UNIVERSIDAD NACIONAL DEL LITORAL

DOCTORADO EN INGENIERÍA

Aprendizaje profundo aplicado al análisis de imágenes médicas

Agostina Juliana Larrazabal

FICH
FACULTAD DE INGENIERÍA
Y CIENCIAS HÍDRICAS

INTEC
INSTITUTO DE DESARROLLO TECNOLÓGICO
PARA LA INDUSTRIA QUÍMICA

CIMEC
CENTRO DE INVESTIGACIÓN DE
MÉTODOS COMPUTACIONALES

sinc(i)
INSTITUTO DE INVESTIGACIÓN EN SEÑALES
SISTEMAS E INTELIGENCIA COMPUTACIONAL

Tesis de Doctorado **2021**

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química
Centro de Investigación de Métodos Computacionales
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

APRENDIZAJE PROFUNDO APLICADO AL ANÁLISIS DE IMÁGENES MÉDICAS

Agostina Juliana Larrazabal

Tesis remitida al Comité Académico del Doctorado
como parte de los requisitos para la obtención
del grado de
DOCTOR EN INGENIERÍA
Mención Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2021

Comisión de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria, Paraje "El Pozo", S3000, Santa Fe, Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

APRENDIZAJE PROFUNDO APLICADO AL ANÁLISIS DE IMÁGENES MÉDICAS

Agostina Juliana Larrazabal

Lugar de Trabajo:
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional– sinc(i), FICH-UNL/CONICET.

Director:
Dr. César Ernesto Martínez — sinc(i)-CONICET-UNL

Co-directora:
Dra. Cecilia Elizabet García Cena — CAR-UPM

Jurado Evaluador:
Dr. José Ignacio Orlando — PLADEMA-CONICET-UNICEN
Dr. Claudio Delrieux — CONICET-UNS
Dr. Leandro A. Bugnon — sinc(i)-CONICET-UNL

2021

DECLARACIÓN LEGAL DEL AUTOR

Esta tesis ha sido remitida como parte de los requisitos para la obtención del grado académico de Doctor en Ingeniería - Mención Inteligencia Computacional, Señales y Sistemas ante la Universidad Nacional del Litoral y ha sido depositada en la Biblioteca de la Facultad de Ingeniería y Ciencias Hídricas para que esté a disposición de sus lectores bajo las condiciones estipuladas por el Reglamento de la mencionada Biblioteca.

Citaciones breves de esta tesis son permitidas sin la necesidad de un permiso especial, en la suposición de que la fuente sea correctamente citada. Solicitudes de permiso para la citación extendida o para la reproducción parcial o total de este manuscrito serán concebidos por el portador legal del derecho de propiedad intelectual de la obra.

TESIS POR COMPILACIÓN

La presente tesis se encuentra organizada bajo el formato de Tesis por Compilación, aprobado en la resolución Nº 255/17 (Expte. Nº 888317-17) por el Comité Académico de la Carrera Doctorado en Ingeniería, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral (UNL). De dicha resolución:

"*En el caso de optar por la Tesis por Compilación, ésta consistirá en una descripción técnica de al menos 30 páginas, redactada en español e incluyendo todas las investigaciones abordadas en la tesis. Se deberán incluir las secciones habituales indicadas a continuación en la Sección Contenidos de la Tesis. Los artículos científicos publicados por el autor, en el idioma original de las publicaciones, deberán incluirse en un Anexo con el formato unificado al estilo general de la Tesis indicado en la Sección Formato. El Anexo deberá estar encabezado por una sección donde el tesista detalle para cada una de las publicaciones cuál ha sido su contribución. Esta sección deberá estar avalada por su director de Tesis. El documento central de la Tesis debe incluir referencias explícitas a todas las publicaciones anexadas y presentar una conclusión que muestre la coherencia de dichos trabajos con el hilo conceptual y metodológico de la tesis. Los artículos presentados en los anexos podrán ser artículos publicados, aceptados para publicación (en prensa) o en revisión.*"

Agradecimientos

A mis directores, Dr. César Martínez y Dra. Cecilia García Cena, por darme la confianza y la oportunidad de formar parte de un grupo de investigación. Por apoyarme y guiarme con libertad. Al Dr. Enzo Ferrante, por ser una guía y un pilar fundamental para el desarrollo de esta tesis. Por permitirme recurrir a su capacidad y experiencia científica en un marco de confianza y amistad. A mis compañeros y amigos del sinc(i), tanto por el conocimiento técnico, como por todos los mementos, mates, y charlas compartidas en estos años. A mi novio, Emi Bressan, por transitar conmigo este camino y ser mi cable a tierra. A mi familia por la compañía y apoyo incondicional.
Finalmente, quiero agradecer a las siguientes instituciones:

- sinc(i): Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional.
- Facultad de Ciencias Hídricas de la Universidad Nacional del Litoral.
- CONICET: Consejo Nacional de Investigaciones Científicas y Técnicas.

Agostina J. Larrazabal
Santa Fe, Noviembre de 2021.

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 17 de Marzo de 2022.

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Aprendizaje profundo aplicado al análisis de imágenes médicas"***, desarrollada por la Bioing. Agostina Juliana LARRAZABAL, en el marco de la Mención "Inteligencia computacional, señales y sistemas", certificamos que hemos evaluado la Tesis y recomendamos que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

La aprobación final de esta disertación estará condicionada a la presentación de dos copias encuadernadas de la versión final de la Tesis ante el Comité Académico del Doctorado en Ingeniería.

Dr. José Ignacio Orlando	Dr. Claudio Delrieux	Dr. Leandro Bugnon

Santa Fe, 17 de Marzo de 2022.

Certifico haber leído la Tesis, preparada bajo mi dirección en el marco de la Mención "Inteligencia computacional, señales y sistemas" y recomiendo que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

.....	.....
Dra. Cecilia García Cena	Dr. César Martínez
Codirectora de Tesis	Director de Tesis

Dr. JOSÉ LUIS MACOR
SECRETARIO DE POSGRADO
Facultad de Ingeniería y Cs. Hídricas

Universidad Nacional del Litoral
Facultad de Ingeniería y
Ciencias Hídricas

Secretaría de Posgrado

Ciudad Universitaria
C.C. 217
Ruta Nacional N° 168 – Km. 472,4
(3000) Santa Fe
Tel: (54) (0342) 4575 229
Fax: (54) (0342) 4575 224
E-mail: posgrado@fich.unl.edu.ar

Índice general

Índice de figuras

Índice de tablas

Resumen

El análisis de imágenes médicas es fundamental para la medicina moderna. En los últimos años, el aprendizaje profundo y particularmente las redes neuronales convolucionales (CNN, del inglés *Convolutional Neural Networks*) han logrado grandes mejoras en el desempeño de las técnicas de visión computacional, permitiendo automatizar tareas complejas. En esta tesis se avanzará en el desarrollo de técnicas y metodologías basadas en redes profundas, focalizando en tres aplicaciones en particular: la segmentación de estructuras en imágenes médicas, la estimación de incertidumbre y el seguimiento de mirada en video-oculografía para detección de enfermedades neurológicas.

1. A pesar de su sorprendente habilidad para segmentar imágenes médicas con gran precisión, algunas de las arquitecturas más populares de CNNs todavía confían en estrategias de postprocesamiento (como los campos aleatorios condicionales) para incorporar restricciones de conectividad a las máscaras resultantes. En esta tesis, se presenta Post-DAE, un método de post-procesamiento basado en autocodificadores que incorpora plausibilidad anatómica a las segmentaciones generadas por distintos algoritmos. Post-DAE aprende un espacio bajo-dimensional anatómicamente plausible, y lo usa para imponer restricciones de forma durante el post-procesamiento de máscaras anatómicas obtenidas con métodos arbitrarios. Además, se entrena únicamente con máscaras anatómicas por lo que es independiente de la modalidad y la intensidad de la imagen, logrando un enfoque flexible.

2. Las CNNs han demostrado estar *mal calibradas* y poseer un exceso de confianza en sus predicciones, incluso ante salidas incorrectas, convirtiéndose en modelos poco confiables. Para superar esta limitación, en este trabajo se analizan dos enfoques. El aprendizaje de ensambles ha demostrado no sólo potenciar el desempeño de los modelos individuales, sino también incrementar su calibración. En este escenario, la diversidad de los modelos es un factor clave que facilita a que los mismos convergan a soluciones funcionales diferentes. En este trabajo se propone un nuevo método de entrenamiento de ensambles que fomenta la diversidad de sus modelos constituyentes mediante la incorporación de restricciones de ortogonalidad. Con la incorporación de dos términos de penalización basados en la similitud coseno, se regulariza el entrenamiento secuencial de los ensambles, obteniendo modelos mejor calibrados y con un mejor desempeño predictivo.

