Universidad Zaragoza

Trabajo Fin de Máster

Registro multi-modal de imágenes médicas mediante aprendizaje profundo

Multi-modal medical imaging registration with deep-learning

Autor

David Solanas Sanz

Directores

Mónica Hernández Giménez
Ubaldo Ramón Júlvez

Máster en Ingeniería Informática

Departamento de Informática e Ingeniería de Sistemas
Escuela de Ingeniería y Arquitectura

2023

Resumen

El registro de imágenes se define como el proceso de alinear dos o más imágenes a un sistema de coordenadas común de forma que se maximice su similitud de acuerdo con una métrica adecuada. El objetivo del problema es encontrar una transformación espacial que optimice la superposición entre las estructuras anatómicas o características correspondientes presentes en las imágenes, permitiendo comparaciones y análisis más efectivos. El registro de imágenes médicas es fundamental en múltiples aplicaciones clínicas, como la reconstrucción de imágenes, el seguimiento de la evolución de tumores, y la planificación de ciertas cirugías, entre otros. El auge del aprendizaje profundo ha permitido el desarrollo de diferentes métodos que tratan de resolver el problema del registro de imágenes, aunque todavía existen formulaciones del problema sin explorar. Una de las más importantes es la del registro multimodal.

El principal objetivo de este Trabajo de Fin de Máster es el desarrollo de un sistema de aprendizaje profundo basado en redes neuronales convolucionales y generativas antagónicas (GAN) que, a partir de imágenes cerebrales por resonancia magnética (MRI) de tipo T1 y T2, sea capaz de realizar el registro no-rígido multimodal de dichas imágenes. La obtención de las imágenes para el entrenamiento y test del método se han obtenido del proyecto *Open Access Series of Imaging Studies* (OASIS-3). Se han desarrollado dos sistemas con dos arquitecturas diferentes: el primer sistema usa la arquitectura CycleGAN para abordar el cambio de dominio de imágenes MRI de tipo T2 a tipo T1. El segundo sistema usa la arquitectura U-Net para realizar el registro de dos imágenes MRI de tipo T1. Las imágenes utilizadas en ambos sistemas son volúmenes 3D, que han tenido que ser convenientemente procesadas para mejorar el rendimiento de ambas redes.

Los experimentos realizados en este trabajo demuestran que la combinación de CycleGAN y VoxelMorph es prometedora como solución al problema de registro no-rígido multimodal. La solución propuesta tiene la ventaja de que, gracias al uso de la arquitectura de CycleGAN, el problema de registro multimodal se convierte en un problema unimodal que puede ser abordado por una gran cantidad de métodos de registro. Además, el proceso completo para realizar el registro es significativamente más rápido que el de los métodos tradicionales, lo que lo convierte en una opción preferible en el procesado de conjuntos grandes de datos.

Lista de abreviaturas utilizadas en el trabajo

- CBAM: *Convolutional Block Attention Module*.
- DSC: *Dice Similarity Coeficient*.
- GAN: *Generative Adversarial Networks*.
- GPU: *Graphics Processing Unit*.
- I2I: *Image-to-image*.
- LDDMM: *Large Deformation Diffeomorphic Metric Mapping*.
- lNCC: *Locally Normalized Cross Correlation*.
- MI: *Mutual Information*.
- MRI: *Magnetic Resonance Imaging*.
- MSE: *Mean Squared Error*.
- OASIS: *Open Access Series of Imaging Studies*.
- TC: *Tomografía computarizada*.
- VRAM: *Vídeo Random Access Memory*.

Índice

1. Introducción

1.1. Motivación y contexto

El registro de imágenes es una técnica de visión por computador cuyo objetivo es el de superponer dos o más imágenes de la misma escena tomadas en momentos diferentes desde distintos puntos de vista o con sensores diferentes. El registro de imágenes es fundamental en múltiples campos desde la teledetección (*remote sensing*) para la vigilancia del medio ambiente; cartografía para la actualización de mapas, y previsión meteorológica; hasta la localización y seguimiento de objetivos, control de calidad automático o detección de cambios en videovigilancia [18]. En medicina, el registro se utiliza para el seguimiento de la evolución de tumores, la planificación de ciertas cirugías, la comparación de los datos del paciente con referencias anatómicas, etc.

En particular, el registro de imágenes médicas busca encontrar una transformación espacial óptima de forma que alinee lo mejor posible las estructuras anatómicas subyacentes [9]. Los métodos de registro de imágenes médicas se pueden clasificar desde diferentes perspectivas. Desde el punto de vista de las imágenes que intervienen, se puede diferenciar entre un registro unimodal, multimodal, inter-paciente o intra-paciente. Desde el punto de vista del modelo de deformación, se obtienen métodos de registro rígido, afín y deformable (también conocido como no-rígido) [4]. Desde el punto de vista de la formulación del problema y su forma de optimización se distingue entre los métodos tradicionales y los basados en aprendizaje profundo [37, 26].

Los métodos tradicionales formulan el registro como un problema de optimización continua o discreta que busca un mínimo local de una función de energía de forma que se maximice la similitud de las imágenes después del registro de acuerdo con la métrica elegida. Sin embargo, los métodos iterativos de optimización no lineal pueden llegar a ser muy costosos temporalmente, limitando su aplicación práctica en entornos clínicos [6]. El aprendizaje profundo, impulsado por el éxito de las redes neuronales convolucionales (CNN) en visión artificial, ha cambiado el panorama de la investigación en el registro de imágenes médicas. En la última década se han desarrollado numerosas soluciones para el problema de registro basados en diferentes arquitecturas, funciones de pérdida, etc [2, 3, 4, 5, 6, 10, 12, 14, 16, 19, 26, 32, 33]. Sin embargo, la mayoría de estos métodos se centran en resolver problemas de registro unimodal con parametrización afín o no-rígida. Nos encontramos ante una gran falta de soluciones que aborden el problema de registro multimodal [5, 6, 33] debido a la alta dificultad del problema.

Los métodos basados en Redes Generativas Antagónicas (GAN) han demostrado un gran rendimiento en tareas de traducción de imagen a imagen (I2I), principalmente en imágenes naturales [17, 38]. Estos métodos ofrecen la posibilidad de convertir el registro multimodal en un problema unimodal que pueda ser abordado mediante la amplia gama de métodos de registro unimodal. Sin embargo, esta idea ha sido muy parcialmente explorada. Faltan resultados cuantitativos debido a la escasez de datos adecuadamente procesados en entornos multimodales. Por ello, el potencial de las GAN en estos escenarios aún está por explorarse [19].

Este trabajo se centra en explorar una posible solución al problema de registro no-rígido multimodal (inter- e intra- paciente) basado en métodos de aprendizaje profundo no supervisado. Nuestra solución consiste en encontrar una transformación óptima no rígida, que alinee las estructuras anatómicas de dos imágenes 3D del cerebro por resonancia magnética (MRI) de diferente modalidad (T1 y T2). Para ello, se propone una arquitectura que combina un método de traducción de imagen a imagen (CycleGAN) [17] para la traducción de la modalidad T2 a T1, junto con el uso de un método de registro no supervisado (VoxelMorph) [26] para calcular la transformación. Para evaluar el rendimiento del método propuesto, se ha realizado una comparativa con un método tradicional multimodal (Stationary LDDMM [7, 8]) y con un método del estado del arte basado en deep learning (SynthMorph [33]). La implementación de los sistemas y el preprocesado de las imágenes se ha realizado en Python 3.6.15, mediante el uso de la librería de PyTorch (1.10.1+cu102). El entrenamiento y test de la arquitectura de CycleGAN se ha realizado sobre una tarjeta gráfica NVIDIA GeForce RTX 3090 ti de 24 GBs. En el caso de la arquitectura de VoxelMorph se ha realizado sobre una NVIDIA GeForce GTX 1080 Ti de 11 GBs. Todo el código implementado para este trabajo puede encontrarse en: https://github.com/DavidSolanas/TFM_Unizar.

1.2. Estado del arte

En los últimos años, las redes neuronales convolucionales se han aplicado con un éxito sin precedentes a diferentes problemas de visión por computador. En particular, estos métodos han ganado una creciente popularidad en las soluciones del problema de flujo óptico y registro de imágenes. El registro de imágenes médicas no es una excepción, y se han propuesto numerosos enfoques basados en el aprendizaje profundo en la literatura. Sin embargo, la mayoría de las técnicas utilizadas se limitan a resolver formulaciones planteadas desde el registro tradicional. Además, muy pocos abordan el problema de registro no-rígido multimodal.

Los trabajos de registro basados en aprendizaje profundo se pueden clasificar en aprendizaje supervisado, donde se utilizan transformaciones *ground truth* que suelen provenir de algún método tradicional, o en métodos de aprendizaje no supervisado, que utilizan las métricas tradicionales de similitud de imágenes y regularización. Los métodos de aprendizaje no supervisado se suelen preferir sobre los supervisados, ya que las transformaciones se pueden aprender directamente a partir de pares de imágenes, y así evitar la sobrecarga de calcular las transformaciones que forman parte del *ground truth* para el entrenamiento. Por este motivo, se decidió trabajar en la propuesta de una solución de registro no supervisada.

Los métodos de registro multimodal basados en aprendizaje profundo no supervisado más relacionados con nuestro trabajo son los siguientes:

SynthMorph se propuso en 2022 como un método de registro no-rígido multimodal independiente de la imagen [33]. Para ello, primero se reemplazan los volúmenes a registrar por sus segmentaciones mediante el algoritmo SynthSeg [28], y después se calcula la transformación óptima para alinear las estructuras anatómicas de las segmentaciones. SynthMorph es capaz de registrar con precisión volúmenes cerebrales 3D de diferentes modalidades. Sin embargo, este método tiene como principal limitación la dependencia con SynthSeg en la obtención de las imágenes de entrada. Si las segmentaciones no se realizan de acuerdo a los requisitos de SynthSeg, entonces no se puede llevar a cabo el registro multimodal.

En 2022 también se propuso un método de registro no-rígido multimodal basado en la traducción de imagen a imagen para simplificar el problema de registro multimodal en uno unimodal [2]. El método propuesto utiliza la combinación de un modelo generativo y un método de registro unimodal (VoxelMorph). Los datos utilizados para el entrenamiento del método en este estudio provienen del desafío Learn2Reg-2021 (https://learn2reg.grand-challenge.org/Learn2Reg2021/) y corresponden a imágenes por resonancia magnética (MRI) y tomografías computarizadas (TC) abdominales intra-paciente, así como TC pulmonares intra-pacientes. Los resultados demuestran que la arquitectura propuesta en este trabajo logra un gran rendimiento al realizar el registro.

SNMBID se propuso en 2022 como un método para generar datos de entrenamiento para que los modelos aprendan registro multimodal [6]. El objetivo de este trabajo es la generación de imágenes cerebrales de modalidad no finita fusionando aleatoriamente algunas estructuras anatómicas finas y muestreando las intensidades para cada estructura anatómica fina utilizando una distribución gaussiana aleatoria. Además, se propone una mejora de la arquitectura Autoencoder Variacional 3D (*Variational Autoencoder*, VAE) para obtener transformaciones más realistas como *ground truth*. SNMBID puede ser utilizado para entrenar y evaluar otros métodos de registro de imágenes cerebrales.

