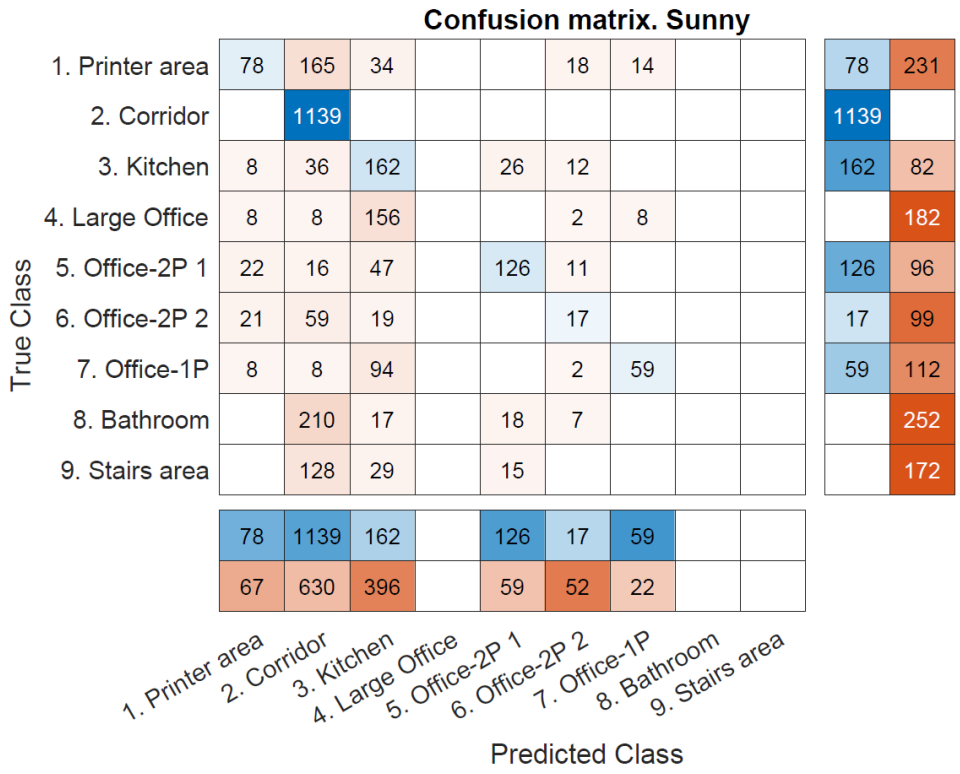


Figura 4.4: Matriz de confusión –clasificación de las estancias– para la condición lumínica de soleado de la CNN reentrenada sin aumento de datos.

Una vez presentados los resultados iniciales del modelo visual, se llevarán a cabo los seis experimentos propuestos para realizar su posterior comparación y discusión, con la intención de conocer qué aumento de datos presenta los mejores resultados en la CNN reentrenada para la tarea de localización de estancias.

4.1. Experimento 1: Efecto de foco de luz

En este primer experimento se realiza un aumento de datos del modelo visual, el cual es el correspondiente a aplicar un efecto de foco de luz en las imágenes originales. El resultado de este *dataset* es el que se presenta en la figura 4.5.

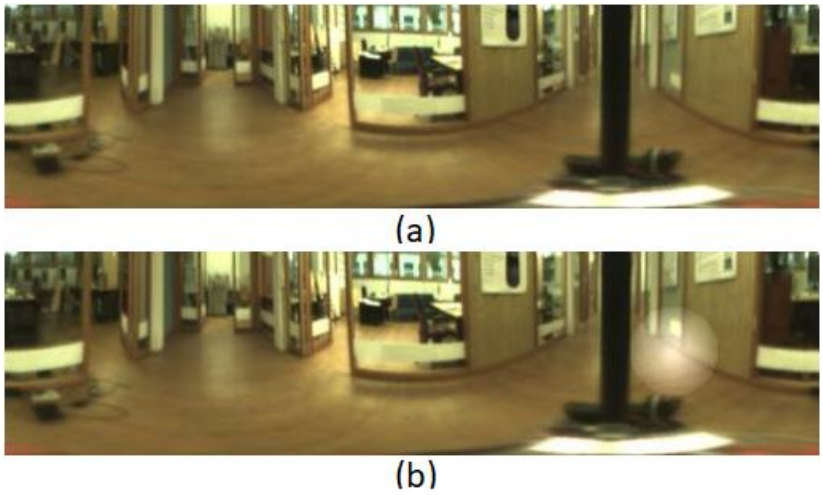


Figura 4.5: Ejemplo de aumento de datos. La figura (a) muestra la imagen original capturada dentro del entorno de Friburgo, mientras que la figura (b) es el resultado de aplicar sobre la primera el efecto de foco de luz.

Seguidamente, se lleva a cabo el entrenamiento de nuestra red neuronal convolucional, la cual fue readaptada tal como se explicó en la sección 3.2, con imágenes correspondientes a este aumento de datos. El proceso de entrenamiento de la CNN es el mostrado en la figura 4.6, que se expone como ejemplo únicamente en este experimento debido a que en el resto de ensayos las gráficas tienen una evolución similar.

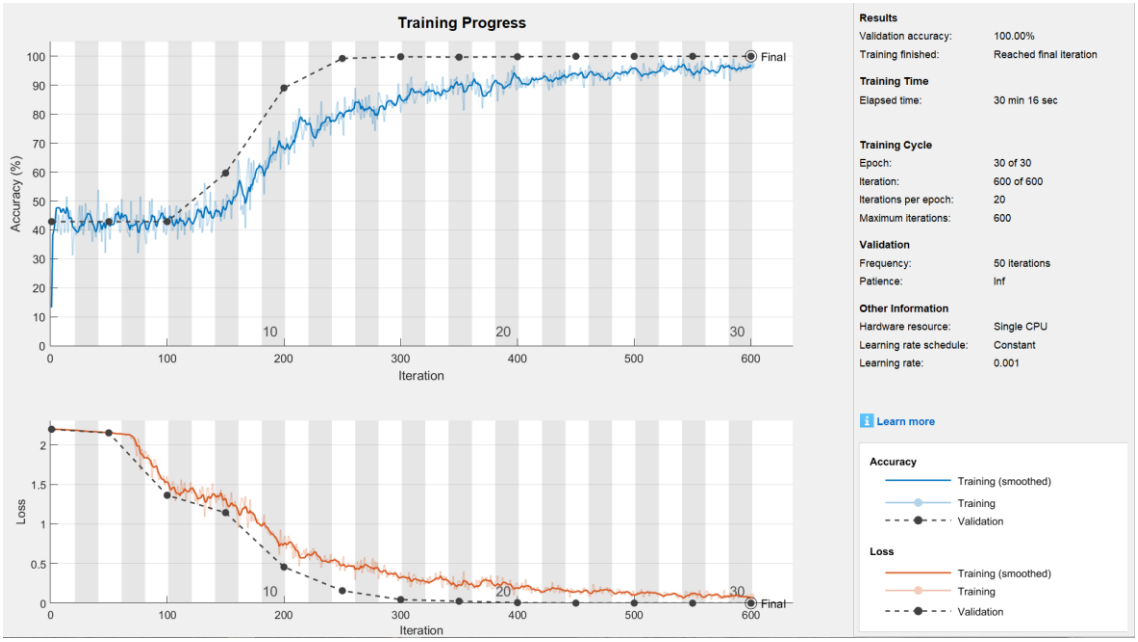


Figura 4.6: Proceso de entrenamiento de la CNN con readaptación de capas y con aumento de datos (en este caso, se ha aplicado el efecto de foco de luz).

De esta manera, una vez terminado el entrenamiento, se realiza el testeo de la CNN reentrenada para clasificar las diferentes estancias con las tres condiciones de iluminación propuestas. En la figura 4.7 se muestran los porcentajes de precisión obtenidos. Cabe recalcar que para cualquier tipo de iluminación la red neuronal convolucional reentrenada sobrepasa porcentajes mayores al 91% de los valores, ya que se obtiene una precisión del 99.17%, 97.16% y 91.38% para los *datasets* de nublado, noche y soleado, respectivamente. Por tanto, podemos ver cómo la precisión de los resultados alcanzados es significativamente mayor con cualquier tipo de iluminación en comparación con los resultados obtenidos cuando a la red neuronal convolucional no se le ha aplicado ningún aumento de datos.

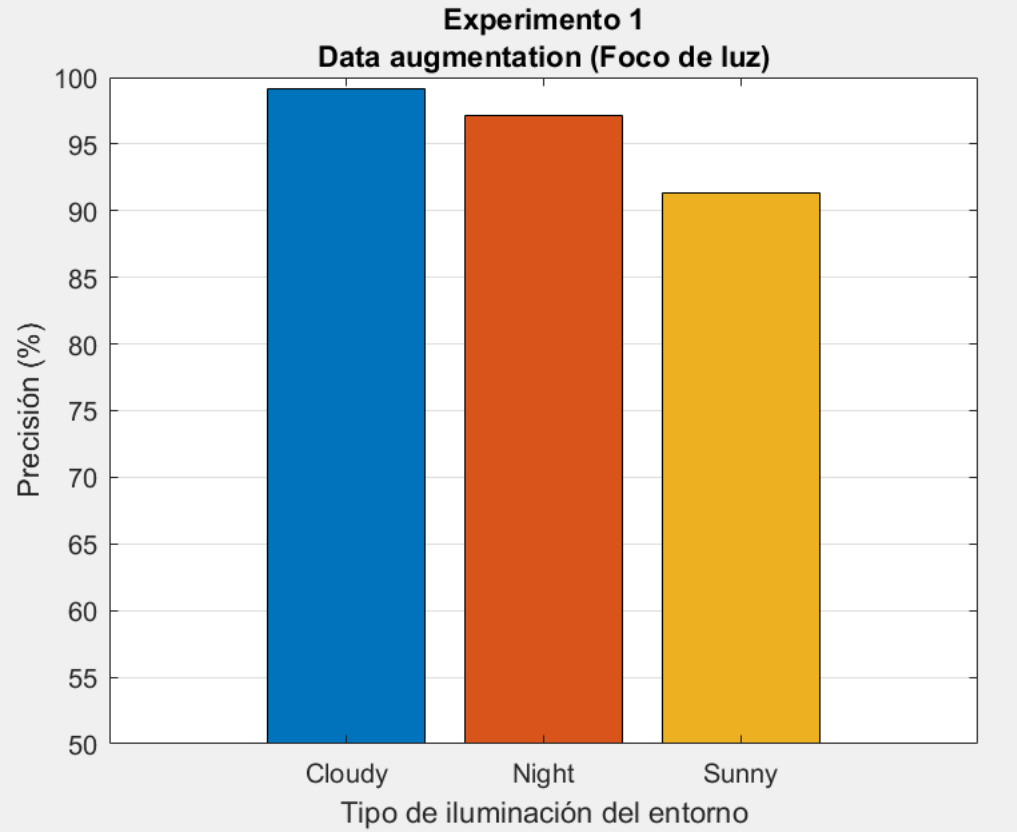


Figura 4.7: Precisión media del proceso de clasificación de estancias de la CNN reentrenada para llevar a cabo la tarea de localización gruesa con aumento de datos (efecto de foco de luz).

Asimismo, se muestra en la figura 4.8 –para la condición de nublado–, en la figura 4.9 –para la categoría de noche– y en la figura 4.10 –para la condición

de soleado– la matriz de confusión para la tarea de clasificación de estancias obtenida tras finalizar el proceso de testeo. Tal como podemos ver, se aprecia claramente que la CNN reentrenada es capaz de clasificar las imágenes correspondientes a todo tipo de estancias. A su vez, se puede observar cómo las matrices de confusión obtenidas son más robustas ante cualquier tipo de iluminación en comparación con las obtenidas para el testeo con la red neuronal convolucional entrenada sin *data augmentation* (figura 4.2, figura 4.3 y figura 4.4). Además, nótese que en este experimento la red neuronal convolucional reentrenada sí ha sido capaz de clasificar las imágenes correspondientes a las estancias de *Large office*, *Bathroom* y *Stairs area*.

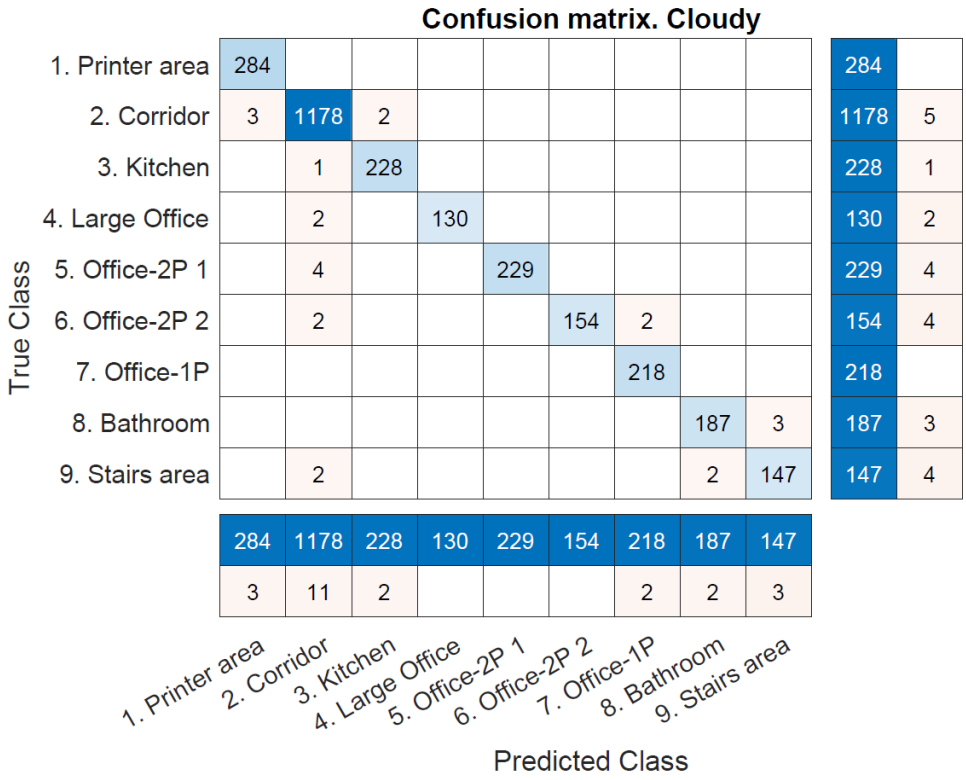


Figura 4.8: Matriz de confusión –clasificación de las estancias– para la condición lumínica de nublado de la CNN reentrenada con aumento de datos (efecto de foco de luz).