Como un enfoque alternativo, se propone el método de entropía máxima para predicciones erróneas (MEEP), una estrategia de entrenamiento que penaliza selectivamente la sobre-confianza de las predicciones erróneas. Para esto, se incorpora un término de regularización que promueve el incremento de entropía del conjunto de píxeles mal clasificados. Los resultados experimentales demuestran que asociando

MEEP con las funciones de pérdida más conocidas se obtienen mejoras significativas tanto en la calibración del modelo, como en la calidad de las segmentaciones.

3. Recientemente, el seguimiento de ojos mediante video-oculografía (VOG) ha comenzado a utilizarse como herramienta para el diagnóstico de una amplia variedad de enfermedades neurológicas y mentales. Para esta aplicación, se utilizan los denominados métodos basados en características, los cuales utilizan características extraídas de imágenes de alta resolución tomadas sobre la región ocular. La principal debilidad de estos métodos es que la cabeza del usuario debe permanecer estática para evitar errores de estimación. En algunos pacientes, los movimientos involuntarios no pueden evitarse, y es necesario medirlos. En esta tesis, se abordó la medición de la posición de la cabeza del usuario, como un primer paso para mejorar el seguimiento de ojos en aplicaciones de alta precisión. Se diseñó una CNN de regresión en cascada para estimar, en dos etapas, las coordenadas de la esquina del ojo. Por último, se adicionó información temporal para mejorar la precisión y disminuir el costo computacional. El desempeño se evaluó de forma cuantitativa y cualitativa, obteniendo mejoras significativas.

Abstract

Medical imaging is a fundamental component of the modern medicine and it has become essential to have automated systems that can analyze and manage the large amount of information available quickly and efficiently. In recent years, deep learning and in particular convolutional neural networks (CNN), have achieved great improvements in the performance of computer vision techniques, allowing the automation of complex tasks. This thesis will advance the development of techniques and methodologies based on deep networks, focusing on three applications in particular: medical imaging segmentation, uncertainty estimation and video-oculography gaze tracking for neurological disease detection.

1. CNNs proved to be highly accurate to perform anatomical segmentation of medical images. However, some of the most popular CNN architectures still rely on post-processing strategies (e.g. Conditional Random Fields) to incorporate connectivity constraints into the resulting masks. Post-DAE, a post-processing method based on denoising autoencoders, is introduced in this thesis to improve the anatomical plausibility of arbitrary biomedical image segmentation algorithms. Post-DAE learn a low-dimensional space of anatomically plausible segmentations, and use it to impose shape constraints by post-processing anatomical segmentation masks obtained with arbitrary methods. The proposed approach is independent of image modality and intensity information since it employs only segmentation masks for training, making the approach very flexible.

2. On the other hand, despite the astonishing performance of deep-learning based approaches in medical image segmentation tasks, it has recently been observed that they tend to produce overconfident estimates, even in situations of high uncertainty, leading to unreliable models. Two different approaches were proposed in this thesis to improve this limitation. Ensemble learning has shown to not only boost the performance of individual models but also reduce their miscalibration by averaging the independent predictions. In this scenario, model diversity has become a key factor, which facilitates individual models converging to different functional solutions. In this work, Orthogonal Ensemble Networks is designed to explicitly enforce model diversity by means of orthogonal constraints. A new pairwise orthogonality constraint was resorted to regularize a sequential ensemble training process, resulting on improved predictive performance and better calibrated model outputs.The experimental results show that the approach produces more robust and well-calibrated ensemble models and can deal with challenging tasks in the context of biomedical image segmentation.

As an alternative approach, Maximum Entropy on Erroneous Predictions (MEEP) was proposed. MEEP is a training strategy for segmentation networks which selectively penalizes overconfident predictions, focusing only on misclassified

pixels. In particular, It was designed a regularization term that encourages high entropy posteriors for wrong predictions, increasing the network uncertainty in complex scenarios. The experimental results demonstrate that coupling MEEP with standard segmentation losses leads to improvements not only in terms of model calibration, but also in segmentation quality.

3. Recently, video-oculographic gaze tracking has begun to be used in the diagnosis of a wide variety of neurological diseases, such as Parkinson and Alzheimer. For this application, the so-called feature-based methods are used. They use geometrically derived eye features from high-resolution eye images captured by zooming into the user's eyes. The main weakness of these methods is that the head of the user must remain motionless to avoid estimation errors. In some patients, some involuntary movements cannot be avoided and it is necessary to measure them. In this thesis, the measurement of head position was tackled as a way to improve the gaze tracking on these precision demanding medical applications. As a first stage, the eye corners coordinates were obtained as a reference point. The problem was handled as a regression problem using a coarse-to-fine cascaded convolutional neural network in order to accurately regress the coordinates of the eye corner. Finally, temporal information was added to increase accuracy and decrease computation time. The accuracy of the estimation was calculated and subjective performance was also evaluated through video inspection. In both cases, satisfactory results were obtained.

Capítulo 1

Introducción

1.1. Descripción del problema

Con el avance del procesamiento de imágenes y la visión computacional, el análisis de imágenes médicas ha cobrado una gran importancia para la medicina moderna, ya que permite diagnosticar, monitorear y tratar problemas médicos de manera temprana y no invasiva. Sin embargo, debido a la variedad de patologías y al constante aumento en el número de imágenes disponibles, se hace indispensable contar con sistemas automáticos que permitan analizar y manejar esta gran cantidad de información de una manera rápida y eficiente. Es por esto, que investigadores y médicos han comenzado a beneficiarse de los sistemas asistidos por computadora.

Desde que las imágenes pudieron descargarse en una computadora, se ha investigado la forma de poder analizarlas automáticamente aplicando la tecnología disponible. Los primeros sistemas utilizaban métodos básicos de procesamiento de imágenes, como filtros para detección de bordes, o modelos matemáticos basados en regiones [1]. A principios de los años '90, los enfoques de aprendizaje de máquina basados en extracción de características se convirtieron en la aproximación mas utilizada por un largo período. Sin embargo, la efectividad de estos métodos dependía en gran medida de expertos en el tema que debían definir, diseñar y extraer de las imágenes las características significativas relacionadas con la tarea que se deseaba realizar. La complejidad de estos enfoques ha sido considerada como una limitación para su desarrollo, ya que imposibilitaba que personas no expertas puedan explorar técnicas de aprendizaje computacional para sus propios estudios.

En los últimos años, gracias a la sustancial mejora en el hardware disponible, el aprendizaje profundo (DL, del inglés *Deep Learning*) ha atenuado estos obstáculos, incluyendo la ingeniería de extracción de características dentro del procedimiento completo de aprendizaje del modelo. Con este enfoque, en lugar de extraer las características de forma manual, los modelos descubren las representaciones relevantes para cada problema, aprendiéndolas desde los datos [2].

Particularmente, las redes neuronales convolucionales (CNN, del inglés *Convolutional Neural Network*) son un tipo de red neuronal que ha logrado desempeños muy altos en una gran variedad de aplicaciones de la inteligencia artificial [3]. Básicamente, una CNN puede pensarse como un tipo de red neuronal que utiliza muchas copias idénticas de la misma neurona. Este esquema se denomina de parámetros compartidos, y permite a la red generar modelos computacionales muy grandes manteniendo el número de parámetros que deben aprenderse relativamente bajo. Además, la es-

tructura profunda de estas redes permite extraer características discriminativas bajo múltiples niveles de abstracción [4], aprendiendo con cada capa representaciones más abstractas.

En el campo de las imágenes médicas, la aparición de estas redes ha conseguido grandes mejoras en el desempeño de las técnicas de visión computacional, permitiendo automatizar tareas complejas como segmentación de imágenes [5], registración [6, 7], anotación de imágenes [8], creación de sistemas de apoyo diagnóstico guiado por computadora [9, 10], detección de lesiones y puntos de referencia, entre muchos otros [11].

En esta tesis se avanzará en el desarrollo de técnicas y metodologías basadas en redes profundas, focalizando en tres aplicaciones en particular: la segmentación de estructuras en imágenes médicas, la estimación de incertidumbre y el seguimiento de mirada en video-oculografía para detección de enfermedades neurológicas.

1.1.1. Segmentación de imágenes médicas

La segmentación de imágenes médicas consiste en identificar los píxeles (o vóxeles) pertenecientes a órganos, estructuras anatómicas o lesiones a partir de una imagen médica. Por ejemplo, una tomografía computada (TC), una resonancia magnética (RM), o una radiografía como en el caso de la Fig. 1.1.