RCV-Net se ha propuesto en noviembre de 2023 como una mejora de la arquitectura de VoxelMorph [26, 32] para el registro no-rígido multimodal a partir de MRI ponderadas en T1 y T2 [5]. RCV-Net incluye el módulo de atención de bloque convolucional (*Convolutional Block Attention Module*, CBAM [13]) durante el proceso de convolución con el objetivo de mejorar las capacidades de extracción de información durante el entrenamiento. Los resultados demuestran que la arquitectura propuesta en este trabajo supera a otros trabajos del estado del arte.

Finalmente, se propuso en 2022 un estudio de la aplicabilidad de los métodos modernos de traducción de imagen a imagen (I2I) para la tarea de registro ***rígido*** de imágenes médicas multimodales en 2D y 3D [19]. En este trabajo, se comparó el rendimiento de varios métodos de traducción I2I basados en Redes Generativas Antagónicas (GAN), posteriormente combinados con métodos de registro unimodal representativos, para así evaluar la efectividad de la traducción de modalidad para el registro rígido de imágenes médicas multimodales. Las conclusiones obtenidas de este trabajo indican que un enfoque de aprendizaje para mapear las modalidades a un "terreno común" en lugar de una modalidad a la otra directamente, podría ser un enfoque muy prometedor.

1.3. Objetivos del trabajo y organización

El objetivo de este trabajo es estudiar la problemática de las soluciones de aprendizaje profundo en el registro no-rígido de imágenes médicas de diferentes modalidades. Como primer objetivo, se plantea realizar un estudio de las limitaciones de los métodos directamente basados en el entrenamiento a partir de parejas de imágenes de modalidades diferentes y las soluciones basadas en segmentaciones. Como segundo objetivo, se implementa una solución propia basada en métodos de transferencia de estilo entre modalidades mediante la arquitectura CycleGAN. Dicha solución se evalúa respecto al estado del arte actual del problema.

La estructura de la memoria es la siguiente: En la Sección 1 se presenta la motivación y contexto del problema a resolver, el estado del arte y los objetivos que se pretenden conseguir en este trabajo. En la Sección 2 se presentan los métodos implementados para resolver el problema planteado. Más concretamente, se detalla el conjunto de datos utilizado, junto con el preprocesado propuesto, el método implementado para la transferencia de estilo entre modalidades mediante la arquitectura CycleGAN, el método utilizado de registro no-rígido VoxelMorph, y la solución final desarrollada. En la Sección 3 se detallan las distintas métricas para la evaluación de los métodos desarrollados y los resultados obtenidos. Por un lado, se presentan los resultados obtenidos con la implementación 3D de CycleGAN. Por otro lado, los resultados de los métodos de registro no-rígido desarrollados junto con los resultados obtenidos con un método tradicional y un método del estado del arte actual. Finalmente, en la Sección 4 se detallan las conclusiones obtenidas tras el desarrollo de este Trabajo de Fin de Máster.

1.4. Diagrama de Gantt

Este trabajo ha sido desarrollado a lo largo de un año. La primera fase del trabajo ha comprendido el estudio previo del problema y la familiarización de las herramientas disponibles para abordarlo. A partir de ahí el trabajo se ha focalizado principalmente en la obtención y preprocesado de los datos, así como los entrenamientos de los distintos modelos de aprendizaje profundo a utilizar. Esta fase ha comprendido la gran mayoría del esfuerzo y dedicación del trabajo. Finalmente, la validación de los modelos entrenados y el desarrollo de la memoria del trabajo. En la Figura 1.1 se puede observar el diagrama de Gantt de este trabajo.

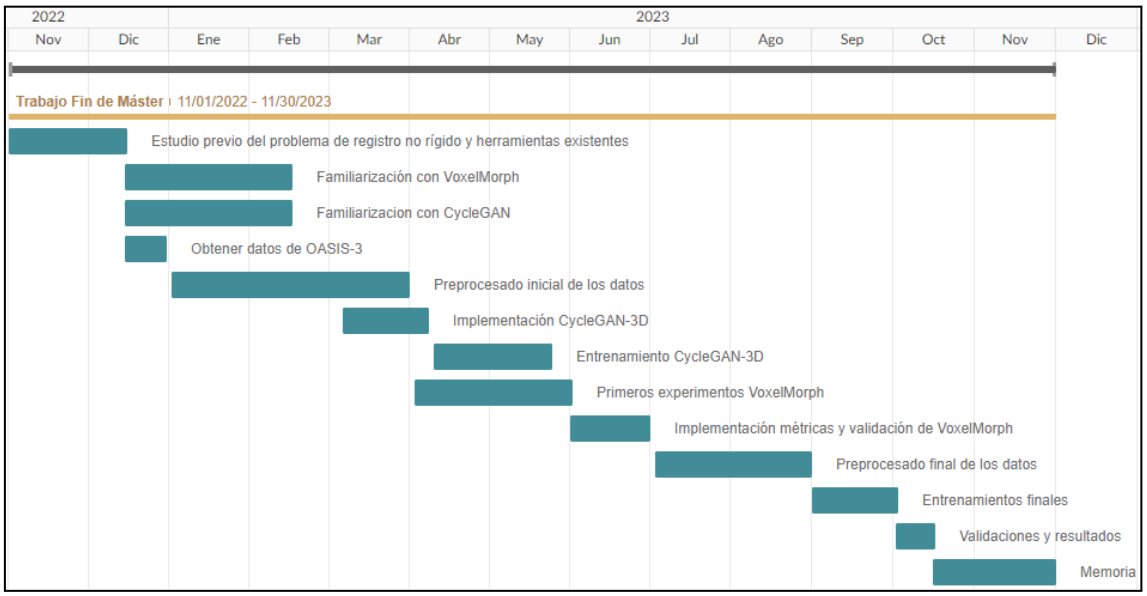


Figura 1.1: Diagrama de Gantt del Trabajo de Fin de Máster.

2. Métodos

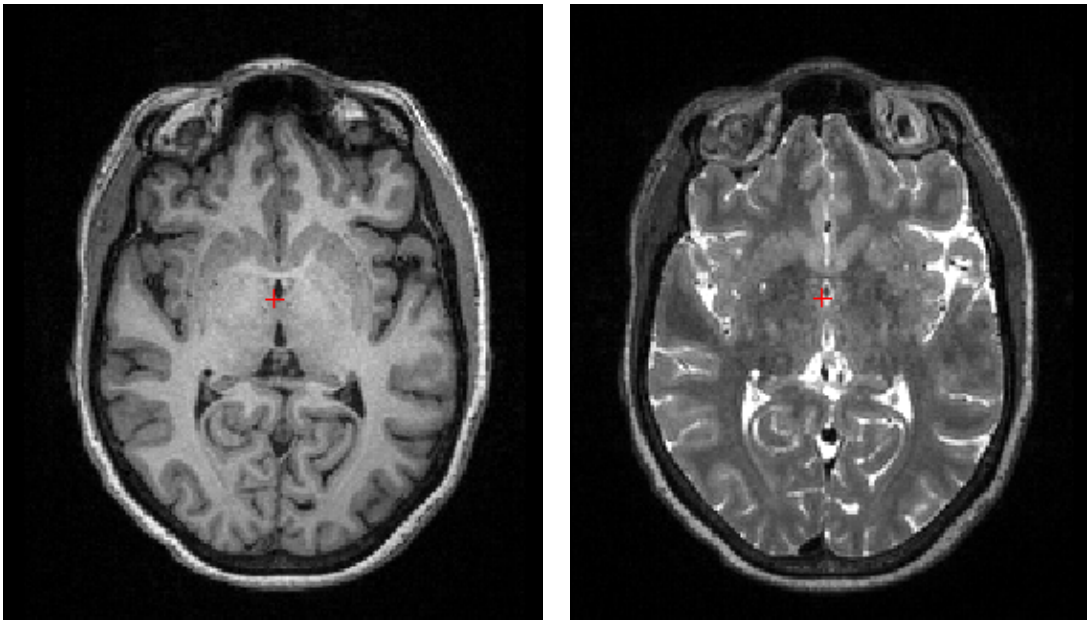
En esta sección se presenta la combinación de métodos propuesta en este trabajo para resolver el problema planteado. En primer lugar, se describe el conjunto de datos utilizados y el preprocesado realizado para el entrenamiento de CycleGAN y VoxelMorph. A continuación, se presentan los métodos utilizados para la traducción entre las diferentes modalidades de imagen y el registro no-rígido de las mismas. Finalmente, se describe la *pipeline* adoptada para realizar el registro multimodal.

2.1. Datos

Para la generación de los modelos y la evaluación del método de registro se ha utilizado una base de datos de resonancias magnéticas 3D del cerebro ponderadas en T1 y T2 (weighted T1 - T2 magnetic resonance image, MRI). Las imágenes han sido obtenidas del proyecto OASIS-3 (https://www.oasis-brains.org/). OASIS-3 es una colección de imágenes de resonancia magnética y datos clínicos de 1098 participantes que se recogieron en varios estudios en curso en el Centro de Investigación de la Enfermedad de Alzheimer Knight de la Universidad de Washington a lo largo de 15 años [11]. Su utilización es libre a cambio del reconocimiento correspondiente con el objetivo de avanzar en el conocimiento de la enfermedad.

La base de datos utilizada en este TFM contiene 1615 pares de volúmenes T1-T2 de 1023 pacientes distintos, se descartan 75 pacientes de los 1098 originales por no tener una MRI de cada modalidad disponible. Los volúmenes tienen una dimensión de 176 x 256 x 256 vóxeles de tamaño 1 mm en los ejes (x, y, z). En la Figura 2.1 se puede observar un ejemplo de un par de imágenes T1 y T2 obtenidas de un mismo paciente.

Las imágenes de resonancia magnética (MRI) ponderadas en T1 y T2 son dos modalidades diferentes que proporcionan información sobre los tejidos del cuerpo. Las principales diferencias entre ellas radican en el modo en que los distintos tejidos reaccionan a las variaciones de la secuencia de pulsos magnéticos, así como en las características anatómicas y patológicas que resalta cada modalidad. Los tejidos con alto contenido en grasa son brillantes en las imágenes ponderadas en T1, y los tejidos con alto contenido en agua se destacan en las ponderadas en T2. Esto provoca notables variaciones en sus contrastes, que llaman la atención sobre diferentes características anatómicas. En el cerebro, las imágenes de T1 son útiles para resaltar estructuras como el líquido cefalorraquídeo y para visualizar eficazmente la anatomía cerebral básica y el contraste entre tejidos blandos y huesos. Sin embargo, las enfermedades como la inflamación, el edema y las lesiones de los tejidos blandos que incluyen cambios en el contenido de agua pueden identificarse con ayuda de las imágenes T2. La modalidad T1, en general, ofrece una clara representación visual de la anatomía estructural, mientras que la modalidad T2 destaca cambios en la composición de los tejidos.



(a) Modalidad A. (b) Modalidad B.

Figura 2.1: Corte axial de un par de imágenes T1 y T2 de un paciente extraídas del conjunto de datos OASIS-3 utilizado en este trabajo. (a) Modalidad A: imagen de MRI T1, (b) Modalidad B: imagen de MRI T2.

2.2. Preprocesado propuesto

Dada la gran variabilidad de los datos proporcionados por OASIS, se ha constatado la necesidad de realizar un preprocesado de los datos para que los modelos se comporten correctamente durante el aprendizaje. Se propone un preprocesado consistente en los siguientes pasos. En primer lugar, se ha utilizado la herramienta FreeSurfer [20], junto con su comando recon-all, para realizar los 4 primeros pasos del procesado que ofrece [21]. Estos son:

1. Corrección de Movimiento y Conformidad: Corrección de posibles movimientos del sujeto durante la adquisición de las imágenes de MRI.