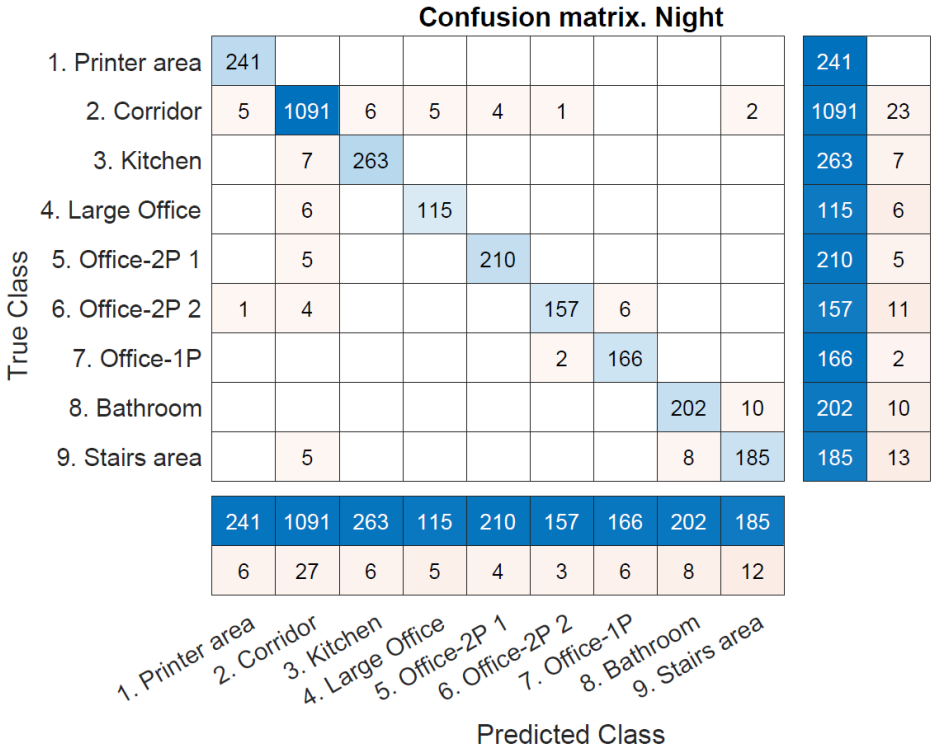


Figura 4.9: Matriz de confusión –clasificación de las estancias– para la condición lumínica de noche de la CNN reentrenada con aumento de datos (efecto de foco de luz).

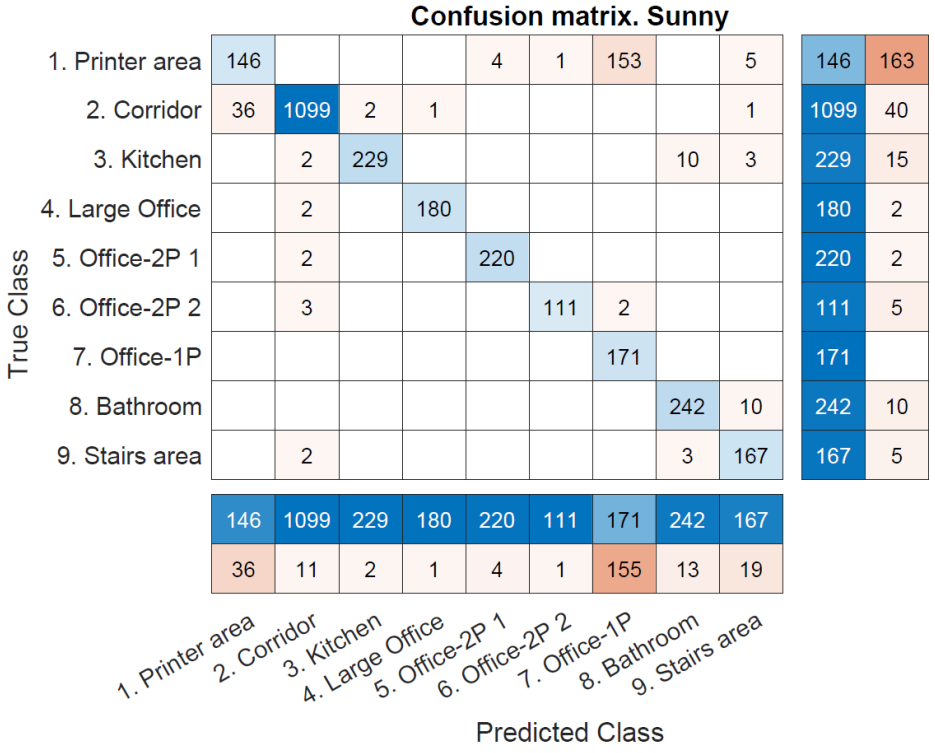


Figura 4.10: Matriz de confusión –clasificación de las estancias– para la condición lumínica de soleado de la CNN reentrenada con aumento de datos (efecto de foco de luz).

Tal como podemos ver en las matrices de confusión para la condición lumínica de nublado y de noche, no hay puntos de interés destacados debido a que la precisión media del proceso de clasificación de estancias ronda la excelencia. Por ello, nos centraremos en los resultados obtenidos para el tipo de iluminación de soleado, ya que podemos observar que la mayor confusión entre la clasificación de habitaciones viene dada por las correspondientes a *Printer area* y *1 - person office*. Esto es debido a que, aunque son estancias que se encuentran alejadas entre sí, comparten tanto una distribución como un mobiliario parecido, lo que puede suscitar a pensar que son el mismo entorno.

4.2. Experimento 2: Efecto de foco de sombra

En el siguiente experimento, al igual que en el anterior, se realiza un aumento de datos del modelo visual, el cual es el correspondiente a aplicar un efecto de foco de sombra en las imágenes originales. El resultado de este *dataset* es el que se presenta en la figura 4.11.

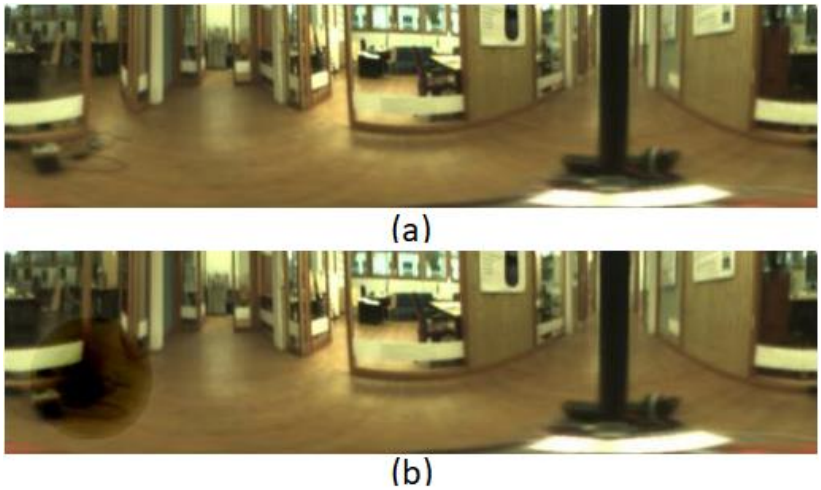


Figura 4.11: Ejemplo de aumento de datos. La figura (a) muestra la imagen original capturada dentro del entorno de Friburgo, mientras que la figura (b) es el resultado de aplicar sobre la primera el efecto de foco de sombra.

Seguidamente, se lleva a cabo el entrenamiento de nuestra red neuronal convolucional, la cual fue readaptada tal como se explicó en la sección 3.2, con imágenes correspondientes a este aumento de datos. De esta manera, una vez terminado el entrenamiento, se realiza el testeo de la CNN reentrenada para

clasificar las diferentes estancias con los tres tipos de iluminación propuestos. Es por ello por lo que podemos observar en la figura 4.12 los porcentajes de precisión obtenidos. A través de esta se puede ver que para cualquier tipo de iluminación la red neuronal convolucional reentrenada sobrepasa porcentajes mayores al 84% de los valores, ya que se obtiene una precisión del 99.03%, 97.04% y 84.54% para los *datasets* de nublado, noche y soleado, respectivamente. Por tanto, podemos ver cómo la precisión de los resultados alcanzados es significativamente mayor en cualquier tipo de iluminación en comparación con los hallados en la figura 4.1 –resultados sin previo aumento de datos–.

Además, podemos ver que los mejores resultados para el *dataset* de soleado se han obtenido en el experimento 1 –efecto de foco de luz–, ya que se ha alcanzado una precisión del 91.38% frente al 84.54% que logramos en este ensayo. Sin embargo, para los *datasets* restantes –nublado y noche– no se encuentran mejoras significativas.

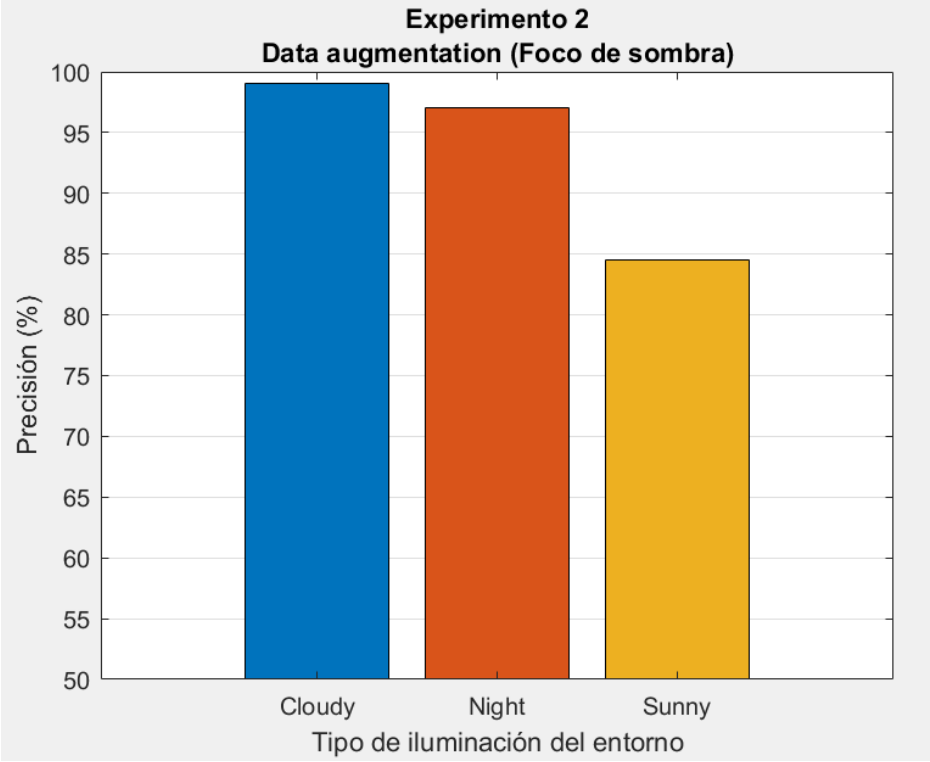


Figura 4.12: Precisión media del proceso de clasificación de estancias de la CNN reentrenada para llevar a cabo la tarea de localización gruesa con aumento de datos (efecto de foco de sombra).

Asimismo, se muestra en la figura 4.13 –para la condición de nublado–, en la figura 4.14 –para la categoría de noche– y en la figura 4.15 –para la condición de soleado– la matriz de confusión de los resultados de clasificación de estancias obtenidos tras finalizar el proceso de testeo. Tal como podemos ver, se aprecia claramente que la CNN reentrenada es capaz de clasificar las imágenes correspondientes a todo tipo de estancias. A su vez, se puede observar cómo las matrices de confusión obtenidas son más robustas ante cualquier tipo de iluminación en comparación con las graficadas en la figura 4.2, en la figura 4.3 y en la figura 4.4.

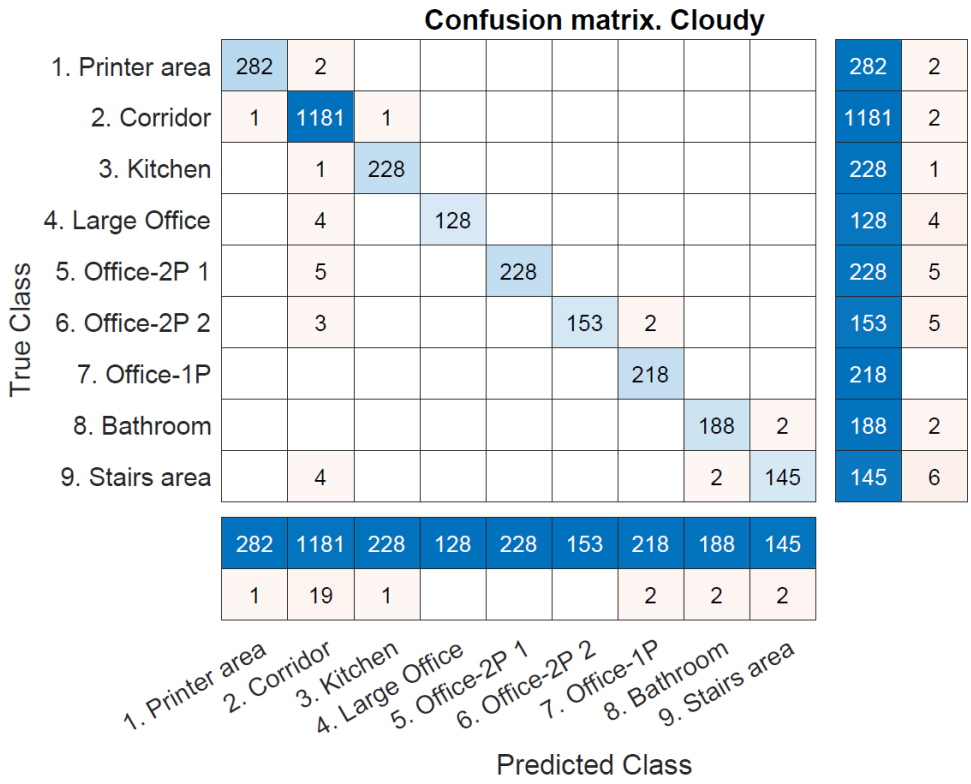


Figura 4.13: Matriz de confusión –clasificación de las estancias– para la condición lumínica de nublado de la CNN reentrenada con aumento de datos (efecto de foco de sombra).