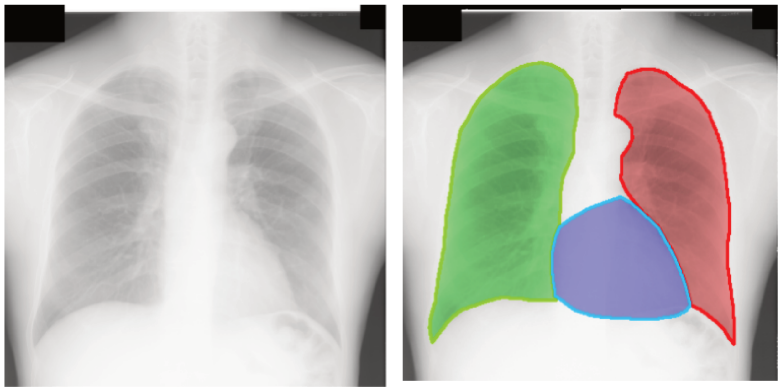


Figura 1.1: Segmentación de pulmones y corazón en una radiografía de tórax.

Esta tarea constituye un paso muy importante en distintos procedimientos médicos como la localización de patologías, el estudio de estructuras anatómicas, el diagnóstico asistido por computadora o la planificación de radioterapia. Por lo tanto, obtener segmentaciones precisas y anatómicamente plausibles resulta imprescindible e influye directamente en la calidad general del sistema completo.

Las CNNs han demostrado ser capaces de segmentar imágenes médicas con gran precisión [12, 13, 14]. Sin embargo, a pesar de que las estructuras anatómicas suelen presentar una gran regularidad en su posición, forma y topología, las predicciones a nivel de píxel de la mayoría de las arquitecturas de redes convolucionales no están diseñadas para tener en cuenta estas propiedades [15]. Por lo tanto, si no se incorpora información del contexto se puede llegar a predicciones erróneas en áreas con intensidades similares. Surge entonces la necesidad de desarrollar nuevos métodos que incorporen antecedentes anatómicos de las estructuras al proceso de segmentación, y de esa forma obtener segmentaciones anatómicamente plausibles y con mayor precisión.

1.1.2. Estimación de incertidumbre

A pesar del éxito y su sorprendente habilidad para aprender características altamente discriminativas, las CNNs han demostrado estar *mal calibradas* y poseer un exceso de confianza en sus predicciones, incluso ante salidas incorrectas [16]. En otras palabras, se ha encontrado una baja correlación entre la confianza asociada a sus predicciones y la precisión real de las mismas [17]. Esto resulta en un problema importante, que puede tener consecuencias catastróficas en sistemas de toma de decisiones como el diagnóstico médico, que dependen de las probabilidades de las predicciones. En este escenario, una vez que un método de segmentación es desarrollado y puesto en práctica, no es posible hacer una evaluación cuantitativa de rutina sobre su desempeño, sin requerir una inspección visual de un experto humano [18].

Sin embargo, como se muestran en [19], la incertidumbre inferida de algunos modelos de segmentación podría dar una noción sobre la confianza asociada a determinadas máscaras y resaltar regiones con posibles errores, para poder derivarlas selectivamente a un especialista humano.

Siguiendo este esquema, para poder aumentar la confiabilidad de los modelos de segmentación de imágenes médicas basados en CNNs e incorporarlos en la práctica clínica estándar, es crucial desarrollar modelos precisos y calibrados que sean capaces de informar cuándo sus predicciones fallan.

1.1.3. Seguimiento de ojos para diagnóstico de enfermedades

Recientemente, el seguimiento de ojos mediante video-oculografía (VOG) ha comenzado a utilizarse como herramienta para el diagnóstico de una amplia variedad de enfermedades neurológicas y mentales. En esta aplicación, la estimación debe realizarse con alta precisión, por lo que se utilizan los llamados métodos basados en características, más precisamente los métodos basados en regresión 2D. Aquí, debe estimarse una función que permita mapear el espacio bidimensional de las características a la dirección de la mirada o a las coordenadas de la pantalla en donde el usuario está mirando.

Estas técnicas son muy utilizadas en aplicaciones de videojuegos, interfases no táctiles y otras. Sin embargo, para el diagnóstico médico se requieren niveles de precisión muy superiores en la estimación de la mirada. Surge así una limitación funcional importante: la cabeza de la persona debe permanecer estática, de lo contrario se producen errores entre la dirección estimada y la dirección real. A pesar de que suelen usarse sistemas de sujeción, en personas con ciertas enfermedades neurológicas los movimientos cefálicos involuntarios persisten. Por lo tanto, ser capaz de medir estos movimientos resulta muy importante, no sólo para corregir los errores de estimación que producen, sino también porque esta medición parece ser otro indicador de la presencia o del progreso de ciertas enfermedades.

Dado que este tipo de dispositivos generalmente sólo cuentan con cámaras y no con otro tipo de sensores, los movimientos de la cabeza deben medirse a partir de puntos de referencia detectados en las imágenes. Debido a que las esquinas del ojo son los puntos del contorno cuya posición se ve menos afectada por los movimientos del globo ocular o de los párpados, los cambios en su posición constituyen una buena referencia para inferir los movimientos de la cabeza.

1.2. Objetivos

El objetivo general de esta tesis es desarrollar nuevos métodos para mejorar el desempeño y la credibilidad de las redes convolucionales profundas al ser aplicadas al análisis de imágenes médicas.

Los objetivos específicos son:

- Desarrollar nuevos métodos de post-procesamiento basados en autocodificadores para corregir segmentaciones erróneas, incorporando antecedentes topológicos de estructuras anatómicas.
- Desarrollar nuevas estrategias de entrenamiento de ensambles para incorporar diversidad en los modelos y obtener predicciones calibradas.
- Diseñar e implementar nuevas funciones de pérdida que permitan entrenar CNNs calibradas y con mejor estimación de la confianza asociada a sus predicciones.
- Desarrollar una base de videos de video-oculografía para el diagnóstico de enfermedades neurológicas y mentales, junto a nuevos métodos para localización de puntos de referencia del contorno del ojo y estimación de su posición.
- Analizar relaciones existentes entre los movimientos cefálicos detectados y los errores en la estimación de la mirada durante pruebas en pacientes.

1.3. Organización de la tesis

Esta tesis se encuentra organizada bajo el formato de *tesis por compilación* de la siguiente forma:

- En la presente sección se describió la motivación y los problemas a abordar y se establecieron los objetivos planteados para llevar adelante el trabajo.
- Debido a las distintas aplicaciones abordadas, se decidió organizarlas en las siguientes secciones:
  - En la sección 2 se aborda la segmentación de estructuras anatómicas en imágenes médicas.
  - En la sección 3 se aborda el análisis de incertidumbre aplicado a la segmentación de imágenes médicas.
  - En la sección 4 se aborda el seguimiento de ojos aplicado al diagnóstico de enfermedades neurológicas.

  En cada una de estas secciones se presentan los antecedentes y las técnicas existentes en el estado del arte para cada aplicación. Se introducen los conceptos teóricos necesarios para la comprensión del problema, se detalla brevemente la metodología empleada y se describen los métodos propuestos y resultados obtenidos.
- Finalmente, en la sección 5 se presentan las conclusiones generales y específicas de la tesis.

Capítulo 2

Segmentación de estructuras anatómicas

2.1. Antecedentes

Múltiples alternativas se han propuesto para incorporar antecedentes anatómicos al proceso de segmentación de imágenes médicas (vea [20] para una revisión completa). Una estrategia muy popular para integrar información de forma y topología en métodos de DL para segmentación de imágenes, consiste en modificar la función de pérdida usada para entrenar el modelo. Por ejemplo, Aïcha et. al [15] incorporaron una regularización de alto orden mediante una función de pérdida sensible a la topología. Mediante un enfoque similar, los autores de [21, 22] utilizaron un autocodificador para aprender representaciones de baja dimensión de la imagen, con las cuales imponer restricciones anatómicas durante el entrenamiento de la red. La principal desventaja de estos enfoques es que sólo pueden usarse durante el entrenamiento de redes neuronales y su uso no puede extenderse a otros métodos como por ejemplo los basados en bosques aleatorios (RF, del inglés *Random Forest*).

Otros métodos considerados en la literatura son los métodos de post procesamiento basados en campos aleatorios de Markov [14], o en campos aleatorios condicionales [13]. Estos métodos se basan en el supuesto de que los objetos suelen ser continuos, y por lo tanto, a los píxeles cercanos se les debe asignar la misma clase de objeto. Sin embargo no incorporan antecedentes anatómicos por lo que no garantizan la plausibilidad anatómica de las estructuras.