2. NU (Normalización de Intensidad No Uniforme): Normalización de la intensidad de las imágenes, asegurando que el brillo y el contraste sean consistentes en todo el conjunto de datos.

3. Cálculo de la Transformación Talairach: Transformación que mapea las imágenes cerebrales a un espacio cerebral común, llamado espacio de Talairach.

4. Normalización de Intensidad 1: Normalización de intensidad que ayuda a garantizar que la intensidad de los datos de MRI sea consistente entre sujetos.

En segundo lugar, se ha utilizado la herramienta ROBEX [22] para la extracción del cráneo de los volúmenes pues los métodos de registro muestran mejores resultados en el interior del cerebro si la imagen no tiene el cráneo presente. Aunque también se consideró FreeSurfer para esta tarea, ROBEX mostró mejores resultados en las resonancias ponderadas en T2.

En tercer lugar, los volúmenes sin cráneo se han registrado de forma afín al atlas multimodal ICBM 152 [23] con la versión multimodal de la herramienta ANTs [24]. De esta forma se consigue que todos los datos se encuentren alineados de forma afín a su correspondiente modalidad en el atlas de referencia. Este registro ha resultado ser fundamental en el preprocesado, permitiendo la mejora de los resultados obtenidos con las distintas redes neuronales.

Finalmente, se ha hecho una revisión manual para comprobar que los datos utilizados habían completado el preprocesado correctamente, ya que el registro afín tiene cierta probabilidad de fallar en función de la calidad de las extracciones realizadas por ROBEX. De los volúmenes finales, se ha elegido un subconjunto de 448 pares de volúmenes T1-T2 para el entrenamiento (402 pares), validación (17 pares) y test (29 pares) de las redes neuronales. Tras el preprocesado y selección de imágenes, los volúmenes se encuentran uniformizados a un sistema de coordenadas común de dimensión de 193 x 239 x 263. No obstante, debido al alto consumo de recursos de la fase de entrenamiento de los métodos utilizados en este trabajo, los volúmenes se han recortado y submuestreado finalmente a un tamaño de 128 x 128 x 128.

Con los volúmenes finales de los conjuntos de validación y test, se han calculado las segmentaciones de dichos volúmenes para la evaluación de los métodos de registro no rígido. Una segmentación es una asignación de las intensidades de la imagen y estructuras anatómicas específicas. En este trabajo se ha utilizado la herramienta de segmentación SynthSeg [28], consistente en un método de aprendizaje profundo especializado en la segmentación de estructuras cerebrales robusto a cualquier modalidad. En la Figura 2.2 se muestra un ejemplo de una segmentación de un sujeto del conjunto de datos de test.

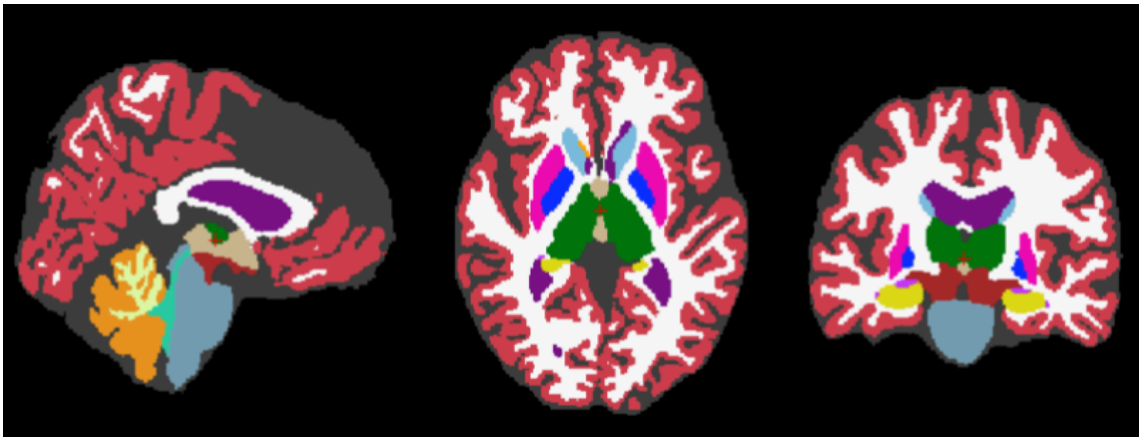


Figura 2.2: Segmentación calculada con SynthSeg a partir de un volumen T1 del conjunto de datos de test. Las distintas estructuras anatómicas aparecen codificadas por colores. Por ejemplo, la materia blanca cerebral en blanco, la corteza cerebral en rojo, el cerebelo en naranja y amarillo, los ventrículos en violeta, etc.

2.3. CycleGAN

El primer paso del método de registro multimodal propuesto es la transformación de un volumen T2 en su equivalente en la modalidad T1. Para ello, se ha utilizado el modelo generativo propuesto en CycleGAN [17]. CycleGAN es una red generativa antagónica (GAN) [25] que puede ser entrenada de forma no supervisada utilizando dos grupos de imágenes o volúmenes no emparejados, para traducir datos

entre dos dominios de intensidades (dominio A y dominio B), que en nuestro caso se corresponden con el dominio T1 y el dominio T2. La idea principal de CycleGAN es hacer que la traducción de la imagen sea "coherente con el ciclo", es decir, si una imagen se traduce del dominio A al B y luego se traduce inversamente de B a A, la salida debe ser lo más parecida posible que la imagen original.

CycleGAN consta de cuatro arquitecturas principales, dos generadores (G_{A2B} y G_{B2A}) y dos discriminadores (D_A y D_B). Los dos generadores producen imágenes del dominio A/B basadas en las imágenes del dominio B/A. Los dos discriminadores distinguen cada imagen como sintética o real. El entrenamiento no supervisado está regularizado por la consistencia del ciclo [1, 17]

$$\begin{aligned} G_{B2A}(G_{A2B}(I_A)) &\approx I_A, \\ G_{A2B}(G_{B2A}(I_B)) &\approx I_B, \end{aligned} \tag{1}$$

donde I_A y I_B son dos imágenes de los dominios A y B.

La principal peculiaridad de las GANs es que ambos componentes (generadores y discriminadores) son entrenados conjuntamente. El discriminador se entrena para identificar las imágenes generadas como falsas, mientras que el generador intenta engañar al discriminador. De este modo, generadores y discriminadores compiten entre sí. Una vez alcanzada la convergencia, el generador debería obtener muestras indistinguibles para el discriminador en el dominio objetivo [25].

2.3.1. Arquitectura

Como se ha mencionado anteriormente, CycleGAN contiene dos generadores (G_{A2B} y G_{B2A}) y dos discriminadores (D_A y D_B). D_B condiciona a que G_{A2B} traduzca I_A en una salida indistinguible del dominio B, y viceversa para D_A y G_{B2A}. En la Figura 2.3 se puede observar un esquema general de la arquitectura de CycleGAN, así como de la arquitectura de los generadores y discriminadores.

Más concretamente, los generadores siguen una arquitectura *encoder-decoder*. La arquitectura *encoder-decoder* transforma la entrada a una representación intermedia (*encoding*), para obtener información de salida a partir de esa representación. Los generadores de CycleGAN están constituidos por varias capas convolucionales (*encoder*), bloques residuales (*encoding*), y capas de deconvolución (*decoder*).

La arquitectura utilizada por los discriminadores en CycleGAN sigue la estructura convolucional habitual. El discriminador está compuesto por capas convolucionales para la extracción de características discriminativas de la imagen. Para cada capa convolucional se utiliza también una capa de normalización, en el caso de CycleGAN se utilizan las capas de normalización por instancia (*instance normalization layer*). Como peculiaridad de CycleGAN, para la última capa del discriminador se utiliza la arquitectura PatchGAN [17]. Con PatchGAN el discriminador analiza pequeños parches locales de la imagen, en lugar de la imagen completa, para determinar si la imagen es real o no. Más que ofrecer una evaluación general, esta arquitectura ayuda a captar las características locales y proporciona información más profunda sobre la autenticidad de los *patches* analizados.

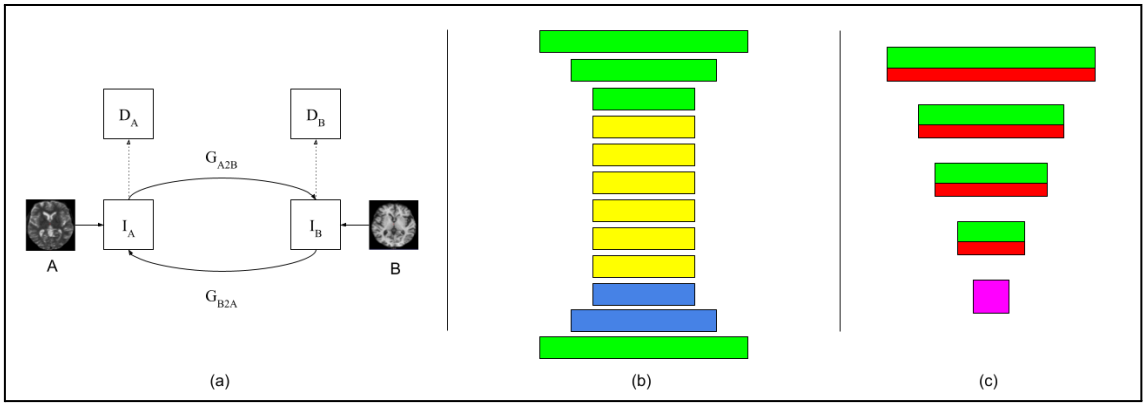


Figura 2.3: Arquitectura de CycleGAN. (a) Arquitectura completa de CycleGAN. (b) Arquitectura de los generadores (G_{A2B} y G_{B2A}). En verde las capas de convolución, en amarillo los bloques residuales (*residual block*) y en azul las capas de convolución transpuesta. (c) Arquitectura de los discriminadores (D_A y D_B). En verde las capas de convolución, en rojo las capas de normalización por instancia (*instance normalization layer*) y en magenta la capa convolucional *PatchGAN*.

2.3.2. Entrenamiento de CycleGAN

El objetivo de CycleGAN es aprender funciones de mapeo entre dos dominios A y B, dadas las imágenes durante el entrenamiento como $\{a_i\}_{i=1}^{N}$ donde $a_i \in$ A y $\{b_j\}_{j=1}^{M}$ donde $b_j \in$ B. Se denota la distribución de los datos como: $a \sim p_{data}(a)$ y $b \sim p_{data}(b)$. Como se ha mencionado anteriormente, CycleGAN está compuesta por dos generadores G_{A2B} y G_{B2A} que realizan sendos mapeos G: A→B y F: B→A, respectivamente. Además tiene dos discriminadores, D_A y D_B, donde D_A tiene como objetivo distinguir entre las imágenes reales $\{a\}$ y las imágenes generadas $\{F(b)\}$, y D_B tiene como objetivo distinguir entre las imágenes reales $\{b\}$ y las imágenes generadas $\{G(a)\}$. El entrenamiento de CycleGAN viene definido por dos funciones de pérdida: la función de pérdida antagónica [25], que iguala la distribución de las imágenes generadas con la distribución de los datos en el dominio objetivo; y la función de pérdida de consistencia del ciclo, que evita que los mapeos aprendidos G y F se contradigan entre sí [17].