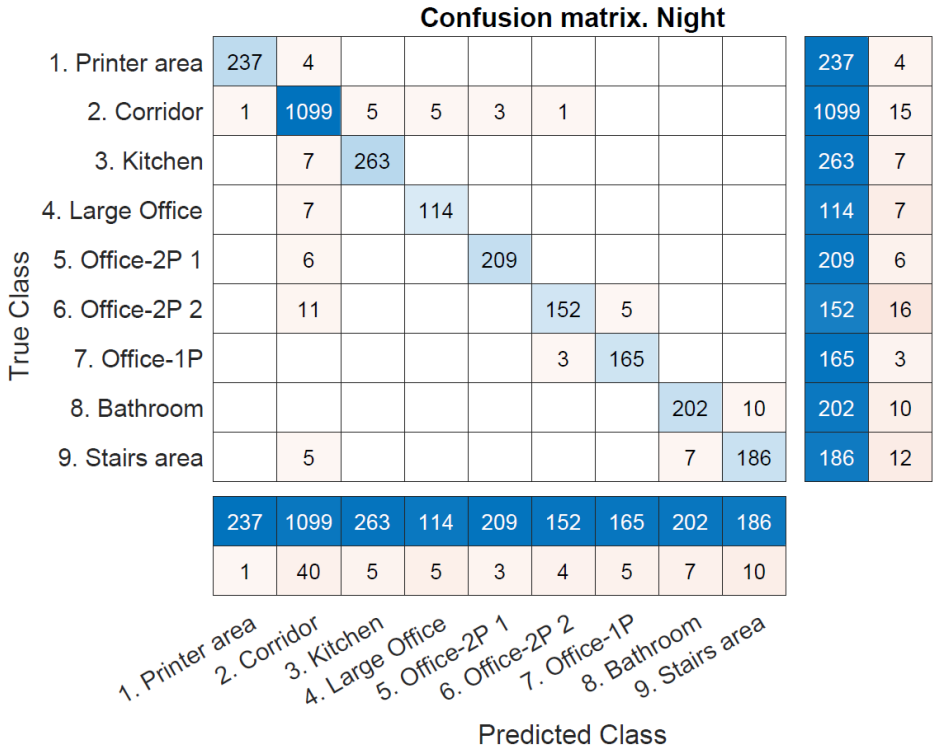


Figura 4.14: Matriz de confusión –clasificación de las estancias– para la condición lumínica de noche de la CNN reentrenada con aumento de datos (efecto de foco de sombra).

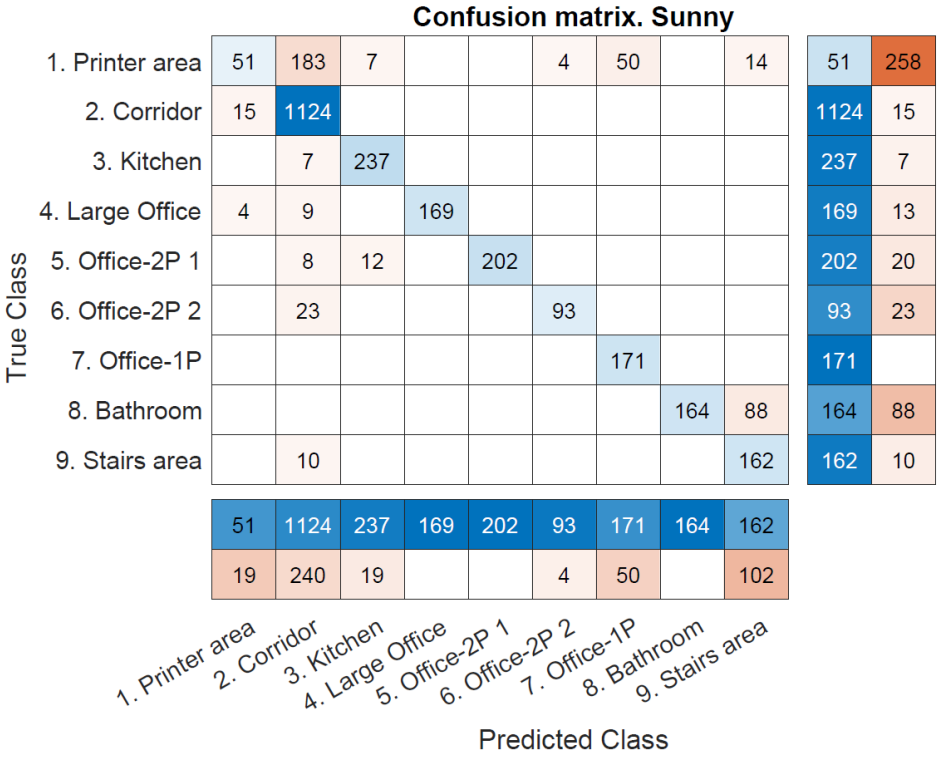


Figura 4.15: Matriz de confusión –clasificación de las estancias– para la condición lumínica de soleado de la CNN reentrenada con aumento de datos (efecto de foco de sombra).

Como podemos observar en las matrices de confusión para el tipo de iluminación de nublado y de noche, no hay información de interés destacada, ya que la precisión media del proceso de clasificación de estancias es realmente elevada. Por tanto, nos centraremos en los resultados obtenidos para la condición lumínica de soleado, ya que podemos ver que la mayor confusión entre la clasificación de habitaciones viene dada por las correspondientes a *Printer area* y *Corridor*. Esto se debe a que, además de ser estancias que se encuentran cercanas la una de la otra, las zonas de unión entre las habitaciones son grandes cristaleras que dejan pasar en gran medida la luz del exterior, por lo que las imágenes se ven realmente influenciadas por este hecho.

4.3. Experimento 3: Efecto de brillo

En este experimento, se realiza un aumento de datos del modelo visual, el cual es el correspondiente a aplicar un efecto de brillo en las imágenes originales. El resultado de este *dataset* es el que se presenta en la figura 4.16.

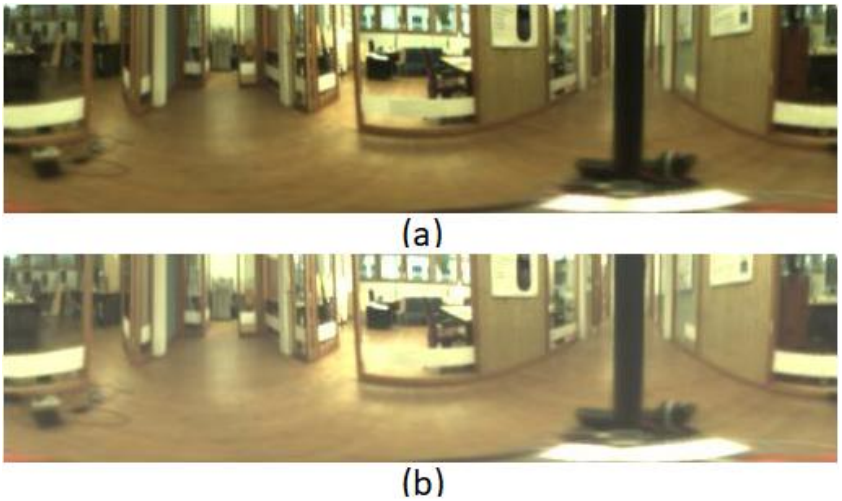


Figura 4.16: Ejemplo de aumento de datos. La figura (a) muestra la imagen original capturada dentro del entorno de Friburgo, mientras que la figura (b) es el resultado de aplicar sobre la primera el efecto de brillo.

Seguidamente, se lleva a cabo el entrenamiento de nuestra red neuronal convolucional, la cual fue readaptada tal como se explicó en la sección 3.2, con imágenes correspondientes a este aumento de datos. De esta manera, una vez terminado el entrenamiento, se realiza el testeo de la CNN reentrenada para

clasificar las diferentes estancias con las tres condiciones de iluminación propuestas. En la figura 4.17 se muestran los porcentajes de precisión obtenidos. Cabe recalcar que para cualquier tipo de iluminación la red neuronal convolucional reentrenada sobrepasa porcentajes mayores al 90% de los valores, ya que se obtiene una precisión del 99.24%, 96.89% y 90.59% para los *datasets* de nublado, noche y soleado, respectivamente. Por tanto, podemos ver cómo la precisión de los resultados alcanzados es significativamente mayor con cualquier tipo de iluminación en comparación con los resultados obtenidos cuando a la red neuronal convolucional no se le ha aplicado ningún aumento de datos.

Además, se puede observar cómo el presente ensayo mejora los resultados que se obtuvieron en el experimento 2 –efecto de foco de sombra– para el *dataset* de soleado debido a que se ha conseguido una precisión del 90.59% frente al 84.54% que se halló anteriormente. Cabe destacar que no se encuentran mejoras significativas para los *datasets* restantes –nublado y noche–.

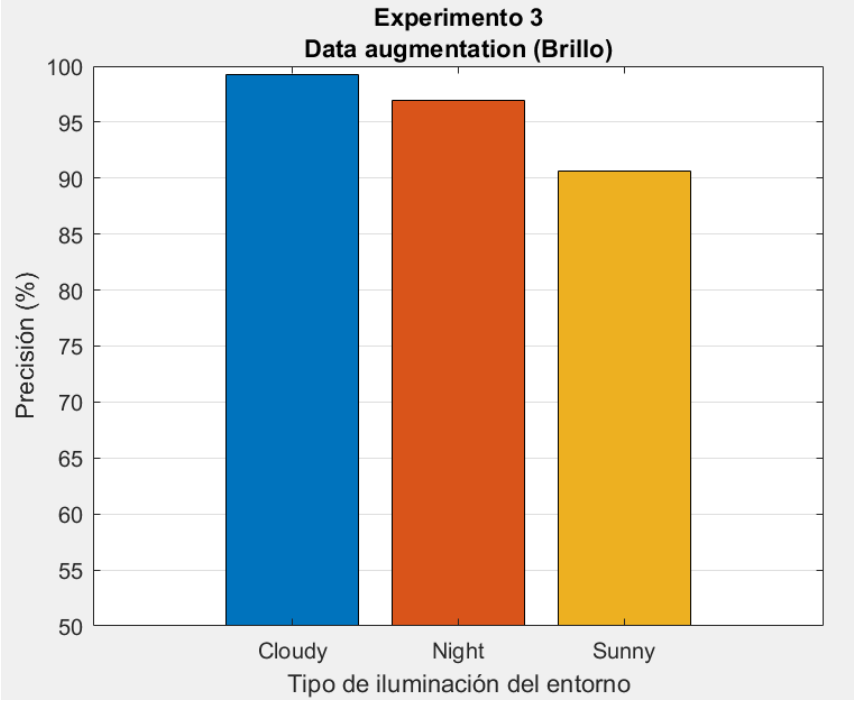


Figura 4.17: Precisión media del proceso de clasificación de estancias de la CNN reentrenada para llevar a cabo la tarea de localización gruesa con aumento de datos (efecto de brillo).

Asimismo, se muestra en la figura 4.18 –para la condición de nublado–, en la figura 4.19 –para la categoría de noche– y en la figura 4.20 –para la condición de soleado– la matriz de confusión para la tarea de clasificación de estancias obtenida tras finalizar el proceso de testeo. Tal como podemos ver, se aprecia claramente que la CNN reentrenada es capaz de clasificar las imágenes correspondientes a todo tipo de estancias. A su vez, se puede observar cómo las matrices de confusión obtenidas son más robustas ante cualquier tipo de iluminación en comparación con las obtenidas para el testeo con la red neuronal convolucional entrenada sin *data augmentation* (figura 4.2, figura 4.3 y figura 4.4).