2.2. Métodos propuestos

2.2.1. Descripción del problema

Dada una base de datos formada por máscaras anatómicas independientes $\mathcal{D}_{\mathcal{A}} = \{S_i^A\}_{0 \leq i \leq |\mathcal{D}_{\mathcal{A}}|}$ (independientes en el sentido de que no necesitan estar asociadas a una imagen de intensidad) se busca aprender un modelo que pueda proyectar al espacio anatómicamente plausible, las segmentaciones $\mathcal{D}_{\mathcal{P}} = \{S_i^P\}_{0 \leq i \leq |\mathcal{D}_{\mathcal{P}}|}$ generadas por clasificadores arbitrarios P.

2.2.2. Auto-codificadores para reducción de ruido

Los auto-codificadores para reducción de ruido (DAE, del inglés *denoising autoencoders*) son redes neuronales diseñadas para reconstruir una entrada limpia a partir de una versión corrupta de la misma [23].

La arquitectura estándar de los DAE sigue un esquema codificador-decodificador. El codificador $f_{cod}(S_i)$ es una función que transforma la imagen de entrada S_i en una representación oculta h, conocida como *código*. Luego este código ingresa al decodificador $f_{dec}(h)$, el cual lo mapea nuevamente a las dimensiones de la imagen de entrada.

La dimensión de las representaciones aprendidas $h = f_{cod}(S_i)$ es mucho menor que la dimensión de la entrada. Esto produce un efecto de cuello de botella que fuerza al codificador a concentrar en h sólo la información mas importante de la imagen de entrada para luego poder reconstruirla. Esta propiedad está basada en el supuesto de las Variedades Diferenciales (*manifold assumption* en inglés) [24], el cual establece que datos de alta dimensión (como por ejemplo las máscaras de segmentación) concentran su información en torno a una variedad no lineal de baja dimensión.

2.2.3. Post-procesamiento con DAE

El método de post-procesamiento con DAE (Post-DAE), utiliza un DAE para reconstruir segmentaciones erróneas reproyectándolas a un espacio anatómicamente plausible. Como se muestra en el diagrama de la Figura 2.1 el DAE es entrenado para aprender una representación anatómicamente plausible de las estructuras anatómicas. Luego, dada una máscara S_i^P obtenida a partir de un método arbitrario de segmentación P (por ejemplo, una CNN o un clasificador basado en RF), se la proyecta a la representación aprendida utilizando el codificador f_{cod} y se reconstruye su máscara anatómicamente plausible correspondiente mediante el decodificador f_{dec}.

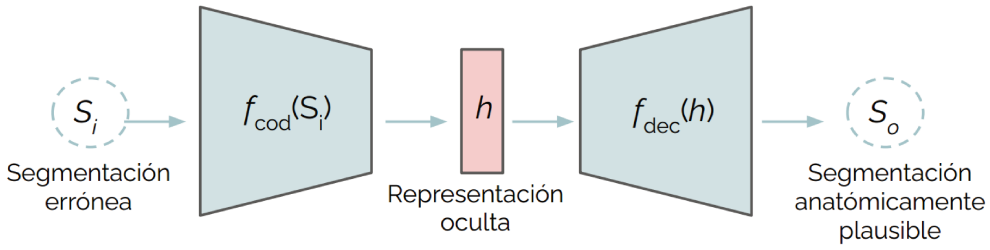


Figura 2.1: Diagrama del Post-DAE utilizado para proyectar segmentaciones erróneas al espacio anatómicamente plausible.

2.2.4. Estrategia de degradación de máscaras

Con el objetivo de obtener un método de post-procesamiento totalmente independiente tanto del método de segmentación inicial como de las características de las imágenes, las segmentaciones *erróneas* usadas para entrenar el DAE se obtuvieron mediante simulación. Para esto, las máscaras originales S_i se degradaron artificialmente aplicando las siguientes funciones $\phi(S_i)$:

- Adición y sustracción de formas geométricas aleatorias (círculos, elipses, lineas) para simular sub y sobre segmentaciones.
- Operaciones morfológicas (erosión, dilatación).
- Permutación aleatoria entre píxeles cercanos a los bordes de las estructuras.

Se utilizaron segmentaciones binarias (pulmones, ventrículo izquierdo) y multi-clase (pulmones, corazón).

2.2.5. Comparación con métodos del estado del arte

A modo de comparación, se utilizó un método basado en campos aleatorios condicionales [25] (CRF, del inglés *Conditional Random Field*). Luego se compararon los resultados obtenidos en dos escenarios diferentes que incluyen radiografía de tórax e imágenes de resonancia magnética cardíaca.

A diferencia del Post-DAE, que sólo utiliza las segmentaciones durante el post-procesamiento y puede ser utilizado sin importar las propiedades de las imágenes, el CRF incorpora información de intensidad de las imágenes originales y por lo tanto debe reajustarse con cada conjunto de datos. Para medir la calidad de las segmentaciones y comparar los métodos se utilizaron el coeficiente Dice (DSC) y la distancia de Hausdorff (HD).

2.2.6. Métodos de segmentación

Para generar las segmentaciones iniciales, se entrenaron versiones binarias y multi-clase de dos métodos estándares: un clasificador basado en RF y una CNN basada en la arquitectura UNet [12].

A fin de evaluar el desempeño del Post-DAE al post-procesar segmentaciones de distintas calidades, se usaron diferentes versiones de los métodos propuestos. En el caso del RF, se varió la profundidad del árbol durante la inferencia. En cambio, para la UNet, se guardaron los modelos cada 5 épocas durante el entrenamiento y se utilizaron todos estos modelos para predecir las segmentaciones de los datos de prueba. En todos los casos se compararon los resultados de post-procesar las máscaras con Post-DAE y con CRF.

Para una descripción detallada de los métodos y su implementación puede referirse a la sección IV-D del Anexo B.

2.3. Experimentos y resultados

2.3.1. Bases de datos

El método se validó en dos escenarios diferentes. Se estudió la segmentación de pulmón y corazón a partir de radiografías de tórax, usando una base de datos de la Sociedad Japonesa de Tecnología Radiológica (JRST). Estas estructuras presentan una alta variabilidad entre sujetos, lo que hace a la tarea de aprendizaje de representaciones aún más desafiante. También se estudió la segmentación del ventrículo izquierdo (LV) a partir de imágenes de RM usando una versión de la base de datos cardíaca de Sunnybrook. Para una descripción detallada de las bases de datos puede referirse a la sección IV-A del Anexo B.

2.3.2. Resultados

En la Figura 2.2 y en la Tabla 2.1 se muestran los resultados obtenidos al post-procesar las segmentaciones generadas con la UNet y el RF respectivamente. En la Figura 2.3 se muestran los resultados cualitativos para algunos ejemplos de cada método y cada estructura anatómica. En todos los casos se observa una mejora consistente al usar el Post-DAE como método de post-procesamiento, que se acentúa para las máscaras de baja calidad. En estos casos se obtienen mejoras sustanciales tanto en términos de DSC como en HD, proyectando las máscaras de segmentación erróneas al espacio anatómicamente plausible. Cuando las máscaras originales son de buena calidad, el Post-DAE continúa mejorando significativamente el HD al eliminar las segmentaciones espurias (por ejemplo, agujeros en el pulmón o burbujas aisladas) que se mantienen incluso en las predicciones de los modelos bien entrenados. También se observa como el Post-DAE supera significativamente al CRF como método de post-procesamiento.

Para un mayor análisis de los resultados obtenidos referirse a la sección V del Anexo B.