La función de pérdida antagónica es aplicada a ambas funciones de mapeo (G y F). Para la función de mapeo G y su discriminador D_B, la función de pérdida está definida por

$$L_{GAN}(G, D_B, A, B) = \mathbb{E}_{b \sim pdata(b)}[\log D_B(b)] + \mathbb{E}_{a \sim pdata(a)}[\log(1 - D_B(G(a))], \quad (2)$$

donde G trata de generar imágenes $G(a)$ que sean similares a imágenes del dominio B, mientras D_B trata de discriminar entre imágenes generadas $G(a)$ e imágenes reales b. G tiene como objetivo minimizar esta función de pérdida frente al discriminador D_B que intenta maximizarlo, es decir

$$\min_G \max_{D_B} L_{GAN}(G, D_B, A, B). \quad (3)$$

Asimismo, para la otra función de mapeo F: B→A y el discriminador D_A se utiliza la misma función de pérdida antagónica

$$\min_{F} \max_{D_A} L_{GAN}(F, D_A, B, A). \quad (4)$$

La función de pérdida de consistencia de ciclo se introduce debido a que el uso de la función de pérdida antagónica por sí sola no garantiza que la función de mapeo aprendida sea capaz de mapear una imagen de entrada a_i a una imagen de salida b_i. Lo que garantiza esta función de pérdida es que las funciones de mapeo aprendidas sean "coherentes con el ciclo". Dicha función de pérdida viene definida por

$$L_{\text{cyc}}(G, F) = \mathbb{E}_{a \sim pdata(a)}[\|F(G(a)) - a\|_1] + \mathbb{E}_{b \sim pdata(b)}[\|G(F(b)) - b\|_1]. \quad (5)$$

De este modo, la función de pérdida utilizada para el entrenamiento de CycleGAN viene definida por

$$L(G, F, D_A, D_B) = L_{\text{GAN}}(G, D_B, A, B) + L_{\text{GAN}}(F, D_A, B, A) + \lambda L_{\text{cyc}}(G, F), \quad (6)$$

donde λ es un parámetro de ajuste entre las funciones de pérdida propias de la GAN y la función de consistencia de ciclo.

2.3.3. Detalles de implementación

El código original de CycleGan [17], admite como entrada imágenes RGB 2D, de modo que ha sido necesario adaptar la red para que funcione con imágenes en escala de grises 3D. Se ha adaptado el código base, implementando un modelo equivalente en 3D, sustituyendo las capas de convoluciones 2D por sus respectivas en 3D. Se ha implementado la gestión de datos, así como distintas funciones de procesado y visualización para tratar con los volúmenes 3D. Para ello, el trabajo realizado en [1] ha servido de guía. La arquitectura adaptada a 3D de los generadores (G_{A2B} y G_{B2A}) y de los discriminadores (D_A y D_B) implementada en este trabajo se encuentra en las Tablas 1 y 2 respectivamente.

Capa	Tipo de capa	Número de filtros	Tamaño de filtro	Stride	Función de activación
1	Submuestreo (Convolución 3D)	32	7x7x7	1	ReLU
2	Submuestreo (Convolución 3D)	64	3x3x3	2	ReLU
3	Submuestreo (Convolución 3D)	128	3x3x3	2	ReLU
4-9	Bloque residual	128	3x3x3	1	-
10	Expansión (Convolución 3D transpuesta)	64	3x3x3	2	ReLU
11	Expansión (Convolución 3D transpuesta)	32	3x3x3	2	ReLU
12	Expansión (Convolución 3D)	1	7x7x7	1	tanh

Tabla 1: Arquitectura de los generadores utilizados en la implementación 3D de CycleGAN.

Capa	Tipo de capa	Número de filtros	Tamaño de filtro	Stride	Función de activación
1	Submuestreo (Convolución 3D)	64	4x4x4	2	LeakyReLU (0.2)
2	Submuestreo (Convolución 3D)	128	4x4x4	2	LeakyReLU (0.2)
3	Submuestreo (Convolución 3D)	256	4x4x4	1	LeakyReLU (0.2)
4	Submuestreo (Convolución 3D)	512	4x4x4	1	LeakyReLU (0.2)
5	Convolución 3D (*PatchGAN*)	1	4x4x4	1	Sigmoidal

Tabla 2: Arquitectura de los discriminadores utilizados en la implementación 3D de CycleGAN.

El entrenamiento de la red se ha realizado durante 200 épocas con una tasa de aprendizaje (*learning rate*) de 0.0002. Durante las 100 primeras épocas se mantiene la tasa de aprendizaje constante y se disminuye linealmente hasta 0 en las siguientes 100 épocas. Se ha utilizado el optimizador Adam [31] con un tamaño de lote (*batch*) de 1. El entrenamiento ha tenido un coste en memoria de 18 GBs y ha durado aproximadamente 9 días en una tarjeta gráfica GeForce RTX 3090 ti de 24 GBs.

2.4. VoxelMorph

Para el registro no rígido de las imágenes se ha utilizado un modelo generado a partir de VoxelMorph, por ser uno de los métodos de registro más populares y versátiles del reciente estado del arte [26, 32]. Más concretamente, se ha utilizado la implementación de VoxelMorph que se proporciona en [27].

VoxelMorph es un método basado en deep-learning para el aprendizaje no supervisado del registro no rígido en pares de imágenes médicas. VoxelMorph tiene una arquitectura basada en U-Net, propuesta originalmente para la segmentación de imágenes [34] y utilizada con éxito en diferentes aplicaciones médicas. El método aprende de forma no supervisada la transformación existente entre parejas de imágenes minimizando diferentes métricas de similitud sobre la intensidad de las imágenes [26]. Estas métricas, junto con la energía de regularización conforman la función de pérdida.

Para este trabajo se ha utilizado la versión de VoxelMorph que parametriza las transformaciones mediante *small deformations* sin control del Jacobiano. Esta elección se justifica debido a que el modelo resultante es más sencillo de controlar durante el entrenamiento. Como inconveniente, la transformación calculada por VoxelMorph no es difeomorfa. Al utilizar *small deformations*, la parametrización de la transformación (ϕ) necesaria para resolver el problema viene dada por

$$\phi = Id + u, \tag{7}$$

donde Id es la transformación identidad, y $u : \Omega \rightarrow \mathbb{R}^d$ es el campo de desplazamiento de la deformación de una imagen x (*moving*) en la imagen y (*fixed*). La transformación ϕ se obtiene a través de la minimización de la función de pérdida L

$$\phi' = arg_{\phi}\ min\ L(x, y, \phi), \tag{8}$$

que es la misma que la utilizada en los métodos tradicionales y se define a partir de la contribución ponderada de la similitud de ambas imágenes x e y y la energía de regularización

$$L(x, y, \phi) = L_{sim}(y, x \circ \phi) + \lambda L_{reg}(\phi), \quad (9)$$

donde λ es un parámetro regularizador utilizado para evitar el sobreajuste durante el entrenamiento y ◦ es la composición de funciones.

Las funciones de pérdida más utilizadas para L_{sim} son el error cuadrático medio (MSE), basado en la suma de diferencias al cuadrado (SSD), la correlación cruzada normalizada local (lNCC) y la información mutua (MI) [26, 36]. MSE es aplicable cuando x e y tienen distribuciones de intensidad de imagen similares. lNCC y MI son más robustas que MSE ante variaciones de intensidad entre las muestras de un conjunto de datos. En particular, MI ha sido propuesta para el problema de registro multimodal en métodos tradicionales.

La minimización de L_{sim} permite estimar una transformación de forma que $x \circ \phi$ se aproxime a y, aunque puede generar una transformación ϕ no suave que no sea físicamente realista. Es por ello que se introduce en la Ecuación 9 la energía de regularización, definida como

$$L_{reg}(\phi) = \int_{\Omega} \|\nabla u(x)\|_2^2 d\Omega \quad (10)$$

de forma que se controle la magnitud del gradiente de los desplazamientos. Para calcular la composición $x \circ \phi$, la arquitectura U-Net se combina con transformador espacial (*Spatial Transformer*), introducido en [35].

2.4.1. Detalles de implementación

Entre las versiones de VoxelMorph que ofrece el código de [27], se ha utilizado VoxelMorph-II, que tiene un mayor consumo de recursos pero produce mejores resultados. Para el entrenamiento, validación y test se han utilizado los mismos volúmenes que en la red CycleGAN (punto 2.2).

El entrenamiento de la red se ha realizado durante 100 épocas con una tasa de aprendizaje (*learning rate*) de 0.0001. La tasa de aprendizaje se actualiza en cada época siguiendo un decaimiento de potencia (*power decay*) hasta 0 en la época 100. Se ha utilizado el optimizador Adam [31] con un tamaño de lote (*batch*) de 1. El entrenamiento ha tenido un coste en memoria de 3.3 GBs y ha durado aproximadamente 16 horas en una tarjeta gráfica NVIDIA GeForce GTX 1080 Ti de 11 GBs.

2.5. Pipeline de registro multimodal

En este trabajo se propone aproximar el problema de registro multimodal mediante la combinación de CycleGAN y VoxelMorph. El método propuesto consiste en realizar un cambio de dominio de T2 a T1 del volumen que se quiere registrar (*moving*), para que esté en el mismo dominio de intensidad que el volumen de referencia (*fixed*) y realizar un registro unimodal mediante VoxelMorph que proporcione la transformación que maximice la similitud entre ambas imágenes.

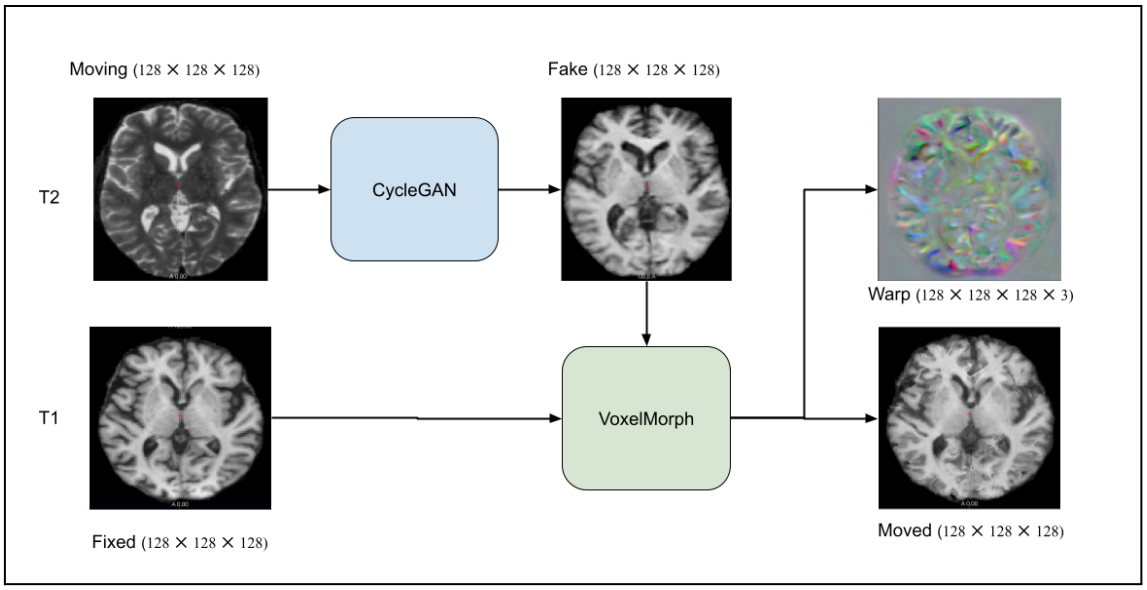


Figura 2.4: Pipeline de registro multimodal implementado. CycleGAN toma la imagen preprocesada *real* T2 (*moving*) y la transforma en la imagen *fake* T1. VoxelMorph toma la salida de CycleGAN (*fake*) y la registra con el volumen preprocesado *real* T1 (*fixed*). En azul, CycleGAN, que transforma el dominio T2 a T1. En verde, VoxelMorph que realiza el registro unimodal en el dominio T1.