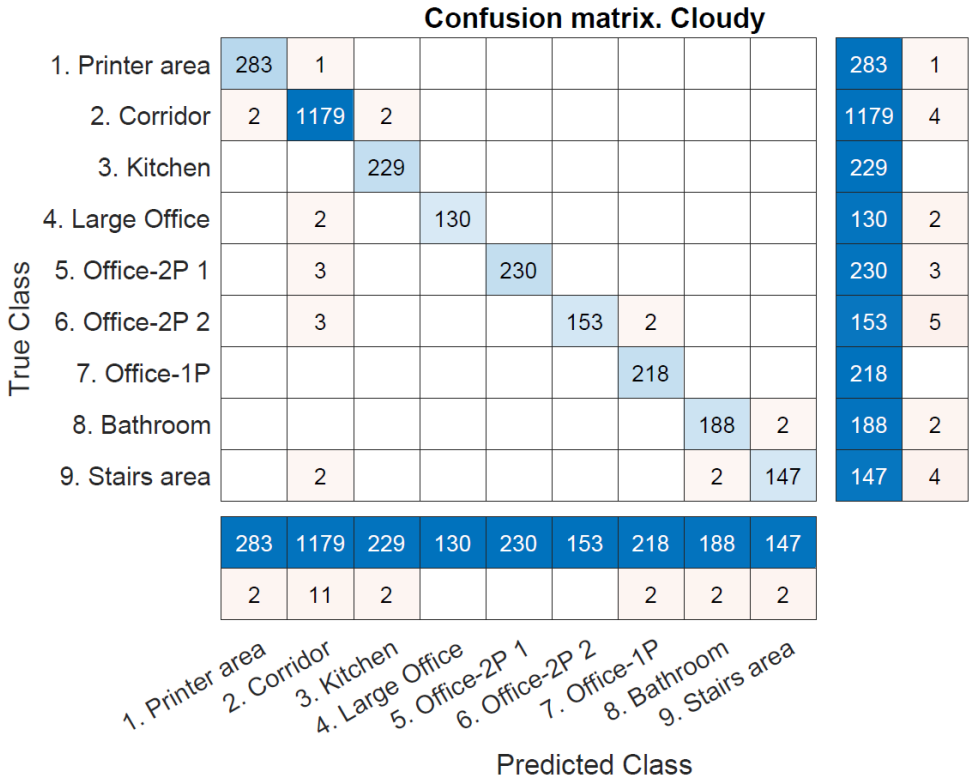


Figura 4.18: Matriz de confusión –clasificación de las estancias– para la condición lumínica de nublado de la CNN reentrenada con aumento de datos (efecto de brillo).

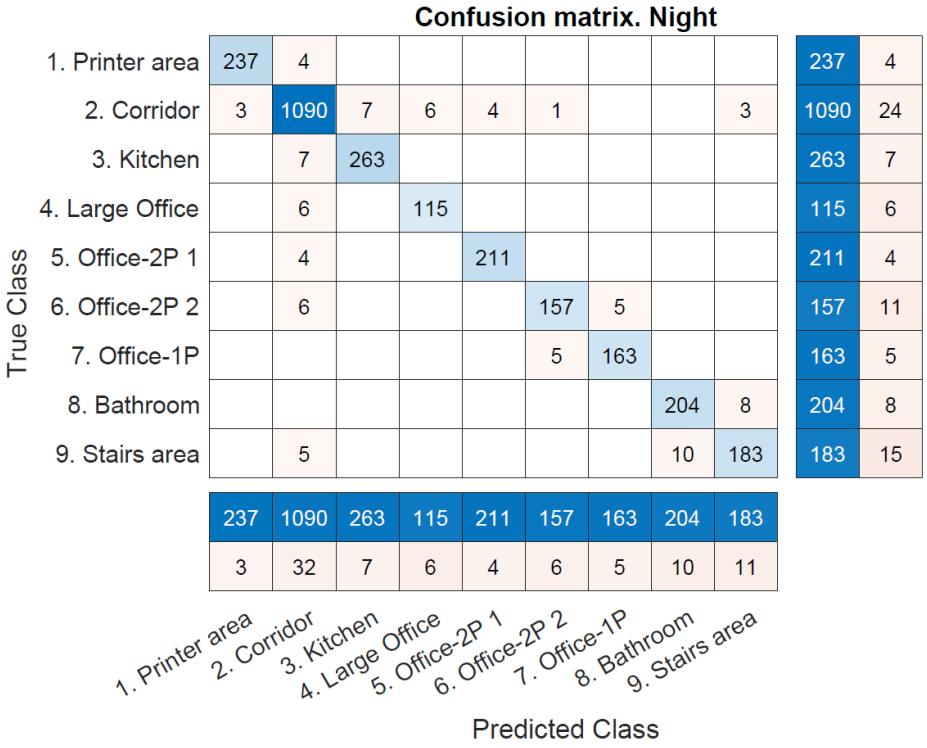


Figura 4.19: Matriz de confusión –clasificación de las estancias– para la condición lumínica de noche de la CNN reentrenada con aumento de datos (efecto de brillo).

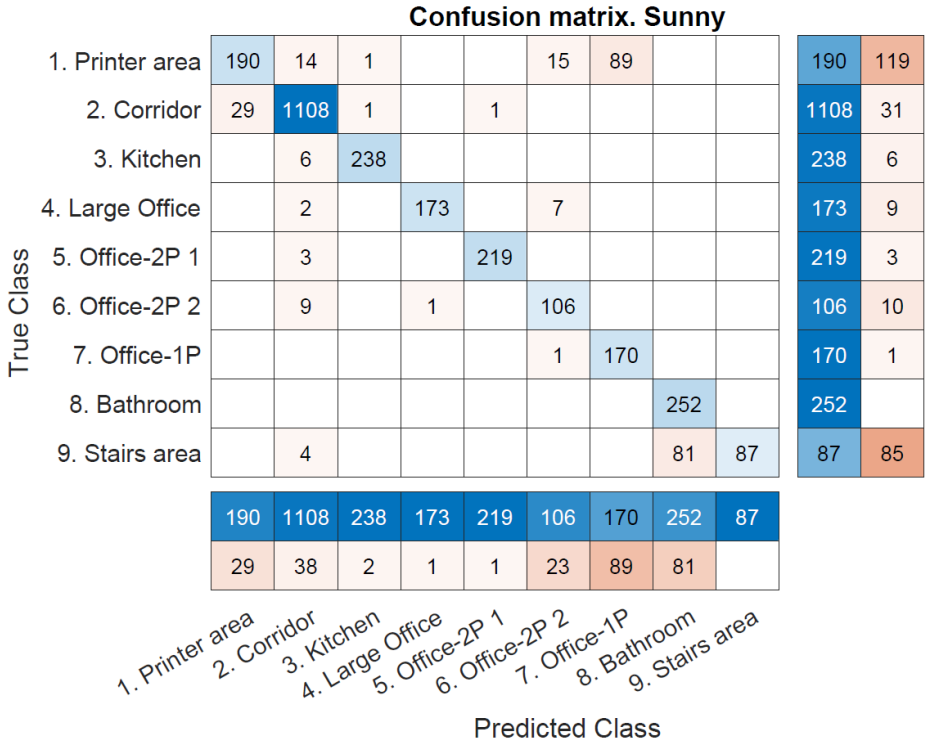


Figura 4.20: Matriz de confusión –clasificación de las estancias– para la condición lumínica de soleado de la CNN reentrenada con aumento de datos (efecto de brillo).

Tal como podemos ver en las matrices de confusión para la condición lumínica de nublado y de noche, no hay puntos de interés destacados debido a que la precisión media del proceso de clasificación de estancias ronda el 100% de los valores. Por ello, nos centraremos en los resultados obtenidos para el tipo de iluminación de soleado, ya que podemos observar que la mayor confusión entre la clasificación de habitaciones viene dada por las correspondientes a *Printer area* y *1 - person office*. Esto es debido a que, aunque son estancias que se encuentran alejadas entre sí, comparten tanto una distribución como un mobiliario parecido, lo que puede originar a pensar que son el mismo entorno. No obstante, cabe remarcar que muy de cerca se presenta también el error en el proceso de clasificación entre las estancias pertinentes a *Stairs area* y *Bathroom*. Esta confusión se origina por el hecho de que ambas habitaciones son contiguas y presentan la misma distribución de paredes blancas, por lo que comparten mucha información común de la escena.

4.4. Experimento 4: Efecto de contraste

En el siguiente experimento, al igual que en los anteriores, se realiza un aumento de datos del modelo visual, el cual es el correspondiente a aplicar un efecto de contraste en las imágenes originales. El resultado de este *dataset* es el que se presenta en la figura 4.21.

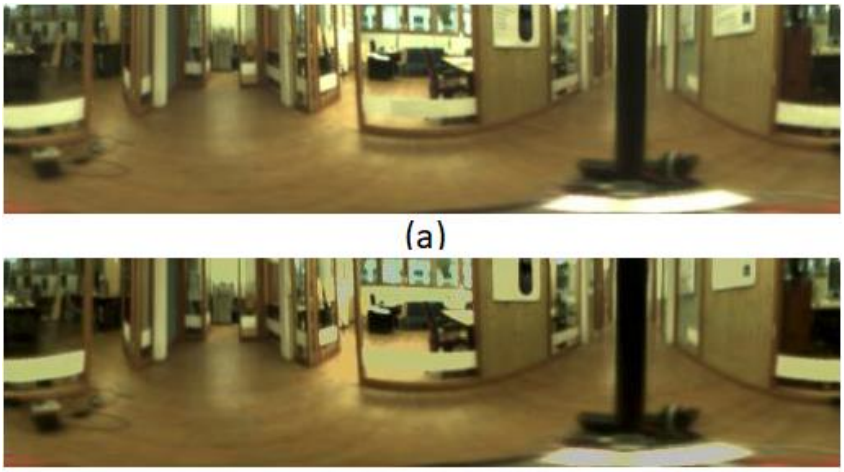


Figura 4.21: Ejemplo de aumento de datos. La figura (a) muestra la imagen original capturada dentro del entorno de Friburgo, mientras que la figura (b) es el resultado de aplicar sobre la primera el efecto de contraste.

Seguidamente, se lleva a cabo el entrenamiento de nuestra red neuronal convolucional, la cual fue readaptada tal como se explicó en la sección 3.2, con imágenes correspondientes a este aumento de datos. De esta manera, una vez terminado el entrenamiento, se realiza el testeo de la CNN reentrenada para clasificar las diferentes estancias con los tres tipos de iluminación propuestos. Es por ello por lo que podemos observar en la figura 4.22 los porcentajes de precisión obtenidos. A través de esta se puede ver que para cualquier tipo de iluminación la red neuronal convolucional reentrenada sobrepasa porcentajes mayores al 88% de los valores, ya que se obtiene una precisión del 98.92%, 97.04% y 88.59% para los *datasets* de nublado, noche y soleado, respectivamente. Por tanto, podemos ver cómo la precisión de los resultados alcanzados es significativamente mayor en cualquier tipo de iluminación en comparación con los hallados en la figura 4.1 –resultados sin previo aumento de datos–.

Además, podemos ver que los resultados empeoran para el *dataset* de soleado en comparación con los obtenidos en el experimento 3 –efecto de brillo–, ya que se ha alcanzado una precisión del 88.59% frente al 90.59% que se logró en ese ensayo. Sin embargo, para los *datasets* restantes –nublado y noche– no se encuentran mejoras significativas.

Asimismo, se muestra en la figura 4.23 –para la condición de nublado–, en la figura 4.24 –para la categoría de noche– y en la figura 4.25 –para la condición de soleado– la matriz de confusión de los resultados de clasificación de estancias obtenidos tras finalizar el proceso de testeo. Tal como podemos ver, se aprecia claramente que la CNN reentrenada es capaz de clasificar las imágenes correspondientes a todo tipo de estancias. A su vez, se puede observar cómo las matrices de confusión obtenidas son más robustas ante cualquier tipo de iluminación en comparación con las graficadas en la figura 4.2, en la figura 4.3 y en la figura 4.4.

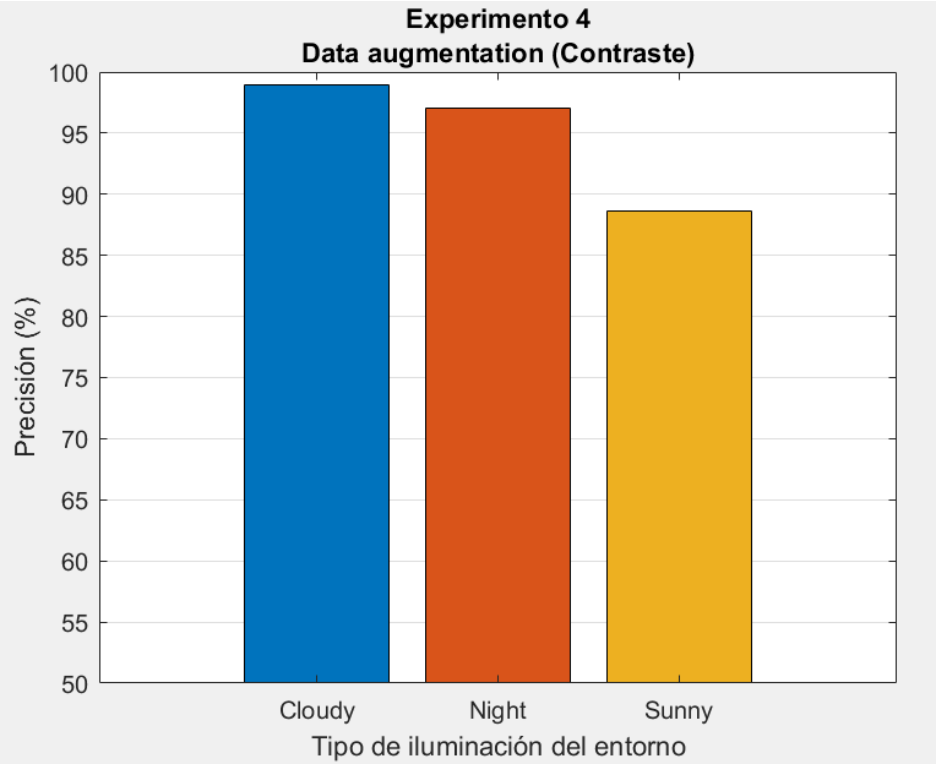


Figura 4.22: Precisión media del proceso de clasificación de estancias de la CNN reentrenada para llevar a cabo la tarea de localización gruesa con aumento de datos (efecto de contraste).