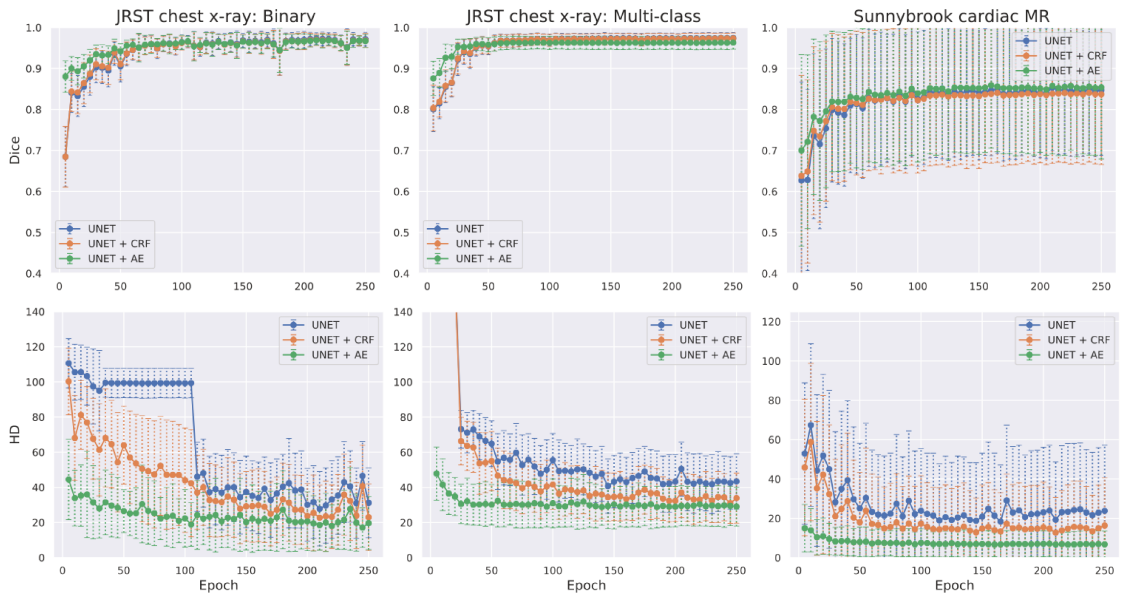


Figura 2.2: Valor medio y desvío estándar de los resultados obtenidos al post-procesar las máscaras generadas con la UNet para las distintas épocas de entrenamiento.

2.3.3. Limitaciones

Se analizó el comportamiento y las limitaciones del Post-DAE al ser aplicado a segmentaciones de estructuras patológicas. Por ejemplo, máscaras de órganos deformados por alguna enfermedad o con grandes oclusiones que luzcan completamente diferente de los casos anatómicamente plausibles. Para este estudio se utilizó una base de datos de radiografías de tórax que incluye pacientes diagnosticados con Tuberculosis (más detallas en la sección V-A del Apéndice B). Cada imagen de rayos-X cuenta con dos máscaras diferentes. Una *máscara de aire* que delimita únicamente la cavidad de aire dentro del pulmón e ignora las partes cubiertas de fluido (Fig.

Tabla 2.1: Valor medio y desvío estándar de los resultados obtenidos al post-procesar las máscaras generadas con el RF. Los números en negrita indican que el Post-DAE supera los otros métodos con significancia estadística de acuerdo al test de Wilcoxon con corrección de Bonferroni.

Segmentaciones		Radiografía de tórax: JRST						S. RM Cardíaca
		Multi-clase					Binaria	Binaria
		Profundidad: 8	Profundidad: 12	Profundidad: 16	Profundidad: 20	Modelo completo	Modelo completo	Modelo completo
Dice	RF	0.858 (0.042)	0.913 (0.033)	0.936 (0.029)	0.949 (0.029)	0.956 (0.028)	0.781 (0.070)	0.46 (0.24)
	RF + CRF	0.860 (0.042)	0.914 (0.032)	0.937 (0.029)	0.950 (0.029)	0.956 (0.027)	0.795 (0.074)	0.44 0.25
	RF + DAE	**0.922 (0.024)**	**0.943 (0.020)**	**0.948 (0.018)**	0.951 (0.018)	0.951 (0.019)	**0.865 (0.056)**	**0.47 (0.25)**
HD	RF	102.26 (11.68)	96.00 (14.31)	88.70 (14.04)	77.45 (12.98)	72.28 (14.07)	91.41 (17.52)	27.73 (9.89)
	RF+CRF	101.17 (12.94)	92.97 (14.72)	81.26 (12.73)	74.51 (13.10)	67.29 (13.53))	80.45 (22.28)	26.87 (10.03)
	RF +DAE	**63.73 (11.85)**	**60.72 (12.20)**	**62.47 (15.16)**	**62.95 (16.84)**	**60.69 (14.12)**	**32.01 (18.44)**	**23.60 (9.88)**

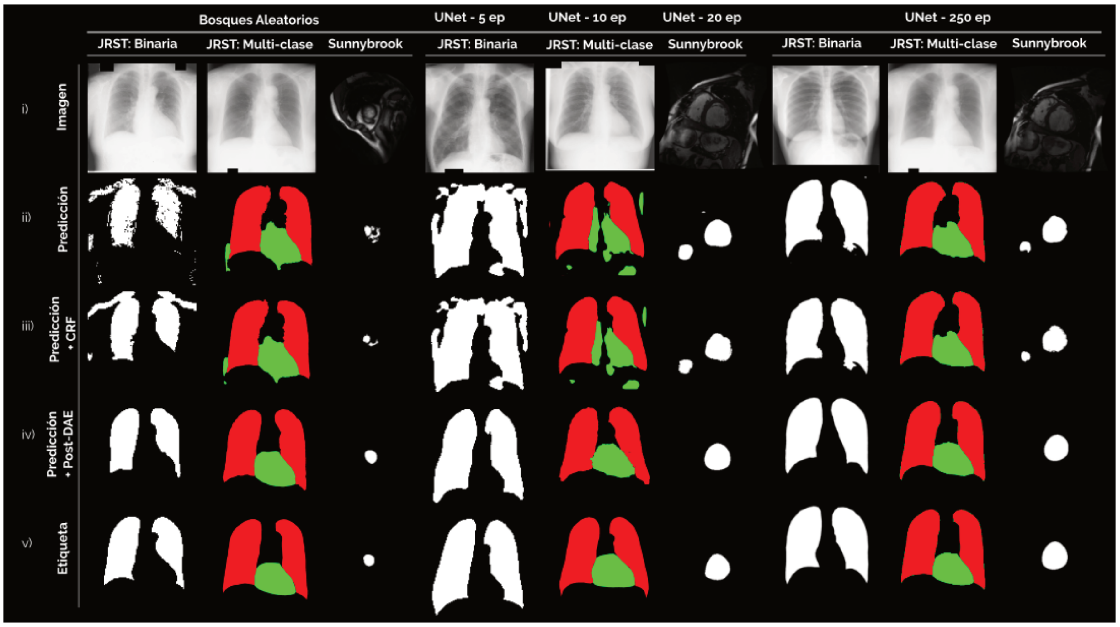


Figura 2.3: Resultados cualitativos. Desde la primer fila a la última pueden observarse: (i) Las imágenes originales (Rayos-X o RM); (ii) Máscaras originales obtenidas con cada uno de los métodos analizados; (iii) máscaras post-procesadas utilizando CRF; (iv) máscaras post-procesadas por el Post-DAE; y (v) máscaras de referencia.

2.4-ii) y una *máscara anatómica* que delimita la forma anatómica que se esperaría observar en ese pulmón, incluyendo las áreas ocluidas (Fig. 2.4-iv).

Se utilizó el Post-DAE previamente entrenado para post-procesar las *máscaras de aire* y se observó que el método tiende a mapearlas a las *máscaras anatómicas* correspondientes (Fig. 2.4 - columnas 1 y 2) . Sin embargo, cuando las anormalidades son muy pronunciadas (Fig. 2.4 - columnas 3 y 4) la máscara anatómica no puede ser reconstruida completamente. Para un análisis cuantitativo referirse a sección V-A del Apéndice B.