La Figura 2.4 muestra la *pipeline* de la solución propuesta. Partiendo de los volúmenes preprocesados de dimensión 128 x 128 x 128 (punto 2.2). El volumen T1 será utilizado como referencia (*fixed*) durante el registro con VoxelMorph. El volumen T2 necesita el paso adicional de cambio de dominio. Para ello, se utiliza la red neuronal CycleGAN. El resultado será un volumen de dimensión 128 x 128 x 128 equivalente en el dominio de T1 (*fake*). Finalmente, se realiza un registro unimodal con VoxelMorph entre el volumen *fake* y *fixed*. VoxelMorph proporciona el volumen *fake* transformado (*moved*) junto con su campo de deformación (*warp*).

Durante el proceso de entrenamiento, ambos modelos son entrenados independientemente con los mismos volúmenes preprocesados. De esta manera, en el momento de realizar el registro multimodal siguiendo la *pipeline* de la Figura 2.4, se pueden utilizar directamente los volúmenes inferidos por CycleGAN para el registro con VoxelMorph. La fase de entrenamiento es computacionalmente costosa en memoria y tiempo. Una vez se tienen los modelos entrenados, la *pipeline* propuesta para realizar el registro multimodal es super eficiente, pues solo se realiza el proceso de inferencia en ambos modelos, tardando del orden de pocos segundos en completar todo el proceso.

3. Resultados

En esta sección se presentan los resultados obtenidos tras entrenar CycleGAN y VoxelMorph. En primer lugar, se definen las métricas utilizadas para la evaluación de los distintos métodos. A continuación, se presentan los resultados obtenidos con CycleGAN sobre el conjunto de datos de test. Finalmente, los experimentos realizados y los resultados obtenidos con el método propuesto para el registro multimodal, comparando los resultados con versiones multimodales del método tradicional *Large Deformation Diffeomorphic Metric Mapping* (Stationary LDDMM [7, 8]) y con un método del estado del arte de registro multimodal (SynthMorph [33]).

3.1. Métricas

Para determinar la precisión de los métodos implementados se han utilizado diferentes métricas. Para la evaluación de la implementación 3D de CycleGAN, se ha utilizado la Distancia de inicio de Fréchet. Para la evaluación de los métodos de registro se han utilizado el Coeficiente de Similitud de Dice y el determinante de la matriz jacobiana, dos métricas ampliamente utilizadas en la evaluación del registro.

3.1.1. Distancia de inicio de Fréchet

La Distancia de inicio de Fréchet (*Fréchet Inception Distance*, FID) es una métrica utilizada normalmente para determinar la calidad de las imágenes creadas por un modelo generativo, en este caso CycleGAN [30]. La métrica FID compara las distribuciones de un conjunto de imágenes reales con las del conjunto de imágenes generadas por la red. Una FID más baja (siendo una medida de distancia) indica una mayor similitud de las imágenes, lo que corresponde a una mayor calidad de la traducción de la imagen. Esta métrica se introdujo en [29] y se define como:

$$d^2(x, y) = |\mu_x - \mu_y|^2 + \mathrm{tr}[\sum_x + \sum_y - 2(\sum_x \sum_y)^{1/2}], \qquad (10)$$

donde x es la imagen generada por la red e y es la imagen real. μ_x, μ_y y $\sum_x, \sum_y$ son las respectivas medias y matrices de covarianza de x e y, y se toma la raíz cuadrada positiva. El operador tr corresponde a la traza (*trace*), que se define como la suma de los elementos de la diagonal principal de una matriz M.

3.1.2. Coeficiente de Similitud de Dice

El Coeficiente de Similitud de Dice (*Dice Similarity Coefficient*, DSC) es una métrica utilizada para comparar la similitud entre dos conjuntos. En el contexto de segmentación y registro de imágenes médicas, es una de las métricas más utilizadas. En este contexto, DSC es utilizado para estimar la superposición espacial de dos segmentaciones. Así, dadas dos segmentaciones S y T, el coeficiente de Dice se define como

$$\mathrm{DSC}\,(S, T) = \frac{2\,|S \cap T|}{|S| + |T|} \qquad (8)$$

Esta métrica proporciona el valor uno si S y T se solapan exactamente y disminuye gradualmente hacia cero en función del solapamiento de los dos volúmenes [8]. En el caso del registro, la utilización de segmentaciones en la evaluación surge de la falta de un *ground truth* establecido.

El algoritmo para calcular el DSC asociado a un método de registro es el siguiente:

1. Registrar un volumen (*moving*) con otro volumen de referencia (*fixed*), utilizando el método correspondiente. La salida será el volumen *moving* registrado (*moved*) junto con su campo de deformación (*warp*).

2. Obtener las segmentaciones calculadas con SynthSeg tanto de *moving* como de *fixed*.

3. Aplicar el *warp* a la segmentación de *moving* para deformarla según el campo de deformación calculado por el método utilizando interpolación *nearest neighbors*.

4. Calcular el DSC (ecuación 8) utilizando la segmentación deformada en el paso 3 y la segmentación de *fixed*.

3.1.3. Determinante de la matriz jacobiana

La matriz jacobiana de una función vectorial es la matriz de todas sus derivadas parciales de primer orden. Para un campo de deformación ϕ^{-1}, su matriz jacobiana es la derivada de primer orden de cada una de sus tres direcciones espaciales (*x, y, z*) respecto a las demás. En cualquier punto dado *p*, la matriz jacobiana se define como:

$$J_{\phi^{-1}}(p) = \begin{pmatrix} \frac{\partial \phi_x^{-1}(p)}{\partial x} & \frac{\partial \phi_x^{-1}(p)}{\partial y} & \frac{\partial \phi_x^{-1}(p)}{\partial z} \\ \frac{\partial \phi_y^{-1}(p)}{\partial x} & \frac{\partial \phi_y^{-1}(p)}{\partial y} & \frac{\partial \phi_y^{-1}(p)}{\partial z} \\ \frac{\partial \phi_z^{-1}(p)}{\partial x} & \frac{\partial \phi_z^{-1}(p)}{\partial y} & \frac{\partial \phi_z^{-1}(p)}{\partial z} \end{pmatrix} \quad (9)$$

A partir del determinante de la matriz jacobiana se puede obtener información acerca del campo de deformación calculado durante el registro. Si el determinante de la matriz jacobiana es positivo, es decir, $|J\phi^{-1}(p)| > 0$, indica que el campo de deformación calculado es localmente difeomorfo [10]. La transformación identidad tiene como valor del determinante 1. Determinantes con valores <1 indican contracción local de la deformación mientras que determinantes con valores >1 indican expansión local. Unos valores altos del determinante indican que el campo de deformación no es localmente suave, lo que en muchas aplicaciones clínicas no es deseable.

3.2. Resultados con CycleGAN

Los resultados obtenidos con nuestro método de CycleGAN entrenado se muestran en la Figura 3.1. Para poder visualizar y comparar de manera sencilla los volúmenes se han extraído las imágenes correspondientes a los 3 cortes del cerebro (sagital, axial y coronal, respectivamente) centrados en la imagen. Se puede observar que las salidas que produce la red son muy similares en apariencia a los volúmenes T1 de referencia correspondientes.

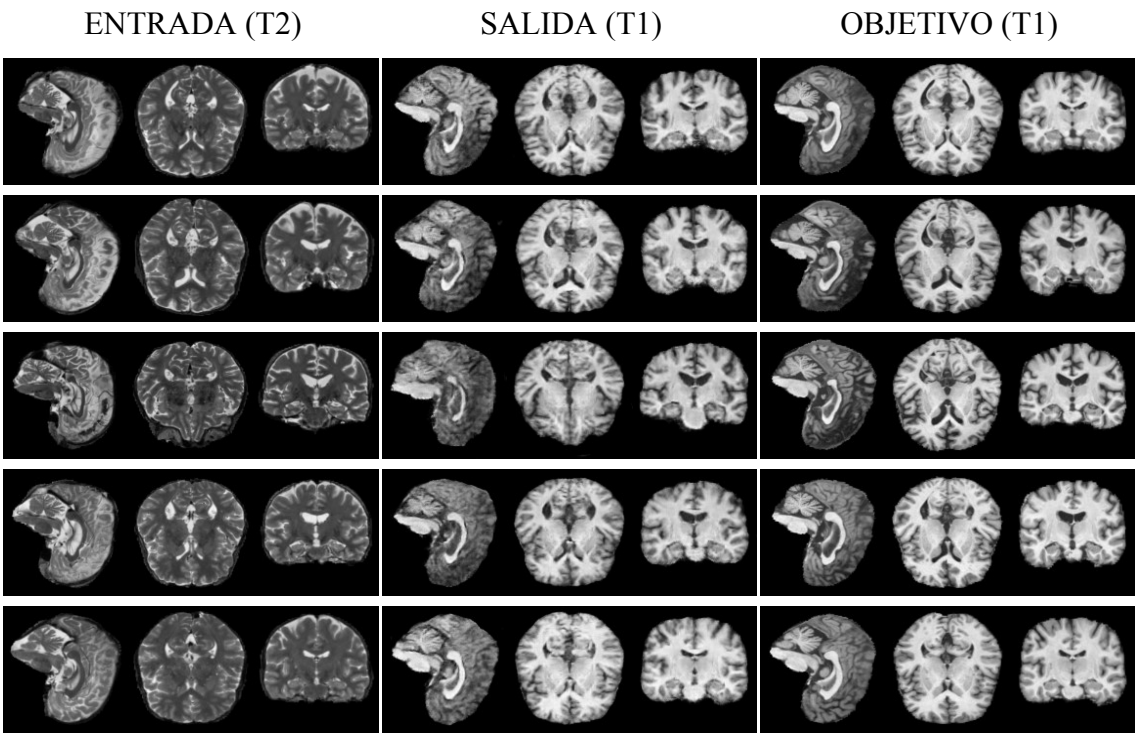


Figura 3.1: Resultados de la inferencia de CycleGAN de resonancias magnéticas ponderadas en T2. Cada fila representa un sujeto distinto del conjunto de datos de test. A la izquierda se encuentran los volúmenes T2 que conforman la entrada de la red. En el centro la salida de la misma, corresponde al cambio de dominio de T2 a T1 del volumen de entrada. A la derecha el volumen objetivo, es el volumen T1 de referencia.

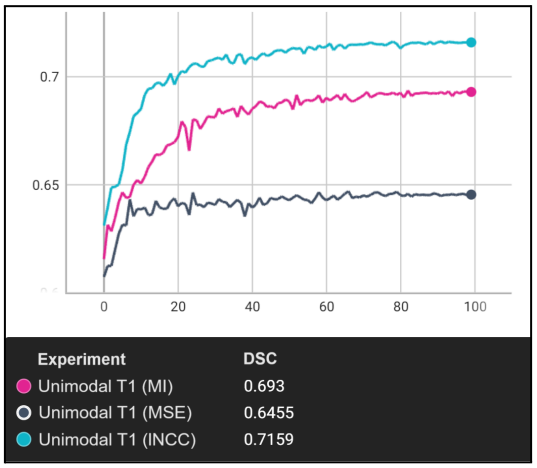
Para el cálculo de la métrica FID se han utilizado imágenes con el mismo formato que en la Figura 3.1. Para todo el conjunto de test, se ha obtenido por un lado la imagen del volumen real T1 y por otro lado la imagen generada por CycleGAN y se ha calculado su FID. Se ha calculado también la FID con las imágenes reales T1 y T2, y así poder cuantificar y comparar el rendimiento de CycleGAN. La FID media obtenida con las imágenes reales T1 y T2 ha sido de 17.999 ± 3.441, mientras que la FID media obtenida comparando la imagen real T1 con la imagen generada por CycleGAN ha sido de 3.325 ± 1.375. Con CycleGAN la diferencia entre los volúmenes T2 y T1 se reduce en un 81.52%.