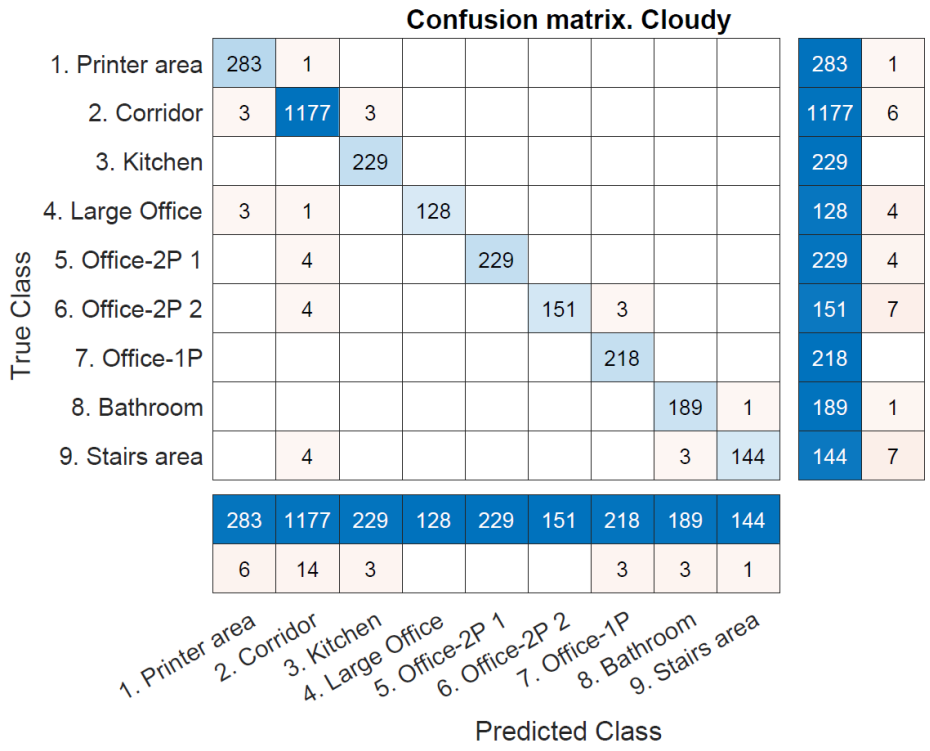


Figura 4.23: Matriz de confusión –clasificación de las estancias– para la condición lumínica de nublado de la CNN reentrenada con aumento de datos (efecto de contraste).

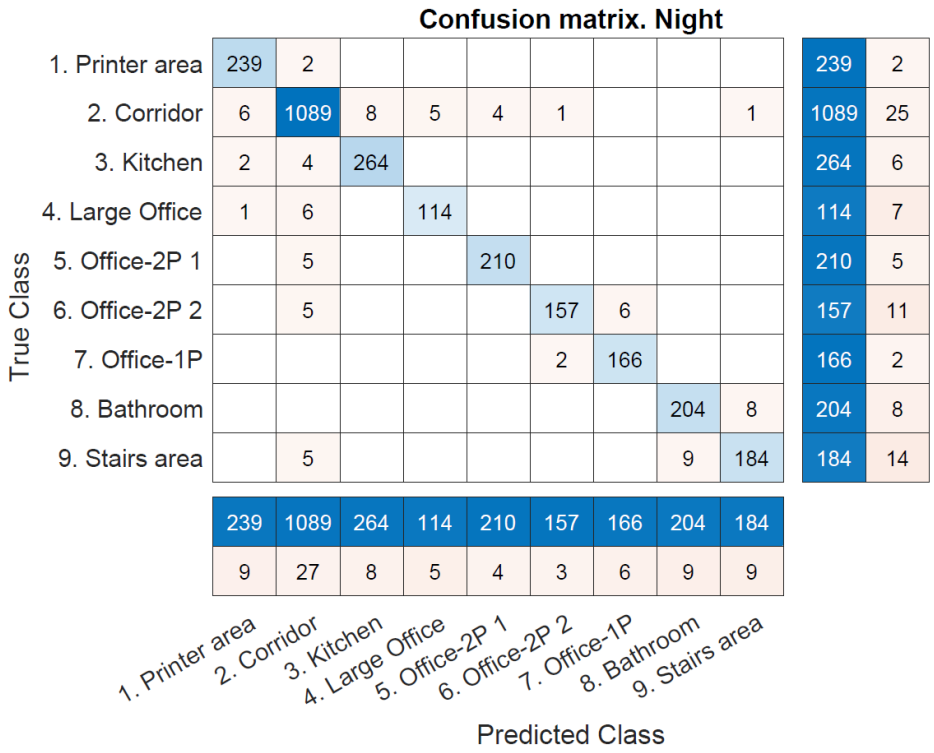


Figura 4.24: Matriz de confusión –clasificación de las estancias– para la condición lumínica de noche de la CNN reentrenada con aumento de datos (efecto de contraste).

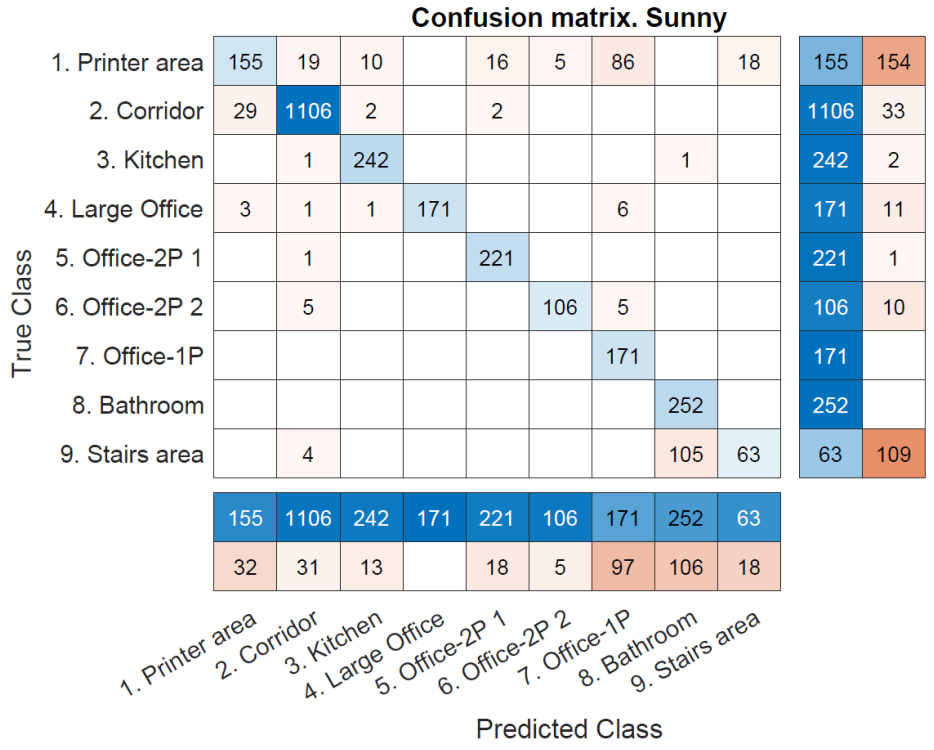


Figura 4.25: Matriz de confusión –clasificación de las estancias– para la condición lumínica de soleado de la CNN reentrenada con aumento de datos (efecto de contraste).

Como podemos observar en las matrices de confusión para el tipo de iluminación de nublado y de noche, no hay información de interés destacada, ya que la precisión media del proceso de clasificación de estancias es realmente alta. Por tanto, nos centraremos en los resultados obtenidos para la condición lumínica de soleado, ya que podemos ver que la mayor confusión entre la clasificación de habitaciones viene dada por las correspondientes a *Stairs area* y *Bathroom*. Esto se debe a que son estancias aledañas y presentan en las paredes la misma tonalidad de color. De la misma manera, la unión de ambas habitaciones está fuertemente iluminada por una ventana que deja pasar en gran medida la luz del exterior, por lo que las imágenes se ven realmente influenciadas por este hecho.

4.5. Experimento 5: Efecto de saturación

En este experimento, se realiza un aumento de datos del modelo visual, el cual es el correspondiente a aplicar un efecto de saturación en las imágenes originales. El resultado de este *dataset* es el que se presenta en la figura 4.26.

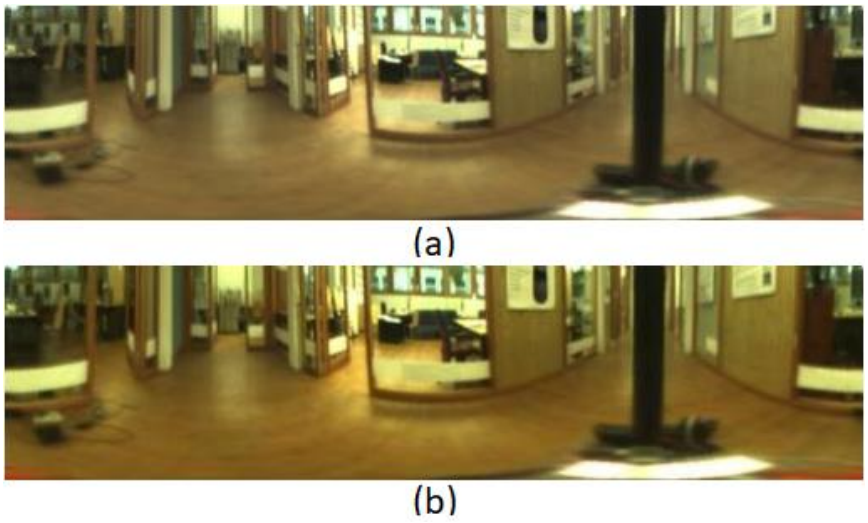


Figura 4.26: Ejemplo de aumento de datos. La figura (a) muestra la imagen original capturada dentro del entorno de Friburgo, mientras que la figura (b) es el resultado de aplicar sobre la primera el efecto de saturación.

Seguidamente, se lleva a cabo el entrenamiento de nuestra red neuronal convolucional, la cual fue readaptada tal como se explicó en la sección 3.2, con

imágenes correspondientes a este aumento de datos. De esta manera, una vez terminado el entrenamiento, se realiza el testeo de la CNN reentrenada para clasificar las diferentes estancias con las tres condiciones de iluminación propuestas. En la figura 4.27 se muestran los porcentajes de precisión obtenidos. Cabe recalcar que para cualquier tipo de iluminación la red neuronal convolucional reentrenada sobrepasa porcentajes mayores al 87% de los valores, ya que se obtiene una precisión del 99.21%, 97.04% y 87.17% para los *datasets* de nublado, noche y soleado, respectivamente. Por tanto, podemos ver cómo la precisión de los resultados alcanzados es significativamente mayor con cualquier tipo de iluminación en comparación con los resultados obtenidos cuando a la red neuronal convolucional no se le ha aplicado ningún aumento de datos.

Además, se puede observar cómo el presente ensayo no mejora los resultados que se obtuvieron en el experimento 4 –efecto de contraste– para el *dataset* de soleado debido a que se ha conseguido una precisión del 87.17% frente al 88.59% que se halló anteriormente. Cabe destacar que no se encuentran mejoras significativas para los *datasets* restantes –nublado y noche–.

Asimismo, se muestra en la figura 4.28 –para la condición de nublado–, en la figura 4.29 –para la categoría de noche– y en la figura 4.30 –para la condición de soleado– la matriz de confusión para la tarea de clasificación de estancias obtenida tras finalizar el proceso de testeo. Tal como podemos ver, se aprecia claramente que la CNN reentrenada es capaz de clasificar las imágenes correspondientes a todo tipo de estancias. A su vez, se puede observar cómo las matrices de confusión obtenidas son más robustas ante cualquier tipo de iluminación en comparación con las obtenidas para el testeo con la red neuronal convolucional entrenada sin *data augmentation* (figura 4.2, figura 4.3 y figura 4.4).

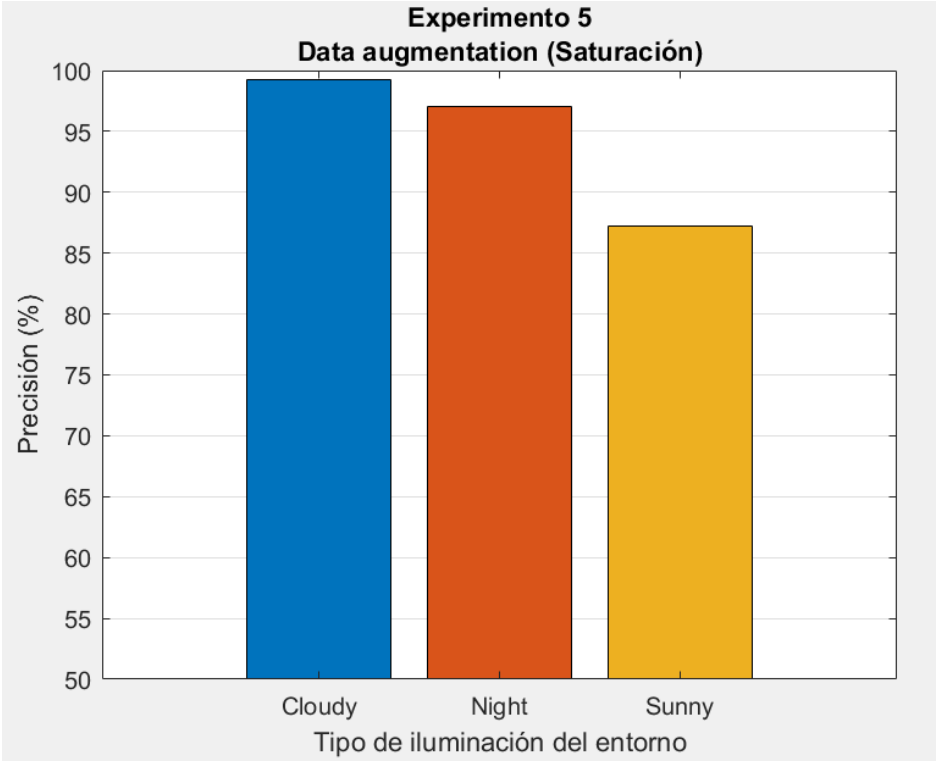


Figura 4.27: Precisión media del proceso de clasificación de estancias de la CNN reentrenada para llevar a cabo la tarea de localización gruesa con aumento de datos (efecto de saturación).