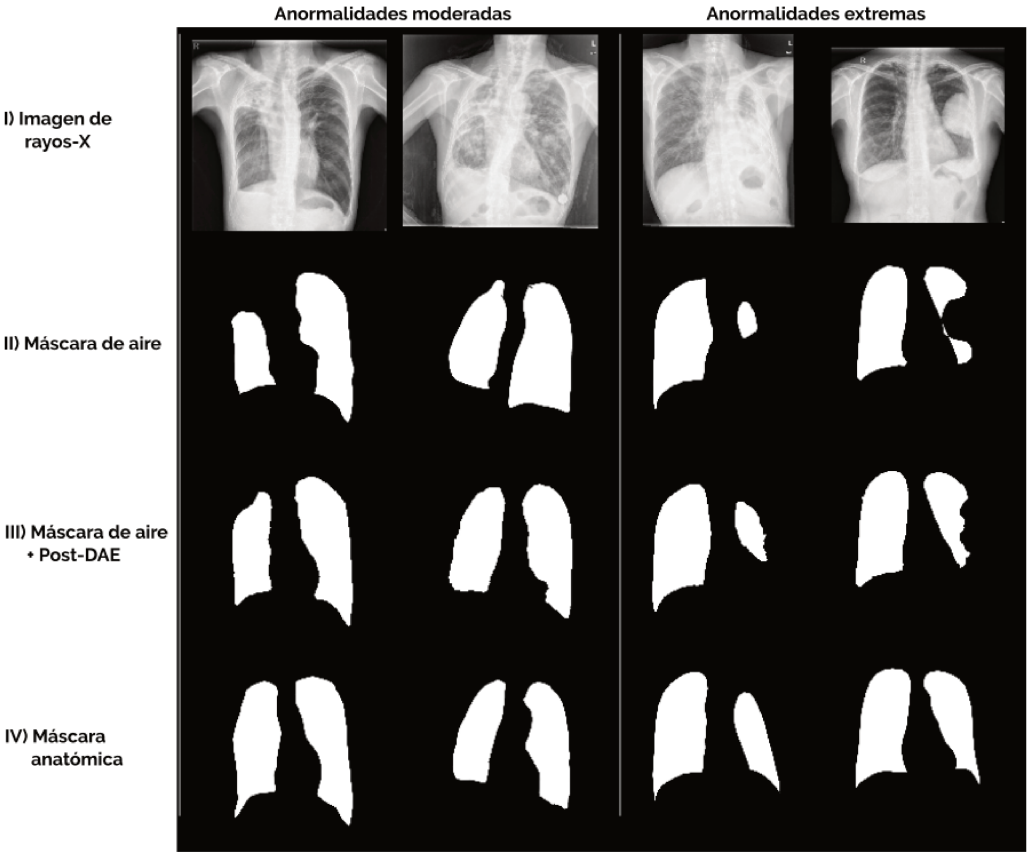


Figura 2.4: Resultados cualitativos obtenidos al aplicar Post-DAE en mascaras de pulmones con anormalidades moderadas y extremas en pacientes con tuberculosis.

Capítulo 3

Análisis de incertidumbre

3.1. Antecedentes

Existen distintos métodos que permiten cuantificar la incertidumbre asociada a las predicciones de las redes neuronales profundas (DNN, del inglés *deep neural networks*). Un ejemplo son las redes neuronales Bayesianas. Los pesos de estas redes no son valores fijos sino distribuciones gaussianas a través de las cuales puede calcularse la incertidumbre del modelo. La desventaja de estas redes es que al ser modelos más complejos, requieren muchos mas parámetros para ser optimizadas, lo que dificulta su adopción en la práctica. Una alternativa muy utilizada es el método de Monte Carlo Dropout, el cual intenta aproximar -en sentido bayesiano- un proceso gaussiano probabilístico al momento de la inferencia. Esto se consigue realizando múltiples iteraciones sobre el conjunto de prueba, aplicando dropout de forma aleatoria. De esta forma se obtiene una distribución de las predicciones del modelo y se puede analizar la incertidumbre asociada. Si bien esta técnica ha sido explorada en el contexto de segmentación de imágenes médicas [26, 27], trabajos recientes [28, 29] han mostrado que las redes neuronales Bayesianas suelen generar predicciones muy similares entre sí. En cambio, un método no bayesiano muy popular que ha demostrado generar predicciones más diversas y por lo tanto obtener mejores resultados en la estimación de incertidumbre, es el ensamble de redes neuronales. Esta es una estrategia muy sencilla para mejorar tanto la robustez como el desempeño en calibración de los modelos predictivos [30, 31].

Un enfoque alternativo consiste en abordar el problema de la calibración al momento del entrenamiento de las redes profundas. Al utilizar funciones de pérdidas basadas en reglas de puntuación como la entropía cruzada (CE), se esperaría recuperar la distribución original de incertidumbre. Sin embargo, debido a la gran capacidad de este tipo de redes y a la poca cantidad de datos disponibles, sobre todo en aplicaciones como el análisis de imágenes médicas, estas funciones tienden a sobre-ajustarse y a sobre-estimar la confianza de sus predicciones [16, 32]. Sumado a esto, se ha observado que entrenar las redes con una función de pérdida basada en el coeficiente dice (DSL), mejora la calidad de las segmentaciones pero empeora aún mas la calibración del modelo. Por ejemplo, los autores de [33] analizaron el comportamiento de las redes de segmentación entrenadas con las dos funciones de pérdida más usadas en la literatura (DSL y CE). En linea con [34, 35], mostraron que la función elegida afecta directamente la calidad de las segmentaciones y la calibración del modelo, destacando que los modelos entrenados con DSL tienden a estar

muy mal calibrados y a ser *sobre-confiados*. Es por esto que se ha hecho énfasis en la necesidad de explorar nuevas funciones de pérdida que mejoren tanto la calidad de la segmentación como de la calibración.

3.1.1. Ensamble de redes neuronales

El ensamble de DNNs es una estrategia simple que se utiliza para incrementar la robustez y la calibración de los modelos predictivos [30, 31]. El enfoque clásico consiste en generar distintas soluciones funcionales entrenando un único modelo bajo condiciones diferentes. Algunas de las técnicas para entrenamiento de ensambles son el corrimiento de dominio [36], el ensamblado por batch [37] y la variación de hiperparámetros [38], entre otras. Una vez entrenados los modelos, se promedian sus predicciones para lograr una disminución de los errores individuales y un aumento en la calibración. Para que el ensamble sea efectivo y robusto, es imprescindible asegurar la *diversidad* entre los modelos constituyentes. Se han propuesto distintas soluciones para este fin, como el uso de variables latentes [39] o la integración de mecanismos de atención que incentiven a diferentes modelos a focalizarse en diferentes partes de los objetos [40].

3.1.2. Penalización de confianza

Distintos trabajos han estudiado la forma de obtener una mejor estimación de la incertidumbre modificando las funciones de pérdida existentes. Pereyra et. al. [41], propusieron regularizar la salida de una DNN integrando un término en la función objetivo para penalizar la baja entropía de las predicciones. Además, los autores propusieron una variante del método, denominada suavizado de etiquetas, que penaliza la baja entropía de manera indirecta, al usar una versión suavizada de las etiquetas para calcular la CE. Por otro lado, los autores de [32] mostraron empíricamente la mejora en calibración que se obtiene al utilizar la *focal loss* como función objetivo.

Si bien estos enfoques son prometedores en problemas de clasificación, su beneficio es menos claro en tareas de segmentación. En un trabajo reciente, Islam et. al. [42] destacaron que una limitación de suavizar las etiquetas con una distribución uniforme es el hecho de no considerar la relación espacial existente entre los píxeles. Por ejemplo, que la incertidumbre del modelo siempre es mayor cerca de los bordes de las estructuras que en el centro de las mismas.

3.2. Métodos propuestos

3.2.1. Calibración en segmentación de imágenes

Dada una base de datos $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})_i\}_{0 \leq i \leq |\mathcal{D}|}$ compuesta de imágenes $\mathbf{x}$ y sus máscaras de segmentación correspondientes $\mathbf{y}$, el objetivo es entrenar un modelo que aproxime la distribución condicional subyacente $p(\mathbf{y}|\mathbf{x})$, que mapee las imágenes de entrada $\mathbf{x}$ a las segmentaciones $\mathbf{y}$. Por lo tanto, $p(y_j = k|\mathbf{x})$ debe indicar la probabilidad de que a un determinado pixel (o voxel) j se le asigne la clase k perteneciente al conjunto de clases posibles $\mathcal{C}$. Esta distribución suele aproximarse utilizando una red neuronal $f_{\mathbf{w}}$ parametrizada por sus pesos $\mathbf{w}$. En otras palabras,

$f_{\mathbf{w}}(\mathbf{x}) = p(\mathbf{y}|\mathbf{x}; \mathbf{w})$. Estos pesos $\mathbf{w}$ se calculan minimizando una determinada función de pérdida sobre los datos de entrenamiento. Si el modelo está bien calibrado, la probabilidad correspondiente a la clase puede verse como la confianza del modelo y puede usarse como una medida de la incertidumbre asociada a la predicción del pixel [33].

3.2.2. Ensamble de redes neuronales

Dado un conjunto de redes de segmentación $\{f_{\mathbf{w}^1}, f_{\mathbf{w}^2} ... f_{\mathbf{w}^N}\}$, la forma más sencilla de construir un ensamble $f_{\mathbf{E}}$ consiste en promediar sus predicciones:

$$f_{\mathbf{E}}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} f_{\mathbf{w}^i}(\mathbf{x}). \tag{3.1}$$

Se ha observado que si se aplican restricciones de ortogonalidad a los filtros durante el entrenamiento de las DNNs, estas aprenden características mas decorrelacionadas y menos redundantes, haciendo un mejor uso de la capacidad del modelo [43, 44]. En este trabajo, se extiende este principio para impulsar la diversidad de los ensambles imponiendo restricciones de ortogonalidad tanto dentro de cada modelo, como entre los distintos modelos del ensamble.