3.3. Resultados de registro multimodal. VoxelMorph y métricas de similitud de imagen.

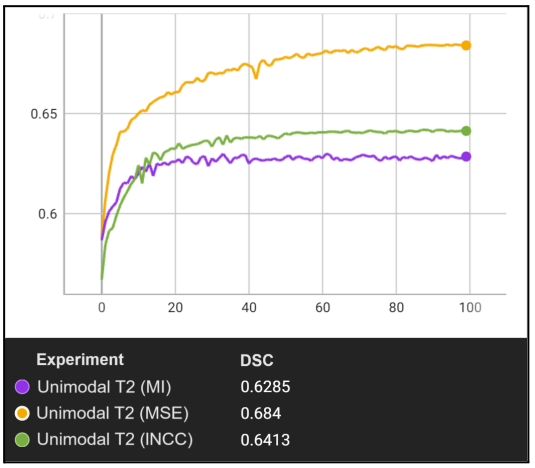
Para motivar la necesidad del método propuesto, se han realizado diferentes experimentos de registro unimodal y multimodal variando la modalidad de imagen y las métricas de similitud. En primer lugar, se han entrenado seis modelos para registro unimodal de T1 con T1 y T2 con T2. Para cada modalidad de MRI se han entrenado tres modelos con diferentes funciones de pérdida para establecer qué métrica de similitud de imagen proporciona los mejores resultados en cada caso. Las funciones de pérdida utilizadas han sido: MSE, lNCC y MI. En segundo lugar, se han entrenado seis modelos adicionales para registro multimodal de T1 con T2 y viceversa.

Durante el entrenamiento de estos modelos se ha realizado una validación de la calidad del registro para poder medir cuantitativamente cómo evoluciona el rendimiento de cada modelo a lo largo de las iteraciones. La validación se ha realizado calculando el DSC al terminar cada época con el conjunto de datos de validación que la red nunca ve durante el entrenamiento. Se han validado por un lado los modelos de registro unimodal (T1 con T1, y T2 con T2), y por el otro lado los modelos de registro multimodal (T2 con T1, y T1 con T2).

En la Figura 3.2 se muestra el DSC durante el entrenamiento para los modelos de registro unimodal con T1 (Figura 3.2a) y con T2 (Figura 3.2b). En la Figura 3.3 se muestra el correspondiente a los modelos de registro multimodal tanto de T1 a T2 (Figura 3.3a) como de T2 a T1 (Figura 3.3b). Se puede observar cómo los modelos unimodales aprenden a realizar un registro que mejora el DSC inicial, especialmente para las métricas lNCC y MI en el caso de T1 y para la métrica MSE en el caso de T2. Por el contrario, los modelos multimodales no consiguen mejorar el DSC inicial. Es más, éste empeora notablemente para las métricas MSE y lNCC durante el aprendizaje. La métrica MI, específicamente propuesta para métodos tradicionales de registro multimodal muestra un estancamiento (*stagnation*) durante el aprendizaje [37]. Con estos experimentos se evidencia la necesidad de abordar el problema con una aproximación necesariamente distinta, como la planteada en este trabajo.

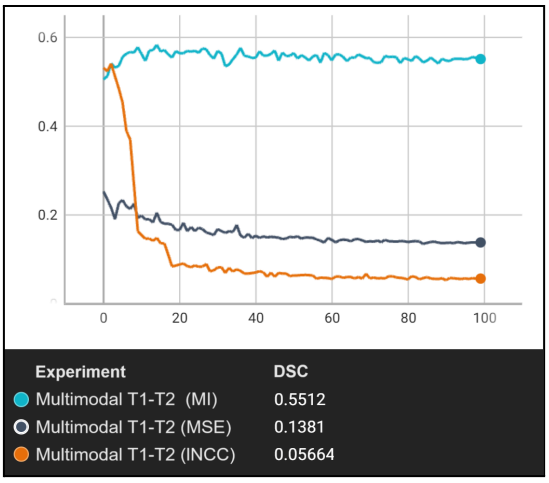


(a) Modelos de registro unimodal con T1.

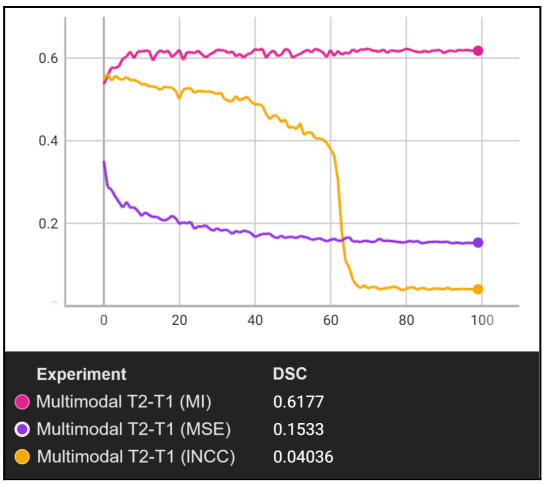


(b) Modelos de registro unimodal con T2.

Figura 3.2: Entrenamiento de los modelos de registro unimodal. Modelos entrenados con las métricas MI, MSE y lNCC durante 100 épocas. Representación del DSC a lo largo del entrenamiento. (a) Modelos de registro unimodal con T1, (b) Modelos de registro unimodal con T2.



(a) Modelos de registro multimodal de T1 a T2.



(b) Modelos de registro multimodal de T2 a T1.

Figura 3.3: Entrenamiento de los modelos de registro multimodal. Modelos entrenados con las métricas MI, MSE y lNCC durante 100 épocas. Representación del DSC a lo largo del entrenamiento. (a) Modelos de registro multimodal con T1, (b) Modelos de registro multimodal con T2.

3.4. Resultados de registro multimodal. CycleGAN + VoxelMorph.

Para evaluar los resultados obtenidos con el método propuesto para registro multimodal, se ha realizado un análisis cuantitativo y otro cualitativo. Dichos análisis se han realizado comparando los resultados obtenidos con nuestra solución y los obtenidos con un método tradicional multimodal (Stationary LDDMM [7, 8]) y con un método del estado del arte basado en deep learning (SynthMorph [33]) que han servido de baseline para nuestro estudio.

El análisis cuantitativo de los resultados de registro se ha llevado a cabo a través del cálculo de las métricas presentadas en 3.1.2 y 3.1.3 para todos los métodos de registro considerados. A partir de estas métricas, se puede obtener un análisis objetivo del rendimiento de los distintos métodos de registro. En la Tabla 3 se pueden observar los resultados obtenidos. El método propuesto obtiene mejor DSC que el método tradicional, pero a costa de peores jacobianos. El método es capaz de aprender que, para minimizar la función de pérdida, puede permitirse jacobianos muy grandes y muy pequeños en la transformación, llegando incluso a jacobianos negativos. Como consecuencia, el método proporciona soluciones que incrementan la similitud entre las imágenes mediante transformaciones localmente no difeomorfas. El método tradicional obtiene un peor DSC, ya que sacrifica similitud entre las imágenes para que el campo de deformación calculado sí sea difeomorfo.

Cabe destacar que SynthMorph es el método que mejores resultados obtiene, además de que el registro que realiza es difeomorfo. Sin embargo, este rendimiento es debido a la utilización de información durante el aprendizaje que le favorece. Por un lado, SynthMorph es entrenado con segmentaciones en lugar de imágenes MRI, por lo que el problema a resolver es significativamente más sencillo. Además, obtiene mucha información sobre cómo resolver el problema en la fase de entrenamiento, ya que las

segmentaciones que utiliza se obtienen con un algoritmo entrenado con esas mismas segmentaciones (SynthSeg). Por otro lado, la función de pérdida utilizada está basada en la métrica de evaluación (DSC), de forma que se está entrenando para optimizar esa métrica. Finalmente, una limitación importante de SnythMorph es que si SynthSeg no puede calcular las segmentaciones de acuerdo con las especificaciones del problema clínico, entonces no se puede realizar el registro, ya que depende de las segmentaciones calculadas por SynthSeg.

Métodos	**DSC**	**Jacobianos**		
		Min	*Max*	*# <0*
Stationary LDDMM (lNCC)	0.5703 (0.0797)	0.617 (0.103)	1.943 (0.401)	**0 (0)**
Stationary LDDMM (MI)	0.4474 (0.0834)	0.076 (0.440)	11.360 (4.516)	2 (9)
SynthMorph	**0.744 (0.049)**	0.071 (0.075)	4.634 (1.833)	**0 (0)**
CycleGAN + VoxelMorph (lNCC)	**0.686 (0.078)**	-39.251 (30.529)	105.857 (61.558)	79031 (34983)
CycleGAN + VoxelMorph (MI)	0.669 (0.073)	-26.187 (15.983)	84.540 (38.568)	80637 (34941)
CycleGAN + VoxelMorph (MSE)	0.663 (0.087)	-5.981 (3.740)	38.380 (25.404)	**23308 (14324)**

Tabla 3: Resultados obtenidos con los distintos métodos de registro. Separados por una línea negrita, se muestran los resultados de los métodos de referencia y los del método propuesto. Los resultados obtenidos representan la media y desviación estándar (en paréntesis) de las distintas métricas para todo el conjunto de datos de test. Se resaltan en negrita los mejores resultados obtenidos.

El análisis cualitativo de los resultados de registro se ha llevado a cabo a través de la visualización y comparación de los volúmenes registrados con los volúmenes de referencia, así como la visualización de los campos de deformación calculados con los métodos considerados. En la Figura 3.4 pueden observarse cortes axiales de un resultado de registro seleccionado. En la columna *warped* se han marcado en rojo algunas similitudes visibles a simple vista con el volumen *fixed*. En la Figura 3.5 pueden observarse las diferencias entre ambas modalidades del resultado de registro de la Figura 3.4, tanto el volumen real T2 como el volumen equivalente inferido por CycleGAN.

Cabe destacar que el método propuesto consigue una mayor similitud visual al realizar el registro con la imagen de referencia. Esto es debido, tal y como se ha mencionado en el análisis cuantitativo, a que el campo de deformación calculado por el método propuesto no cumple con la restricción de ser difeomorfo. El método tradicional obtiene una menor similitud entre imágenes que el método propuesto, debido a que su campo de deformación calculado sí es difeomorfo. Se aprecia también que el método de SynthMorph, al ser un método de registro invariante de contraste, funciona correctamente independientemente de la modalidad de los volúmenes a registrar.