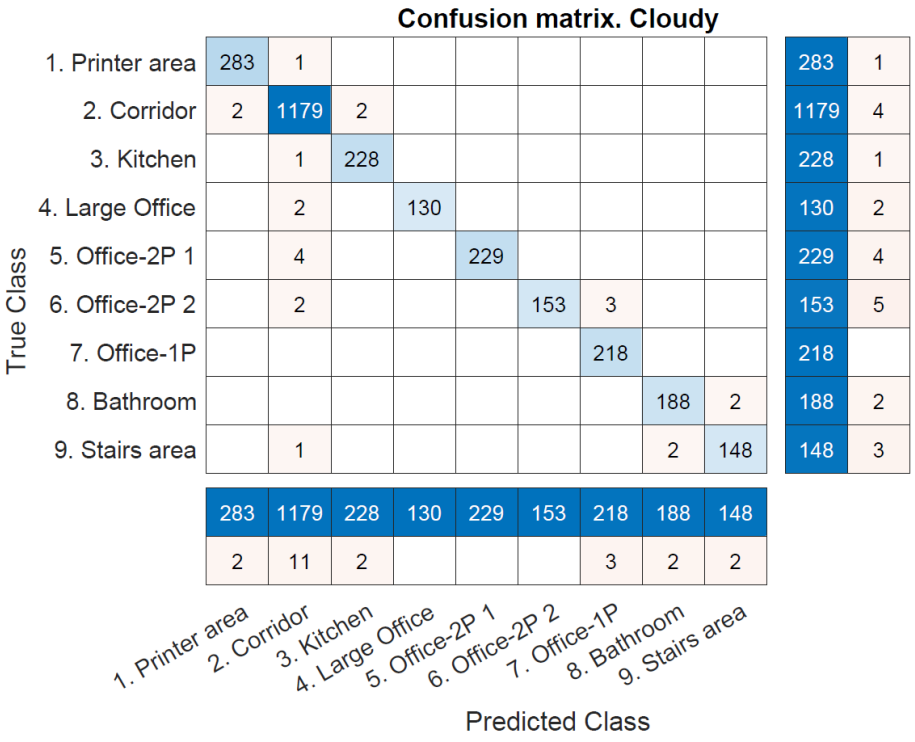


Figura 4.28: Matriz de confusión –clasificación de las estancias– para la condición lumínica de nublado de la CNN reentrenada con aumento de datos (efecto de saturación).

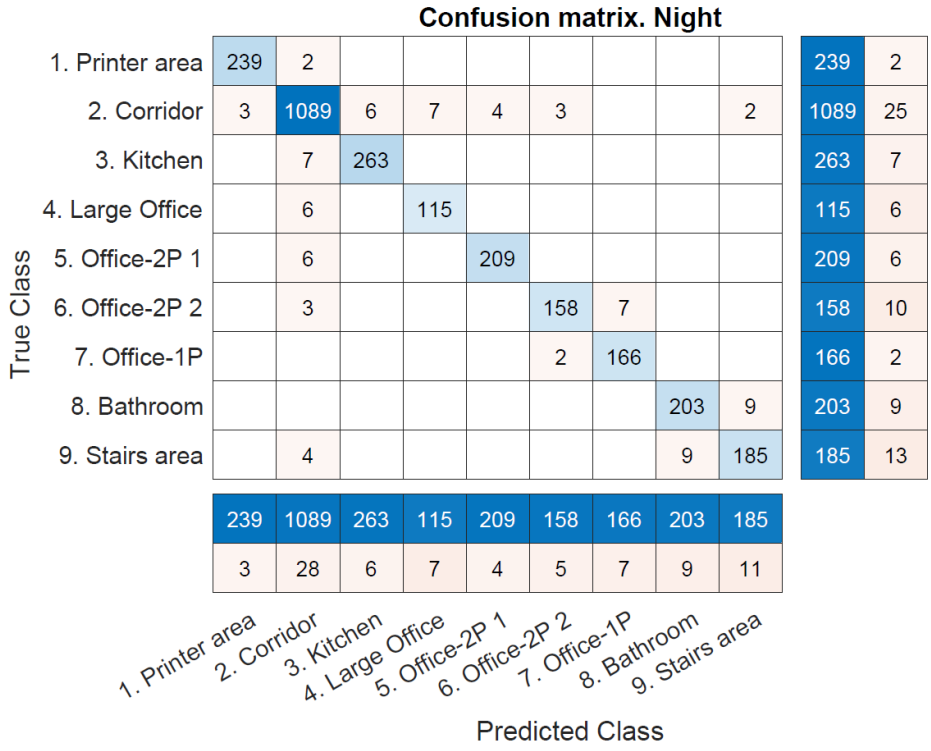


Figura 4.29: Matriz de confusión –clasificación de las estancias– para la condición lumínica de noche de la CNN reentrenada con aumento de datos (efecto de saturación).

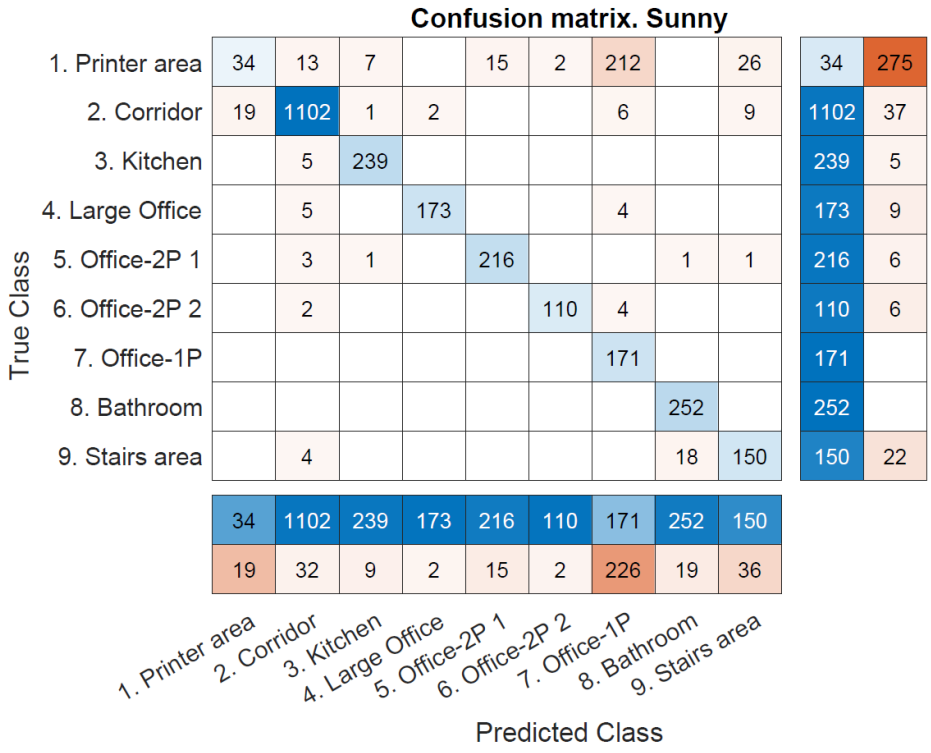


Figura 4.30: Matriz de confusión –clasificación de las estancias– para la condición lumínica de soleado de la CNN reentrenada con aumento de datos (efecto de saturación).

Tal como podemos ver en las matrices de confusión para la condición lumínica de nublado y de noche, no hay puntos de interés destacados debido a que la precisión media del proceso de clasificación de estancias ronda la excelencia. Por ello, nos centraremos en los resultados obtenidos para el tipo de iluminación de soleado, ya que podemos observar que la mayor confusión entre la clasificación de habitaciones viene dada por las correspondientes a *Printer area* y *1 - person office*. Esto es debido a que, aunque son estancias que se encuentran alejadas entre sí, comparten tanto una distribución como un mobiliario parecido, lo que puede motivar a pensar que son el mismo entorno. Cabe remarcar que este experimento tiene el porcentaje más alto del error de confusión entre estas estancias (68.61%) en comparación con los experimentos 1 (49.51%) y 3 (28.80%) que también presentaban la misma casuística.

4.6. Experimento 6: Cambio de orientación

En situaciones reales es muy común que el robot móvil visite puntos del entorno con una orientación diferente, lo cual puede influir negativamente a la hora de tratar de realizar la localización –conocer la posición 'x - y'–. Es por ello por lo que debemos tratar de que el modelo visual sea invariante antes cambios de orientación.

Por tanto, en el siguiente experimento, se realiza un aumento de datos del modelo visual, el cual es el correspondiente a aplicar cambios de orientación en las imágenes originales. El resultado de este *dataset* es el que se presenta en la figura 4.31.

(a)

(b)

Figura 4.31: Ejemplo de aumento de datos. La figura (a) muestra la imagen original capturada dentro del entorno de Friburgo, mientras que la figura (b) es el resultado de aplicar sobre la primera un cambio de orientación.

Seguidamente, se lleva a cabo el entrenamiento de nuestra red neuronal convolucional, la cual fue readaptada tal como se explicó en la sección 3.2, con imágenes correspondientes a este aumento de datos. De esta manera, una vez terminado el entrenamiento, se realiza el testeo de la CNN reentrenada para clasificar las diferentes estancias con los tres tipos de iluminación propuestos. Es por ello por lo que podemos observar en la figura 4.32 los porcentajes de precisión obtenidos. A través de esta se puede ver que para cualquier tipo de iluminación la red neuronal convolucional reentrenada sobrepasa porcentajes mayores al 88% de los valores, ya que se obtiene una precisión del 99.17%, 97.27% y 88.31% para los *datasets* de nublado, noche y soleado, respectivamente. Por tanto, podemos ver cómo la precisión de los resultados alcanzados es significativamente mayor en cualquier tipo de iluminación en comparación con los hallados en la figura 4.1 –resultados sin previo aumento de datos–.

Además, podemos ver que los resultados mejoran para el *dataset* de soleado en comparación con los obtenidos en el experimento 5 –efecto de saturación–, ya que se ha alcanzado una precisión del 88.31% frente al 87.17% que se logró en ese ensayo. Sin embargo, para los *datasets* restantes –nublado y noche– no se encuentran mejoras significativas.

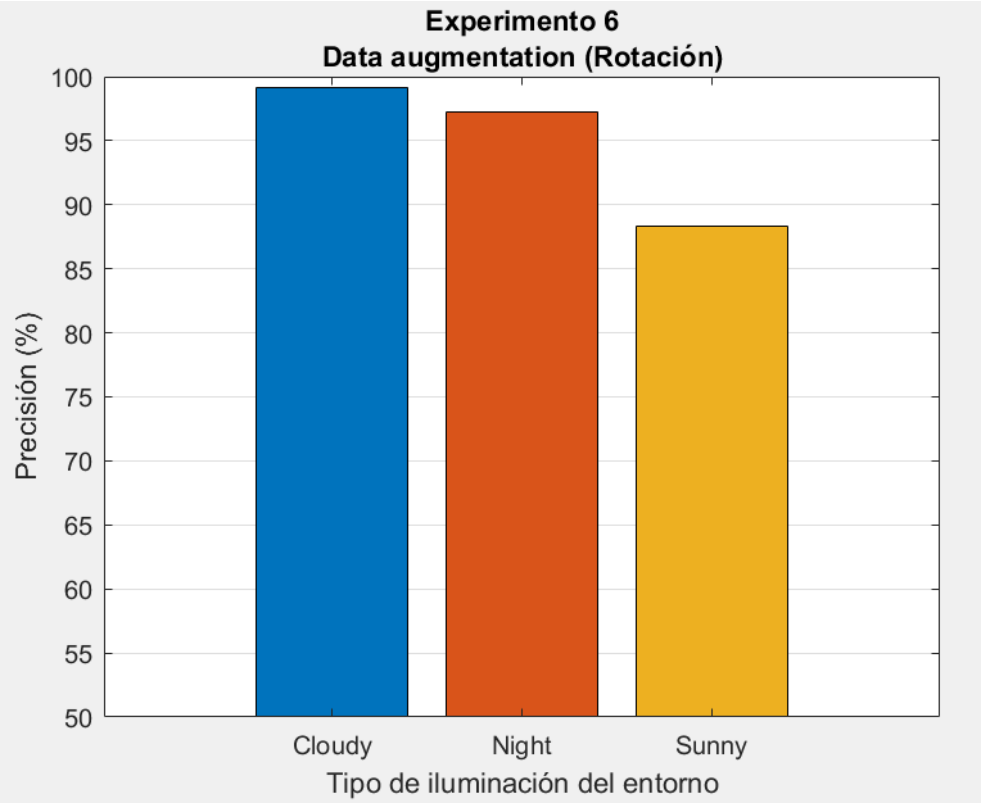


Figura 4.32: Precisión media del proceso de clasificación de estancias de la CNN reentrenada para llevar a cabo la tarea de localización gruesa con aumento de datos (cambio de orientación).

Asimismo, se muestra en la figura 4.33 –para la condición de nublado–, en la figura 4.34 –para la categoría de noche– y en la figura 4.35 –para la condición de soleado– la matriz de confusión de los resultados de clasificación de estancias obtenidos tras finalizar el proceso de testeo. Tal como podemos ver, se aprecia claramente que la CNN reentrenada es capaz de clasificar las imágenes correspondientes a todo tipo de estancias. A su vez, se puede observar cómo las matrices de confusión obtenidas son más robustas ante cualquier tipo de iluminación en comparación con las graficadas en la figura 4.2, en la figura 4.3 y en la figura 4.4.