3.2.3. Restricciones de ortogonalidad

La similitud coseno es una métrica utilizada para cuantificar la ortogonalidad (o decorrelación) entre dos vectores $\mathbf{x}$ y $\mathbf{y}$. Toma valores entre -1 (vectores opuestos) y 1 (vectores iguales), con 0 indicando vectores ortogonales. Esta medida se puede definir como:

$$\mathrm{SIM}_C(\mathbf{x}, \mathbf{y}) = \frac{< \mathbf{x}, \mathbf{y} >}{||\mathbf{x}||\ ||\mathbf{y}||} \tag{3.2}$$

Con el objetivo de impulsar la ortogonalidad entre los filtros, siguiendo el planteo de [43], se propone agregar en la función de pérdida un término de regularización definido como el cuadrado de la similitud coseno. Una de las ventajas de esta medida es que considera tanto las correlaciones positivas como negativas.

3.2.4. Ensambles ortogonales

Se incluyeron dos términos de regularización en la función objetivo. El primero es un término de ortogonalidad *intra-modelo* ($\mathcal{L}_{\mathrm{SelfOrth}}$), que penaliza la correlación entre los filtros de cada capa de una determinada CNN. Siendo l una capa convolucional, este término se calcula como:

$$\mathcal{L}_{\mathrm{SelfOrth}}(\mathbf{w}_l) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathrm{SIM}_C(\mathbf{w}_{l,i}, \mathbf{w}_{l,j})^2, \tag{3.3}$$

donde $\mathbf{w}_{l,i}$ y $\mathbf{w}_{l,j}$ son las versiones vectorizadas de cada uno de los n filtros de la capa l. El segundo es un término de ortogonalidad *inter-modelo* ($\mathcal{L}_{\mathrm{InterOrth}}$), que penaliza la correlación entre los filtros de las distintas CNNs pertenecientes al

ensamble. Siguiendo un esquema secuencial, el término de la capa l del modelo N_e se calcula de la siguiente manera:

$$\mathcal{L}_{\text{InterOrth}}(\mathbf{w}_l; \{\mathbf{w}_l^e\}_{0\leq e<N_e}) = \frac{1}{N_e}\sum_{e=0}^{N_e-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{SIM}_C(\mathbf{w}_{l,i}, \mathbf{w}_{l,j}^e)^2, \tag{3.4}$$

donde $\{\mathbf{w}_l^e\}_{0\leq e<N_e}$ son los parámetros de los $N_e - 1$ modelos previamente entrenados.

Finalmente, la función objetivo del ensamble se define como:

$$\mathcal{L} = \mathcal{L}_{Seg} + \lambda\sum_{l}\Big(\mathcal{L}_{\text{SelfOrth}}(\mathbf{w}_l) + \mathcal{L}_{\text{InterOrth}}(\mathbf{w}_l; \{\mathbf{w}_l^e\})\Big), \tag{3.5}$$

donde $\mathcal{L}_{Seg}$ es la función de pérdida de segmentación (por ejemplo DSL o CE) y λ es el hiper parámetro que controla la influencia de los términos de ortogonalidad.

3.2.5. Entropía máxima para predicciones erróneas

En general, la entropía de una distribución de probabilidad $p(x)$ con variable aleatoria x se puede calcular como:

$$\mathcal{H}(p(x)) = -\sum_i p(x_i)\ln p(x_i), \tag{3.6}$$

donde x_i son los valores posibles que puede tomar la variable aleatoria. Una de las propiedades de la entropía, es que su valor es máximo cuando la distribución p es uniforme, es decir, cuando la incertidumbre es máxima. En este trabajo, se hace uso del concepto de entropía para reducir la confianza (es decir, aumentar la incertidumbre) de las predicciones erróneas de la red.

Se busca entonces entrenar un modelo capaz de advertir la presencia de regiones difíciles de segmentar sin perder la confianza en sus predicciones ante regiones identificables. Para esto, como se detalla en la sección 3 del Anexo D, se propone penalizar exclusivamente la sobre-confianza de los píxeles mal clasificados durante el entrenamiento. Dadas las predicciones de clase a nivel de píxel $\hat{y}_i$ realizadas por una red neuronal, y sus etiquetas asociadas y_i, se define el conjunto de píxeles mal clasificados como $\hat{\mathbf{y}}_w = \{y_i|\hat{y}_i \neq y_i\}$ y se computa el negativo de la entropía para este conjunto:

$$-\mathcal{H}(\hat{\mathbf{y}}_w) = \frac{1}{|\hat{\mathbf{y}}_w|}\sum_{k\in\mathcal{C}, i\in\hat{\mathbf{y}}_w} p(y_i = k|\mathbf{x};\mathbf{w})\log p(y_i = k|\mathbf{x};\mathbf{w}) \tag{3.7}$$

De esta forma, minimizar el término (3.7), como $\text{mín}_{\mathbf{w}} -\mathcal{H}(\hat{\mathbf{y}}_w)$ equivale a maximizar la entropía de $\hat{\mathbf{y}}_w$. Luego, siendo $\hat{\mathbf{y}}$ el conjunto completo de píxeles, la función objetivo puede calcularse como:

$$\mathcal{L} = \mathcal{L}_{Seg}(\mathbf{y}, \hat{\mathbf{y}}) - \lambda\mathcal{H}(\hat{\mathbf{y}}_w) \tag{3.8}$$

donde $\mathcal{L}_{Seg}$ corresponde a la función de pérdida de segmentación (CE o DSL) y λ da la importancia relativa de cada término dentro de la función objetivo.

3.2.6. Alternativa a la maximización de entropía

Maximizar la entropía de $\mathbf{y}_w$ equivale a minimizar la divergencia de *Kullback-Leibler* (KL) entre $\mathbf{y}_w$ y una distribución uniforme $\mathbf{q}$.

Como regularizador alternativo, se implementó una variante de la divergencia de KL que intenta acercar las probabilidades de salida en $\mathbf{y}_w$ a una distribución uniforme (todos los elementos en el vector de salida q iguales a $\frac{1}{K}$), lo cual corresponde al nivel de incertidumbre máximo. Este término puede ser expresado como:

$$\mathcal{D}_{KL}(\mathbf{q}||\hat{\mathbf{y}}_w) \overset{\text{C}}{=} \mathcal{H}(\mathbf{q}, \hat{\mathbf{y}}_w), \tag{3.9}$$

donde el símbolo C= indica igualdad escalando por una constante asociada a la cantidad de clases. Es importante notar que a pesar de que los dos términos (3.7) y (3.9) empujan $\hat{\mathbf{y}}_w$ hacia una distribución uniforme, la dinámica de sus gradientes es diferente y por tanto ambas versiones serán consideradas para evaluar su impacto durante el entrenamiento de la red.

3.3. Experimentos y resultados

3.3.1. Bases de datos

Los ensambles ortogonales se validaron en dos tareas complejas de segmentación de lesiones cerebrales a partir de imágenes de RM: tumores (BraTS 2020 [45, 46, 47]) e hiperintensidades de la sustancia blanca (WMH [48]). A fin de evaluar el segundo método, además de los datos de WMH, se utilizó una base de datos de segmentación de ventrículo izquierdo (LA) [49] que provee 100 imágenes de RM con las máscaras asociadas. Para una descripción detallada de los datos y las particiones, referirse a la sección 4 de los Anexos C y D.

3.3.2. Métricas de calibración

Brier score: Es una métrica muy usada para medir el desempeño en calibración [50], es una regla de puntuación cuyo valor óptimo corresponde a una predicción perfecta. En otras palabras, un sistema perfectamente discriminativo y perfectamente calibrado tendrá un Brier score igual a cero. En el contexto de segmentación de imágenes médicas, para una imagen con N pixeles (voxeles), el Brier score se puede definir como:

$$Br = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{C}|}\sum_{k=1}^{|\mathcal{C}|}\Big(p(y_i = k|\mathbf{x};\mathbf{w}) - \mathbb{1}[y_i = k]\Big)^2, \tag{3.10}$$

donde $\mathbb{1}[y_i = k]$ es una función indicadora que toma el valor 1 cuando y_i (la etiqueta del pixel i) es igual a k, y 0 de otra forma.