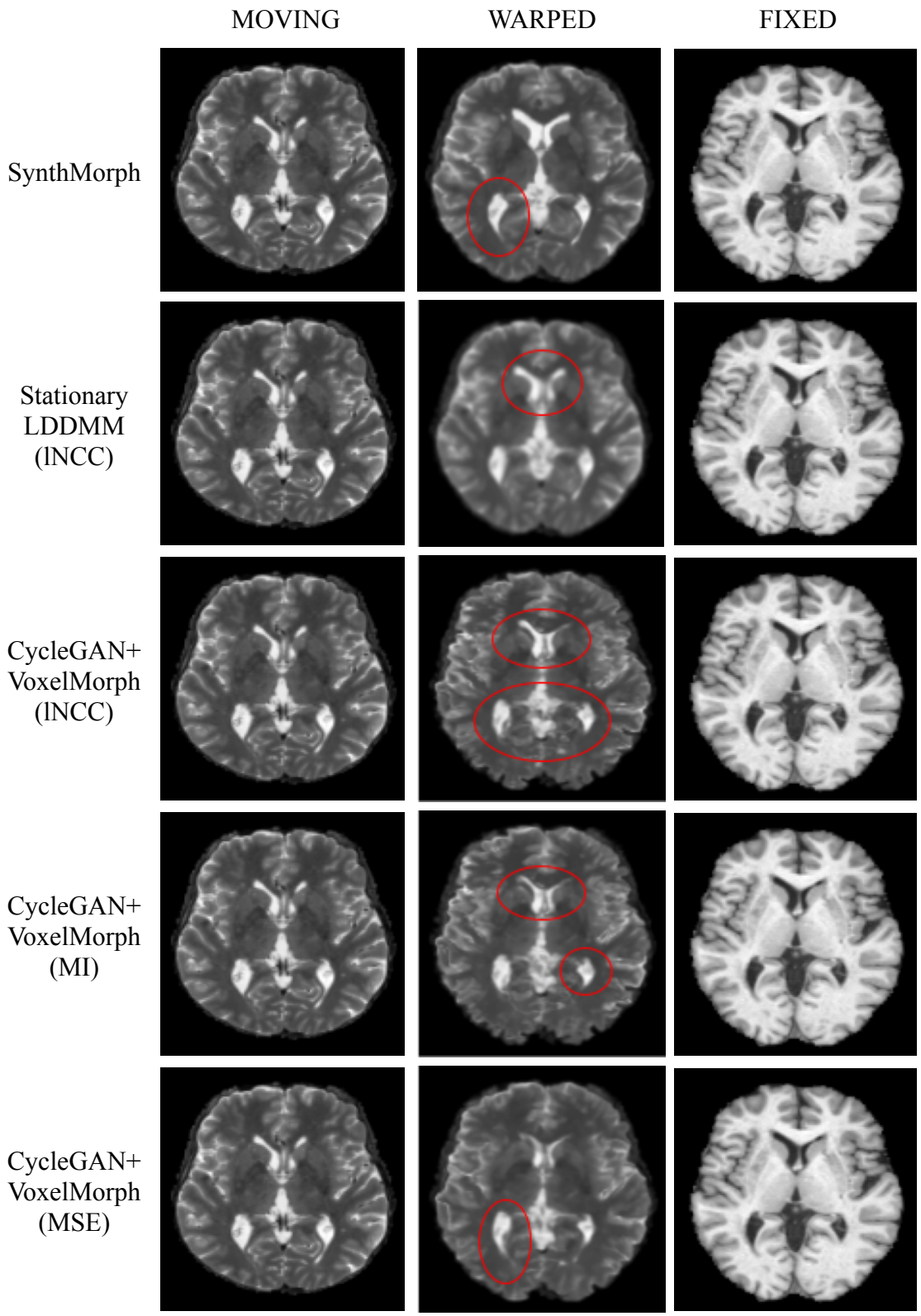


Figura 3.4: Vistas axiales de un resultado de registro seleccionado para todos los métodos evaluados. Para cada método se muestra la vista axial de *moving* (el volumen original a registrar), *fixed* (el volumen objetivo) y *warped* (el volumen *moving* tras aplicar el campo de deformación calculado). En la columna WARPED se han marcado en rojo algunas similitudes visibles a simple vista con el volumen *fixed*.

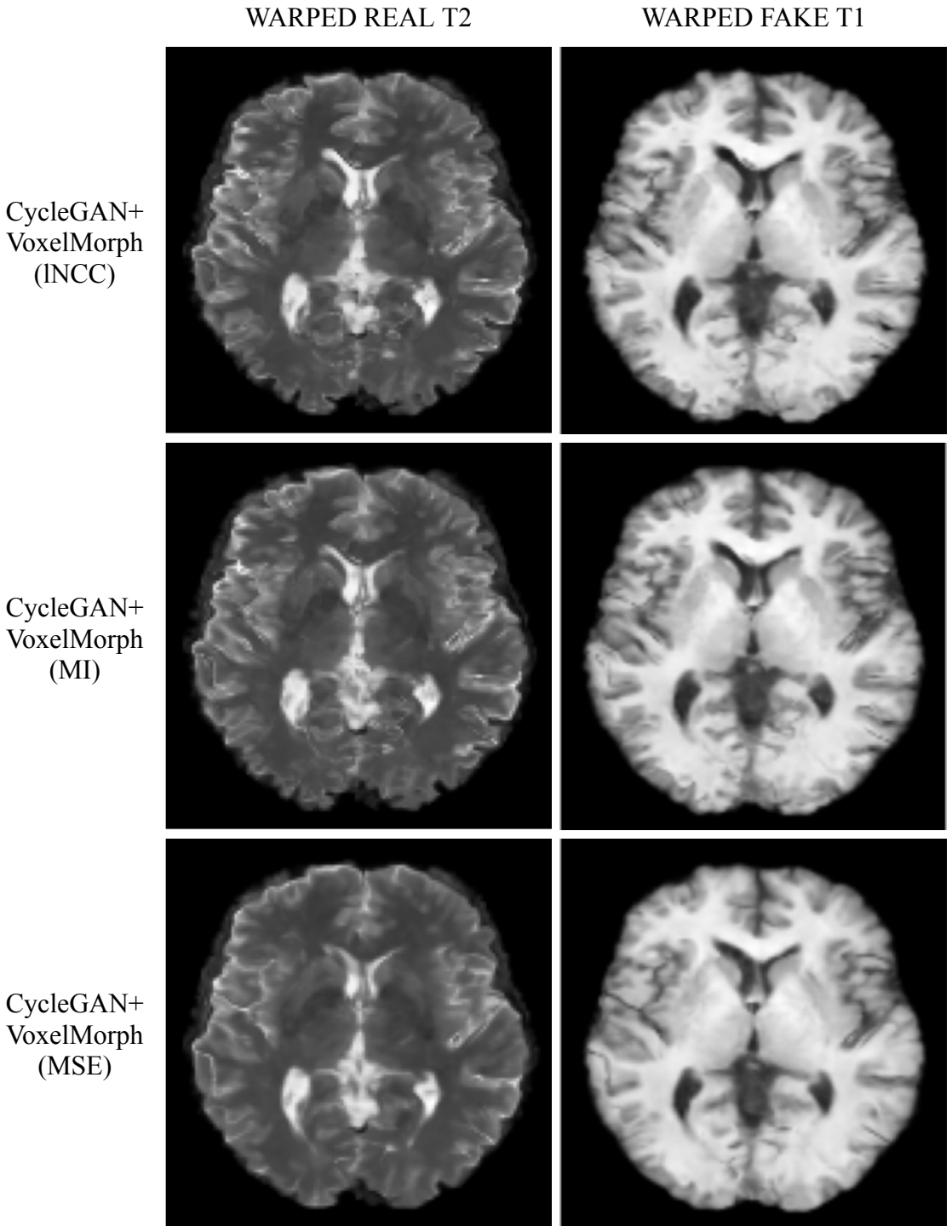


Figura 3.5: Comparativa resultado tras aplicar el campo de deformación al volúmen real y al mismo inferido por CycleGAN. Por un lado, se muestra la vista axial del volumen real T2 registrado (*warped real T2*). Por otro lado, la vista axial del volumen registrado tras aplicar el campo de deformación calculado (*warped real T2*) por CycleGAN (*warped fake T1*).

En la Figura 3.6 se observa un corte axial de los campos de deformación calculados por los distintos métodos de registro a partir del registro realizado con los volúmenes de la Figura 3.4. Esta visualización permite apreciar las diferencias entre el método implementado en este trabajo y los métodos de baseline. Los campos de deformación calculados por los métodos de referencia son difeomorfos, mientras que los del método propuesto no. Esta propiedad es muy deseable en la mayoría de las aplicaciones clínicas pues el modelo de deformación real debe de garantizar que se preserva la topología de las estructuras cerebrales. Esta es, sin duda, una debilidad del método propuesto que se solucionará como trabajo futuro.

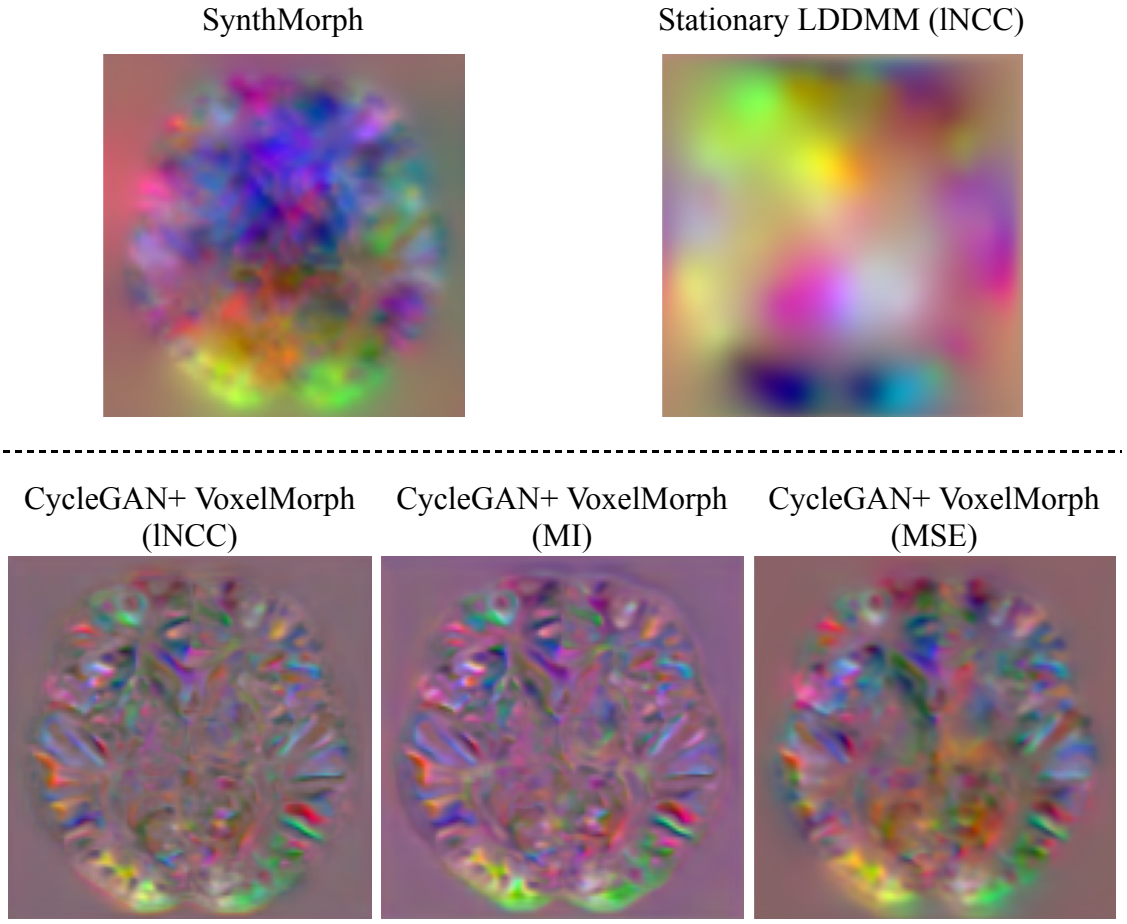


Figura 3.6: Vistas axiales de los campos de deformación calculados por los distintos métodos de registro. Cada vector (*ux, uy, uz*) del campo de deformación es mapeado a un color RGB.