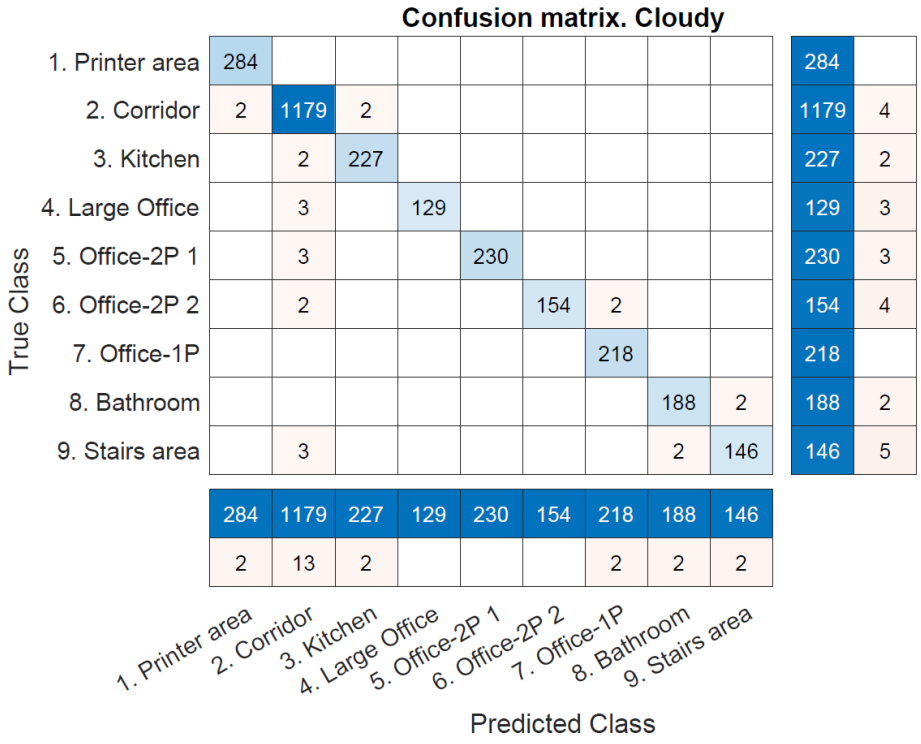


Figura 4.33: Matriz de confusión –clasificación de las estancias– para la condición lumínica de nublado de la CNN reentrenada con aumento de datos (cambio de orientación).

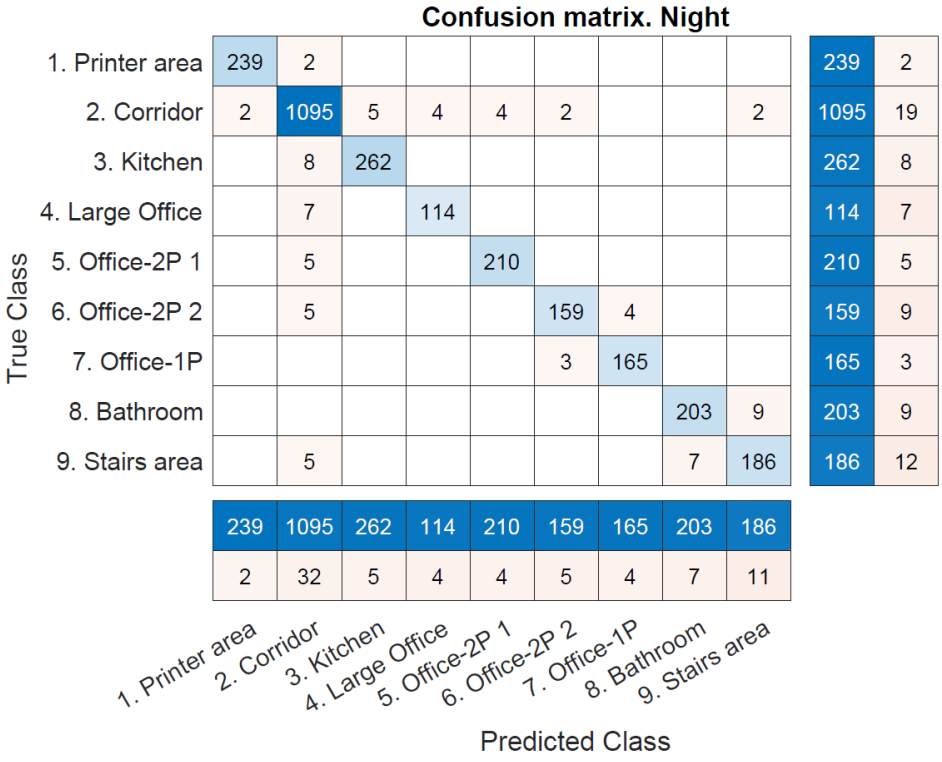


Figura 4.34: Matriz de confusión –clasificación de las estancias– para la condición lumínica de noche de la CNN reentrenada con aumento de datos (cambio de orientación).

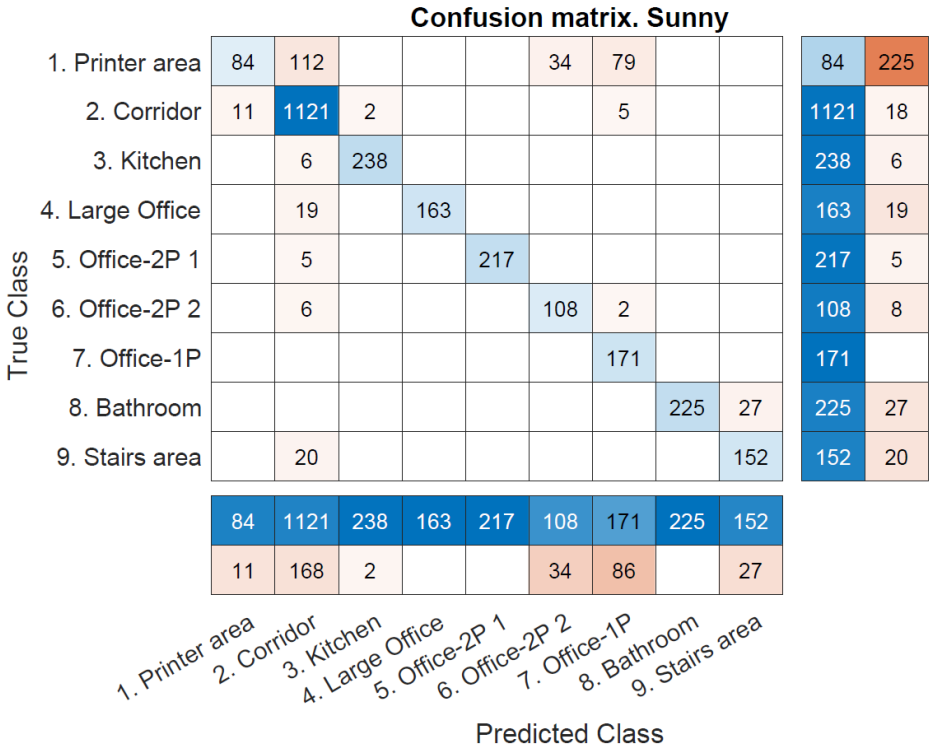


Figura 4.35: Matriz de confusión –clasificación de las estancias– para la condición lumínica de soleado de la CNN reentrenada con aumento de datos (cambio de orientación).

Como podemos observar en las matrices de confusión para el tipo de iluminación de nublado y de noche, no hay información de interés destacada, ya que la precisión media del proceso de clasificación de estancias es realmente elevada. Por tanto, nos centraremos en los resultados obtenidos para la condición lumínica de soleado, ya que podemos ver que la mayor confusión entre la clasificación de habitaciones viene dada por las correspondientes a *Printer area* y *Corridor*. Esto se debe a que, además de ser estancias que se encuentran cercanas la una de la otra, las zonas de unión entre las habitaciones son grandes cristaleras que dejan pasar en gran medida la luz del exterior, por lo que las imágenes se ven realmente influenciadas por este hecho. Cabe remarcar que este experimento tiene el porcentaje más bajo del error de confusión entre estas estancias (36.25%) en comparación con el experimento 2 (59.22%) que también presentaba el mismo caso concreto.

Por último, para poder llevar a cabo una comparativa entre los experimentos realizados en esta investigación, se han tomado los resultados de

precisión para la condición lumínica de nublado de cada uno de los ensayos propuestos –ver figura 4.36–. Las barras azules presentan la precisión obtenida al realizar la clasificación de estancias y los puntos naranjas muestran la relación entre el tiempo de entrenamiento de la CNN y el número de imágenes utilizadas para realizar dicho entrenamiento. Tal como podemos observar, se ha normalizado el tiempo del proceso de entrenamiento para cada una de las imágenes pertenecientes a cada uno de los *dataset* empleados –los cuales fueron mostrados ya en la tabla 3.4– para que así no se vea afectada la duración de dicho proceso por la cantidad de imágenes existentes en el modelo.

De esta forma, observamos cómo, ante cualquier aumento de datos aplicado al *dataset* original, los resultados de precisión en la clasificación de estancias son notablemente mejores. Además, cabe destacar también que, por un lado, a la hora de realizar el entrenamiento de la CNN se emplea un menor tiempo para esta tarea cuando se trabaja con el *dataset* original y, por otro lado, se obtiene una mayor duración en este proceso para el aumento de datos correspondiente a haber aplicado en las imágenes un cambio de orientación.

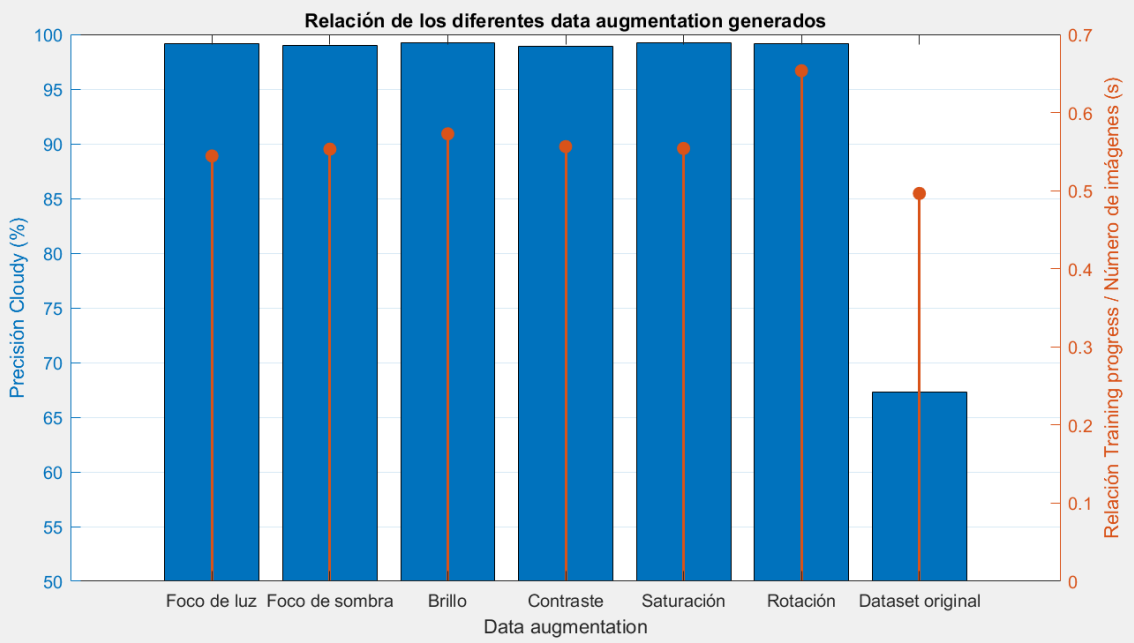


Figura 4.36: Comparativa entre los diferentes efectos visuales evaluados para llevar a cabo el aumento de datos y posterior re-entreno de la CNN para realizar la tarea de recuperación de estancias con el *dataset* de test.

Capítulo 5
Conclusiones y trabajos futuros

En el presente trabajo de fin de máster se han estudiado los problemas en la tarea de localización jerárquica cuando se aplican determinados aumentos de datos al modelo visual para llevar a cabo el entrenamiento de una red neuronal profunda y únicamente se dispone de una cámara omnidireccional como sensor de medida. El problema se ha abordado realizando una serie de pruebas en las que se ha resuelto dicha tarea cuando se aplica a las imágenes existentes una determinada variación visual como un foco de luz, un foco de sombra, un cambio de brillo, una variación de contraste, una alteración de la saturación y una posible rotación de la imagen debido a la orientación con la que el robot hubiese tomado la instantánea. De la misma manera, el robot móvil que realizaba esta tarea estaba equipado por un sistema de visión catadióptrico montado sobre él y las imágenes que capturaba eran la única información que se proporcionaba para resolver el problema de localización. Dichas imágenes pertenecen a un *dataset* contenido en *The COLD Database* [27], el cual posee imágenes omnidireccionales con información de *ground truth* adquiridas bajo tres condiciones de iluminación –nublado, noche y soleado–.

Es por ello por lo que este trabajo plantea llevar a cabo la tarea de la localización de un robot móvil mediante la readaptación y entrenamiento de una red neuronal convolucional por medio de la aplicación de la técnica de aumento de datos para mejorar el proceso de entrenamiento de la CNN. Para ello, se evalúan las mejorías de aplicar los efectos visuales para implementar el aumento de datos del modelo visual para, de esta forma, realizar una localización jerárquica tanto con cambios de iluminación en el entorno como sin ellos.

Además, la manera de llevar a cabo dicha tarea es por medio del método conocido como *room retrieval* o recuperación de la estancia.

Para medir la mejoría presentada al aplicar los diferentes efectos visuales a la hora de llevar a cabo el proceso de *data augmentation*, se ha realizado un estudio acerca de cuán robusta es la CNN empleada, por lo que para ello se han obtenido sus resultados de clasificación de estancias comprobando así su eficiencia ante el cambio sustancial de la apariencia visual del entorno.

A continuación, se presentan las conclusiones de cada uno de los seis experimentos realizados.

Experimento nº 1

En este experimento, por una parte, podemos corroborar que el aumento de datos correspondiente al efecto de foco de luz otorga mejores resultados en la precisión del proceso de clasificación de estancias en comparación con los obtenidos cuando no se aplica *data augmentation* al modelo visual. Cabe destacar que los mejores resultados para la condición lumínica de soleado se hayan obtenido en este experimento (91.38%), ya que trabajábamos con incrementos locales de intensidad en la imagen y, tal como como se pudo ver en las investigaciones de Céspedes [8], para este tipo de iluminación siempre se alcanzaban los resultados más bajos. Por otra parte, concluimos que la precisión en la clasificación de estancias para la condición lumínica de nublado (99.17%) y de noche (97.16%) ha sido exitosa. Sin embargo, en cuanto a la iluminación de soleado, obtenemos que la mayor confusión en la clasificación de habitaciones viene dada entre *Printer area* y *1 - person office*. Esto es debido a que, aunque son estancias que se encuentran alejadas entre sí, comparten una distribución parecida del mobiliario, lo que puede originar a pensar que son el mismo entorno. Así pues, se obtiene que para el 49.51% de los casos en los que la estancia real es el *Printer area*, la CNN obtiene como resultado la *1 - person office*.

Experimento nº 2

En primer lugar, se demuestra que la clasificación de estancias para los tipos de iluminación de nublado (99.03%) y de noche (97.04%) ha sido muy buena. Cabe recalcar que los peores resultados para la condición lumínica de soleado se hayan obtenido en este experimento (84.54%). Para dicha condición, extraemos que la mayor confusión en la clasificación de habitaciones viene dada entre *Printer area* y *Corridor*. Esto es a causa de que, además de ser estancias que se encuentran cercanas la una de la otra, presentan en su zona de unión una gran cristalera que deja pasar en gran medida la luz del exterior, por lo que las imágenes se ven realmente influenciadas por este hecho. Asimismo, se determina que para el 59.22% de los casos en los que la habitación real es el *Printer area*, la red neuronal convolucional concluye que se trata del *Corridor*.

Experimento nº 3

En este experimento, por un lado, se obtiene que la precisión de los resultados para la condición lumínica de soleado es del 90.59%. Así pues, hallamos que las mayores confusiones en la clasificación de estancias vienen dadas tanto entre *Printer area* y *1 - person office* como entre *Stairs area* y *Bathroom*. Esta segunda es debida a que ambas habitaciones son contiguas y presentan la misma distribución de paredes blancas, De la misma manera, la unión de ambas habitaciones está fuertemente iluminada por una ventana que deja pasar en gran medida la luz del exterior, hecho que influencia notoriamente la apariencia de las imágenes adquiridas en ese espacio. Por tanto, se concluye que para el suceso *Printer area* / *1 - person office* la CNN está tomando valores erróneos el 28.80% de las veces. Además, para el 47.09% de los casos en los que la habitación real es el *Stairs area*, la red neuronal convolucional toma como resultado el *Bathroom*. Por otro lado, concluimos que la clasificación de estancias para la condición lumínica de nublado (99.24%) y de noche (96.89%) ha sido acertada.

Experimento nº 4

Por una parte, en cuanto a la iluminación de soleado, obtenemos que la mayor confusión en la clasificación de habitaciones viene dada entre *Stairs area* y *Bathroom*. Por ello, se determina que para esta casuística la CNN no clasifica

correctamente el 61.05% de las imágenes. Por otro lado, se logra que la precisión en la clasificación de estancias para la condición lumínica de nublado, de noche y de soleado sea de 98.92%, 97.04% y 88.59%, respectivamente.

Experimento nº 5

En primer lugar, se demuestra que la clasificación de habitaciones para los tipos de iluminación de nublado (99.20%) y de noche (97.04%) ha sido exitosa. En cuanto a la iluminación de soleado, la cual ha obtenido una precisión en sus resultados del 87.17%, hallamos que la mayor confusión en la clasificación de estancias viene dada entre *Printer area* y *1 - person office*. Así pues, se obtiene que para el 68.61% de las veces la red neuronal convolucional obtiene resultados no acertados.

Experimento nº 6

Para finalizar, por un lado, cabe destacar que los mejores resultados para la condición lumínica de noche se hayan obtenido en este experimento (97.27%). Seguidamente, en cuanto a la iluminación de soleado, obtenemos que la mayor confusión en la clasificación de habitaciones viene dada entre *Printer area* y *Corridor*. Por tanto, la CNN está tomando elecciones falsas en el 36.25% de las decisiones.

Los métodos de descripción visual basados en redes neuronales convolucionales buscan minimizar el efecto de *visual aliasing*, pero este es irremediable al haber utilizado en este trabajo únicamente información visual para resolver la tarea de localización. Es por ello por lo que algunos de los errores obtenidos en la resolución de esta tarea son producidos por dicho efecto.

Asimismo, debido al valor de estos resultados experimentales, los cuales animan a profundizar en este campo de estudio por su gran precisión en la estimación de la posición del robot móvil en el modelo visual, se pueden plantear nuevas líneas de trabajo que permitan avanzar en este campo, ya que se ha comprobado que el aumento de datos es una técnica a tener en cuenta en el entrenamiento de modelos visuales para llevar a cabo la resolución de la tarea de localización.

En cuanto a las posibles líneas futuras de investigación, existen varios estudios que se podrían realizar para ampliar este trabajo de fin de máster. Estos serían los siguientes:

- Aplicar todos los efectos por separado a las imágenes del modelo visual, es decir, únicamente implementaríamos una única perturbación a la imagen original. De este modo, entrenaríamos la red neuronal convolucional con todos los efectos al mismo tiempo.

- Combinar aleatoriamente diversos efectos en cada una de las imágenes del modelo visual, es decir, en este caso sí que llevaríamos a cabo varias perturbaciones en la misma imagen. De la misma manera, entrenaríamos la red neuronal convolucional con este *dataset* generado, ya que hay que tener en cuenta que en un ambiente de trabajo real podemos encontrar varios efectos aplicados a la misma instantánea.

- Adaptar el tamaño de entrada de las imágenes de nuestro *dataset* a 227 x 227 –en este caso es para *Places*– para así poder hacer *transfer learning* con la CNN que queramos trabajar. De este modo, realmente aprovechamos los parámetros que la red neuronal convolucional había aprendido cuando fue entrenada y no perdemos así el valor de los pesos preestablecidos.

- Llevar a cabo un proceso de *transfer learning* con el mejor *data augmentation* que obtenemos de los seis experimentos que en el presente trabajo se han realizado. De esta forma, podremos preservar el tamaño de la imagen de entrada y, por tanto, conservar la información aprendida por la CNN cuando se realizó su entrenamiento.

Referencias

[1] J. G. Armas and A. J. Cerda, “Implementación de un robot móvil para desplazamientos en ambientes no estructurados empleando visión artificial”, Trabajo de Fin de Grado, Escuela Superior Politécnica de Chimborazo, Riobamba, Chimborazo, Ecuador, 2020.

[2] N. Ayache and F. Lustman, “Trinocular stereo vision for robotics”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 1, pp. 73-85, January 1991.

[3] J. J. Cabrera, S. Cebollada, M. Ballesta, L. M. Jiménez, L. Payá and Ó. Reinoso, “Entrenamiento, optimización y validación de una CNN para localización jerárquica mediante imágenes omnidireccionales”, in *Proceedings of the XLII Jornadas de Automática*, Castellón, España, 2021, pp. 640-647.

[4] A. Carrio, C. Sampedro, A. Rodríguez and P. Campoy, “A Review of Deep Learning Methods and Applications for Unmanned Aerial Vehicles”, *Journal of Sensors*, vol. 2017, no. 2, pp. 1-13, August 2017.

[5] S. Cebollada, L. Payá, M. Flores, V. Román, A. Peidró and Ó. Reinoso, “A Deep Learning Tool to Solve Localization in Mobile Autonomous Robotics”, in *Proceedings of the 17th International Conference on Informatics in Control, Automation and Robotics*, Lieusaint, Paris, France, 2020, pp. 232-241.

[6] S. Cebollada, L. Payá, X. Jiang and Ó. Reinoso, “Development and use of a convolutional neural network for hierarchical appearance-based localization”, *Artifcial Intelligence Review*, vol. 55, no. 3, pp. 2847-2874, September 2021.

[7] S. Cebollada, L. Payá, V. Román and Ó. Reinoso, “Hierarchical Localization in Topological Models Under Varying Illumination Using Holistic Visual Descriptors”, *IEEE Access*, vol. 7, no. 1, pp. 49580-49595, April 2019.

[8] O. J. Céspedes, “Localización de un robot móvil utilizando información visual y redes neuronales convolucionales”, Trabajo de Fin de Grado, Ingeniería de Sistemas y Automática, Universidad Miguel Hernández de Elche, Elche, Alicante, España, 2020.

[9] M. Flores, “Desarrollo de algoritmos de odometría visual mediante una cámara de 360 grados”, Trabajo de Fin de Máster, Ingeniería de Sistemas y Automática, Universidad Miguel Hernández de Elche, Elche, Alicante, España, 2018.

[10] R. A. García and M. Arias, “Prototipo virtual de un robot móvil multi-terreno para aplicaciones de búsqueda y rescate”, *ResearchGate*, pp. 337-351, October 2016.

[11] J. Gaspar, N. Winters and J. Santos-Victor, “Vision-based navigation and environmental representations with an omnidirectional camera”, *IEEE Transactions on Robotics and Automation*, vol. 16, no. 6, pp. 890-898, December 2000.

[12] A. Gil, Ó. Reinoso, M. Ballesta, M. Juliá and L. Payá, “Estimation of Visual Maps with a Robot Network Equipped with Vision Sensors”, *Sensors*, vol. 10, no. 5, pp. 5209-5232, May 2010.

[13] G. Giralt, R. Sobek and R. Chatila, “A multi-level planning and navigation system for a mobile robot: a first approach to HILARE”, in *Proceedings of the 6th International Joint Conference on Artificial Intelligence*, Tokyo, Japan, 1979, pp. 335-337.

[14] J. Guo and S. Gould, “Deep CNN Ensemble with Data Augmentation for Object Detection”, *arXiv*, vol. 1506.07224, pp. 1-5, June 2015.

[15] P. -T. Jiang, C. -B. Zhang, Q. Hou, M. -M. Cheng and Y. Wei, "LayerCAM: Exploring Hierarchical Class Activation Maps for Localization", *IEEE Transactions on Image Processing*, vol. 30, pp. 5875-5888, June 2021.

[16] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097-1105, January 2012.

[17] A. C. Murillo, J. J. Guerrero and C. Sagues, “SURF features for efficient robot localization with omnidirectional images”, in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Roma, Italy, 2007, pp. 1-8.

[18] J. Neumann, C. Fermuller and Y. Aloimonos, “Eyes from eyes: new cameras for structure from motion”, in *Proceedings of the IEEE Workshop on Omnidirectional Vision. Held in conjunction with ECCV'02*, Copenhagen, Denmark, 2002, pp. 19-26.

[19] L. Payá, “Técnicas de descripción de la apariencia global de escenas: Aplicación a la creación de mapas y localización de robots móviles”, Tesis Doctoral, Ingeniería de Sistemas y Automática, Universidad Miguel Hernández de Elche, Elche, Alicante, España, 2014.

[20] L. Payá, L. Fernández, Ó. Reinoso, A. Gil and D. Úbeda, “Appearance-based Dense Maps Creation - Comparison of Compression Techniques with Panoramic Images”, in *Proceedings of the 6th International Conference on Informatics in Control, Automation and Robotics*, Milan, Italy, 2009, pp. 250-255.

[21] L. Payá, A. Gil and Ó. Reinoso, “A State-of-the-Art Review on Mapping and Localization of Mobile Robots Using Omnidirectional Vision Sensors”, *Journal of Sensors*, vol. 2017, pp. 1-20, April 2017.

[22] L. Payá, A. Peidró, F. Amorós, D. Valiente and Ó. Reinoso, “Modeling Environments Hierarchically with Omnidirectional Imaging and Global-Appearance Descriptors”, *Remote Sensing*, vol. 10, no. 4, pp. 522, March 2018.

[23] P. Sarlin, C. Cadena, R. Siegwart and M. Dymczyk, “From Coarse to Fine: Robust Hierarchical Localization at Large Scale”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 12716-12725.

[24] M. T. Seco, “Robot Localization in Tunnel-like Environments”, Tesis Doctoral, Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Zaragoza, España, 2020.

[25] R. Sim and G. Dudek, “Effective exploration strategies for the construction of visual maps”, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA, 2003, pp. 3224-3231.

[26] R. Sim and J. J. Little, “Autonomous vision-based exploration and mapping using hybrid maps and Rao-Blackwellised particle filters”, in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, 2006, pp. 2082-2089.

[27] M. M. Ullah, A. Pronobis, B. Caputo, J. Luo and P. Jensfelt, “The COLD Database”, KTH Royal Institute of Technology, Stockholm, Sweden, Technical Report TRITA-CSC-CV 2007:1, 2007.

[28] N. Winters, J. Gaspar, G. Lacey and J. Santos-Victor, “Omni-directional vision for robot navigation”, in *Proceedings of the IEEE Workshop on Omnidirectional Vision*, Hilton Head Island, SC, USA, 2000, pp. 21-28.

[29] D. F. Wolf and G. S. Sukhatme, “Semantic Mapping Using Mobile Robots”, *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 245-258, April 2008.

[30] M. Xu, N. Snderhauf and M. Milford, “Probabilistic Visual Place Recognition for Hierarchical Localization”, *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 311-318, April 2021.

[31] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba and A. Oliva, “Learning Deep Features for Scene Recognition using Places Database”, in *Proceedings of the Advances in Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 487-495.