4. Conclusiones

En este trabajo se ha implementado una solución novedosa y prometedora basada en métodos de aprendizaje profundo para resolver el problema del registro no-rígido multimodal de MRI cerebrales 3D, problema para el que actualmente hay pocas soluciones propuestas. Nuestro método hace uso de una combinación de los modelos de CycleGAN y VoxelMorph. CycleGAN realiza un cambio de modalidad de las MRI, transformando las MRI ponderadas en T2 a su equivalente en T1. Por su parte, VoxelMorph se encarga de realizar el registro de los volúmenes. Los conjuntos de datos biomédicos multimodales disponibles públicamente para la evaluación de métodos de registro son muy escasos. Es por ello que en este trabajo, se ha generado un conjunto de datos de MRI cerebrales ponderadas en T1 y en T2. Este conjunto de datos ha sido preprocesado y revisado para poder ser utilizado para la evaluación del registro multimodal de imágenes médicas en otros trabajos siguiendo la idea del *Learn2Reg* challenge (https://learn2reg.grand-challenge.org/).

Los experimentos realizados en este trabajo justifican el uso del método propuesto para abordar el problema de registro no-rígido multimodal. Por un lado, la implementación en 3D realizada de CycleGAN permite realizar un cambio de modalidad de imagen de T2 a T1 con una buena precisión, reduciendo así la complejidad del problema de registro al posibilitar la utilización de métodos uni-modales. La combinación de CycleGAN y VoxelMorph ha resultado ser prometedora como solución al registro no-rígido multimodal.

La evaluación cuantitativa ha mostrado un mejor DSC que con un método tradicional (Stationary LDDMM) aunque por debajo del método de deep-learning elegido como referencia (SynthMorph). El campo de deformación calculado por nuestro método no es difeomorfo, lo que en la práctica clínica no es deseable ya que no garantiza la conservación de la topología anatómica y, en este sentido, SynthMorph sería preferible. No obstante, los excelentes resultados de SynthMorph vienen condicionados por ciertas prácticas realizadas durante su entrenamiento en la que visualiza información del proceso de evaluación, que hacen cuestionable su utilización como referencia o *baseline*.

Como principal línea trabajo futuro, se podría extender el método propuesto para utilizar métodos de deformación y regularizadores que garanticen que el campo de deformación calculado sea difeomorfo y realizar una evaluación exhaustiva de su rendimiento teniendo en cuenta otros métodos del estado del arte. Adicionalmente, se podría estudiar el comportamiento del método propuesto bajo un cambio de dominio que demuestre la validez de la idea no sólo en imágenes MRI cerebrales ponderadas en T1 y T2. También se podría considerar el utilizar una GAN más adecuada para el cambio de modalidad. Esto se debe a que la "consistencia del ciclo" presente en CycleGAN conduce a múltiples soluciones, lo que significa que las imágenes traducidas pueden no mantener la estructura anatómica de las imágenes de origen y pueden contener ciertas imprecisiones [39]. Finalmente, una vez evaluada la validez de la metodología se podría estudiar su usabilidad en aplicaciones clínicas como el diagnóstico asistido por computador mediante sistemas de deep-learning [15, 40].

Como conclusión final, la solución propuesta tiene la ventaja de que, gracias al uso de la arquitectura de CycleGAN, el problema de registro multimodal se convierte en un problema unimodal que puede ser abordado por una gran cantidad de métodos de

registro. Además, el proceso completo para realizar el registro es significativamente más rápido que los métodos tradicionales, lo que lo convierte en una opción preferible en el procesado de conjuntos grandes de datos.

Bibliografía

[1] D. Abramian and A. Eklund, ‘Generating fMRI volumes from T1-weighted volumes using 3D CycleGAN’, arXiv [eess.IV]. 2019.

[2] Z. Chen, J. Wei, and R. Li, ‘Unsupervised Multi-Modal Medical Image Registration via Discriminator-Free Image-to-Image Translation’, ArXiv, vol. abs/2204.13656, 2022.

[3] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, ‘TransMorph: Transformer for unsupervised medical image registration’, Medical Image Analysis, vol. 82, p. 102615, 2022.

[4] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, ‘Deep learning in medical image registration: a review’, Physics in Medicine & Biology, vol. 65, no. 20, p. 20TR01, Oct. 2020.

[5] L. Deng, Q. Lan, Q. Zhi, S. Huang, J. Wang, and X. Yang, ‘Deep learning-based 3D brain multimodal medical image registration’, Medical & Biological Engineering & Computing, Nov. 2023.

[6] Y. He, A. Wang, S. Li, Y. Yang, and A. Hao, ‘Nonfinite-modality data augmentation for brain image registration’, Computers in Biology and Medicine, vol. 147, p. 105780, 2022.

[7] M. Hernandez, M. N. Bossa, and S. Olmos, ‘Registration of Anatomical Images Using Paths of Diffeomorphisms Parameterized with Stationary Vector Field Flows’, International Journal of Computer Vision, vol. 85, no. 3, pp. 291–306, Dec. 2009.

[8] M. Hernandez, U. Ramon-Julvez, and D. Sierra-Tome, ‘Partial Differential Equation-Constrained Diffeomorphic Registration from Sum of Squared Differences to Normalized Cross-Correlation, Normalized Gradient Fields, and Mutual Information: A Unifying Framework’, Sensors, vol. 22, 5 2022.

[9] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, ‘Medical image registration’, Physics in Medicine & Biology, vol. 46, no. 3, p. R1-45, Mar. 2001.

[10] B. Kim, D. H. Kim, S. H. Park, J. Kim, J.-G. Lee, and J. C. Ye, ‘CycleMorph: Cycle consistent unsupervised deformable image registration’, Medical Image Analysis, vol. 71, p. 102036, 2021.

[11] P. J. LaMontagne et al., ‘OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease’, medRxiv, 2019.

[12] H. R. Boveiri, R. Khayami, R. Javidan, and A. Mehdizadeh, ‘Medical image registration using deep neural networks: A comprehensive review’, Computers & Electrical Engineering, vol. 87, p. 106767, 2020.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, ‘CBAM: Convolutional Block Attention Module’, in Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII, Munich, Germany, 2018, pp. 3–19.

[14] A. Q. Wang, E. M. Yu, A. V. Dalca, and M. R. Sabuncu, ‘A robust and interpretable deep learning framework for multi-modal registration via keypoints’, Medical Image Analysis, vol. 90, p. 102962, 2023.

[15] U. Ramon-Julvez, M. Hernandez, E. Mayordomo, and ADNI, ‘Analysis of the Influence of Diffeomorphic Normalization in the Prediction of Stable VS Progressive MCI Conversion with Convolutional Neural Networks’, in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1120–1124.

[16] Y. Zheng et al., ‘SymReg-GAN: Symmetric Image Registration with Generative Adversarial Networks’, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, pp. 5631–5646, 9 2022.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, ‘Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks’, in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251.

[18] B. Zitová and J. Flusser, ‘Image registration methods: a survey’, Image and Vision Computing, vol. 21, no. 11, pp. 977–1000, 2003.

[19] J. Lu, J. Öfverstedt, J. Lindblad, and N. Sladoje, ‘Is image-to-image translation the panacea for multimodal image registration? A comparative study’, PLoS One, vol. 17, no. 11, p. e0276196, Nov. 2022.

[20] B. Fischl, ‘FreeSurfer’, NeuroImage, vol. 62, no. 2, pp. 774–781, 2012.

[21] “recon-all - Free Surfer Wiki.” https://surfer.nmr.mgh.harvard.edu/fswiki/recon-all (accessed Nov. 04, 2023).

[22] “NITRC: Robust Brain Extraction (ROBEX): Tool/Resource Info,” www.nitrc.org. https://www.nitrc.org/projects/robex (accessed Nov. 04, 2023).

[23] “ICBM 152 extended nonlinear atlases (2020) – NIST.” https://nist.mni.mcgill.ca/icbm-152-extended-nonlinear-atlases-2020/ (accessed Nov. 04, 2023).

[24] “Advanced Normalization Tools,” GitHub, Nov. 03, 2023. https://github.com/ANTsX/ANTs (accessed Nov. 04, 2023).

[25] I. Goodfellow et al., ‘Generative Adversarial Networks’, Advances in Neural Information Processing Systems, vol. 3, 06 2014.

[26] G. Balakrishnan, A. Zhao, M. Sabuncu, J. Guttag, and A. V. Dalca, ‘VoxelMorph: A Learning Framework for Deformable Medical Image Registration’, IEEE TMI: Transactions on Medical Imaging, vol. 38, pp. 1788–1800, 2019.

[27] junyuchen245, “GitHub - junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration: TransMorph: Transformer for Unsupervised Medical Image Registration (PyTorch),” GitHub. https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration (accessed Nov. 04, 2023).

[28] B. Billot et al., ‘SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining’, Medical Image Analysis, vol. 86, p. 102789, 2023.

[29] D. C. Dowson and B. V. Landau, ‘The Fréchet distance between multivariate normal distributions’, Journal of Multivariate Analysis, vol. 12, no. 3, pp. 450–455, 1982.

[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, ‘GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium’, in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017, pp. 6629–6640.

[31] D. Kingma and J. Ba, ‘Adam: A Method for Stochastic Optimization’, International Conference on Learning Representations, 12 2014.

[32] G. Balakrishnan, A. Zhao, M. Sabuncu, J. Guttag, and A. V. Dalca, ‘An Unsupervised Learning Model for Deformable Medical Image Registration’, CVPR: Computer Vision and Pattern Recognition, pp. 9252–9260, 2018.

[33] M. Hoffmann, B. Billot, D. N. Greve, J. E. Iglesias, B. Fischl, and A. V. Dalca, ‘SynthMorph: Learning Contrast-Invariant Registration Without Acquired Images’, IEEE Transactions on Medical Imaging, vol. 41, no. 3, pp. 543–558, 2022.

[34] O. Ronneberger, P. Fischer, and T. Brox, ‘U-Net: Convolutional Networks for Biomedical Image Segmentation’, in Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, 2015, pp. 234–241.

[35] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, ‘Spatial Transformer Networks’, Advances in Neural Information Processing Systems 28 (NIPS 2015), 06 2015.

[36] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, ‘Mutual-information-based registration of medical images: a survey’, IEEE Transactions on Medical Imaging, vol. 22, no. 8, pp. 986–1004, 2003.

[37] J. Modersitzki, Fair: Flexible Algorithms for Image Registration. USA: Society for Industrial and Applied Mathematics, 2009.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, ‘Image-to-Image Translation with Conditional Adversarial Networks’, CVPR, 2017.

[39] L. Kong, C. Lian, D. Huang, Z. Li, Y. Hu, and Q. Zhou, ‘Breaking the Dilemma of Medical Image-to-image Translation’, in Neural Information Processing Systems, 2021.

[40] S. Spasov, L. Passamonti, A. Duggento, P. Liò, N. Toschi, and Alzheimer’s Disease Neuroimaging Initiative, ‘A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer’s disease’, Neuroimage, vol. 189, pp. 276–287, Jan. 2019.