Brier Score estratificado: En problemas con altos desbalances de clases, como la segmentación de lesiones cerebrales en donde la mayoría de los pixeles corresponden al fondo, el modelo puede estar bien calibrado desde un punto de vista global pero no estarlo cuando sólo se mira a las clases minoritarias. En este escenario, la clase mayoritaria domina el Brier Score y la mala calibración que existe en la clase de

interés no se ve reflejada. Wallace et al. [51] propusieron el Brier score estratificado para medir la calibración en problemas de clasificación binarios con altos grados de desbalance. En esta tesis, extendemos este concepto a la tarea de segmentación y utilizamos el Brier score estratificado para medir la calibración de cada estructura anatómica de forma individual. Para una determinada imagen con etiqueta $\mathbf{y}$, el Brier score *estratificado* para la clase k (Br^k) se calcula utilizando sólo el subconjunto de píxeles $\mathcal{P}_k = \{p : y_p = k\}$ mediante

$$Br^k = \frac{1}{|\mathcal{P}_k|} \sum_{i \in \mathcal{P}_k} \Big(p(y_i = k|\mathbf{x}; \mathbf{w}) - \mathbb{1}[y_i = k] \Big)^2. \tag{3.11}$$

Diagramas de confianza y error esperado de calibración. Una forma de medir la calidad de la calibración de un modelo, es comparar la confianza de sus predicciones contra la frecuencia de observación de los ejemplos positivos. Para esto, se divide el intervalo $[0, 1]$ en M bins equiespaciados, donde el bin i corresponde al intervalo $\left(\frac{i-1}{M}, \frac{i}{M}\right]$, y B_i denota el conjunto de píxeles cuya confianza pertenece al bin i. Luego, se calcula la frecuencia de ejemplos positivos para cada bin como $freq(B_i) = \frac{1}{|B_i|} \sum_{i \in B_i} \mathbb{1}[\hat{y}_i = k]$ y su confianza como $C_i = \frac{1}{|B_i|} \sum_{i \in B_i} p(y_i = k|\mathbf{x}; \mathbf{w})$. Finalmente, el diagrama de confianza se grafica como la frecuencia de ejemplos positivos en función de la confianza promedio del bin. Se dice que un modelo está perfectamente calibrado cuando estos valores se acercan a la recta identidad.

Para resumir estas estadísticas en un único valor, se calcula el error esperado de calibración (ECE) como:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \, |freq(B_m) - conf(B_m)|$$

3.3.3. Resultados con ensambles ortogonales

Se implementaron 3 configuraciones de ensambles. Un *ensamble aleatorio* donde no se agregó ninguna regularización y los modelos se entrenaron para reducir el error de segmentación $\mathcal{L}_{Seg}$. Un ensamble con ortogonalidad *intra-modelo*, donde se agregó el término $\mathcal{L}_{\text{SelfOrth}}$ a la función objetivo, pero no se impuso ninguna regularización entre los modelos constituyentes del ensamble. Y el ensamble ortogonal propuesto en esta tesis al que se le incorporó el término de ortogonalidad *inter-modelo* en la función objetivo (como se ve en la ecuación 3.5).

Se entrenaron 10 modelos por cada configuración y luego, durante la evaluación, se ensamblaron grupos de 1, 3 y 5 CNNs promediando las predicciones individuales de cada uno. En la Figura 3.1 se muestran los resultados para la segmentación de tumores y de WMH. Para medir el desempeño en segmentación se utilizó el DSC, mientras que la calidad en calibración se midió con el Brier score y el Brier score estratificado.

Se puede observar que con sólo agregar el término de ortogonalidad *intra-modelo* se supera al modelo de referencia en todos los grupos y con todas las métricas. Esta mejora es aún más notable cuando se impone explícitamente la diversidad del ensamble incorporando el término de ortogonalidad *inter-modelo* durante el entrenamiento secuencial. Las mejoras obtenidas tanto en calibración como en segmentación

Training loss		Segmentation performance				Calibration performance					
		dice coefficient		HD		Brier (1^{-4})		Brier^{+}		ECE (1^{-3})	
		WMH	LA	WMH	LA	WMH	LA	WMH	LA	WMH	LA
$\mathcal{L}_{dice}$	–	**0.770 (0.100)**	**0.886 (0.060)**	24.041 (10.845)	**28.282 (11.316)**	6.717 (4.184)	29.182(15.068)	0.257 (0.125)	0.107 (0.090)	0.667 (0.414)	28.861 (15.009)
	$+\mathcal{L}_H(\hat{Y})$ [41]	0.769 (0.099)	0.885 (0.050)	21.608 (8.830)	29.811 (11.168)	6.751 (4.194)	29.019(12.709)	0.249 (0.125)	0.109 (0.077)	0.670 (0.415)	28.458 (12.514)
	$+\mathcal{L}_H(\hat{Y}_w)$	0.758 (0.108)	0.873 (0.069)	21.243 (8.755)	29.374 (10.965)	5.874 (3.875)	24.709(13.774)	0.244 (0.124)	0.103 (0.086)	0.510 (0.350)	18.796 (15.005)
	$+\mathcal{L}_{KL}(\hat{Y}_w)$	**0.770 (0.098)**	0.881 (0.064)	**20.804 (8.122)**	28.415 (12.860)	**5.564 (3.586)**	**23.182(12.464)**	**0.231 (0.114)**	**0.095 (0.077)**	**0.471 (0.318)**	**15.587 (13.391)**
$\mathcal{L}_{CE}$	–	0.755 (0.111)	0.878 (0.070)	21.236 (7.735)	**27.163 (11.967)**	6.462 (4.141)	24.447 (14.876)	0.280 (0.140)	0.108 (0.092)	0.620 (0.400)	18.383 (16.700)
	$+\mathcal{L}_H(\hat{Y})$ [41]	0.760 (0.109)	0.881 (0.070)	23.124 (9.523)	29.464 (14.389)	6.369 (4.018)	23.539 (11.903)	0.242 (0.125)	0.096 (0.070)	4.100 (0.582)	15.590 (14.002)
	$+\mathcal{L}_H(\hat{Y}_w)$	0.770 (0.095)	**0.883 (0.058)**	**19.544 (7.254)**	28.560(13.352)	5.417 (3.547)	**22.506 (11.903)**	0.217 (0.104)	**0.093 (0.071)**	0.436 (0.301)	**15.242 (13.730)**
	$+\mathcal{L}_{KL}(\hat{Y}_w)$	**0.777 (0.093)**	0.876 (0.070)	22.298 (9.566)	28.736 (11.972)	**5.331 (3.478)**	24.085 (13.330)	**0.213 (0.099)**	0.105 (0.090)	**0.422 (0.289)**	17.348 (14.786)
$\mathcal{L}_{FL}$		0.753 (0.113)	0.881 (0.064)	21.931 (8.167)	28.599 (11.968)	5.760 (3.732)	23.928 (11.626)	0.243 (0.130)	0.095 (0.066)	0.438 (0.310)	25.998 (12.740)

Tabla 3.1: Evaluación cuantitativa: Se muestra el valor medio y la desviación estándar de las predicciones obtenidas con las distintas configuraciones de función objetivo para segmentación de WMH y LA. Se indica en negrita el método que obtuvo el mejor desempeño.

demuestran los beneficios de la estrategia propuesta para generar modelos precisos y bien calibrados. Para un mayor análisis de los resultados referirse a la sección 5 del Anexo C.

3.3.4. Resultados con entropía máxima para predicciones erróneas

La efectividad del método se evaluó entrenando dos DNN del estado del arte (UNet [12] y ResUNet [52]) con diferentes configuraciones de su función objetivo. Se usaron la CE y el DSL como funciones de referencia y, para comparar con métodos mas sofisticados, se implementaron la función focal-loss ($\mathcal{L}_{FL}$) y el método de penalización de entropía detallado en [41]. En este caso no solo se usó la versión original del método ($\mathcal{L}_{CE}(\mathbf{y}, \hat{\mathbf{y}}) + \beta\mathcal{L}_H(\hat{\mathbf{y}})$) sino que también se combinó el término de penalización de entropía usando el DSL como función de segmentación. Finalmente, los dos términos de penalización de confianza para predicciones erróneas, ($\mathcal{L}_H(\hat{Y}_w)$ y $\mathcal{L}_{KL}(\hat{Y}_w)$), se combinaron con las funciones de segmentación y se compararon todas las variantes (para una descripción completa referirse a la sección 3 del Anexo D).

En la Tabla 3.1 se muestran los resultados obtenidos con la UNet para la segmentación de WMH y LA. Se observa que al agregar el término propuesto, se mantiene o incluso mejora el desempeño de la segmentación mientras que se obtiene una mejora significativa en calibración para todas las métricas y bases de datos.

En la Figura 3.2 se muestran ejemplos de los mapas de probabilidad obtenidos para cada función de pérdida. Vale la pena destacar la mejora significativa que se logra cuando el DSL es usado en la función objetivo. En esos casos, se observa claramente que las predicciones realizadas por los modelos de referencia, tienden a ser sobre-confiadas y asignan probabilidades iguales a 0 o a 1. Sin embargo, cuando se agrega el regularizador propuesto, el modelo tiende a usar un mayor rango de valores, asignando valores cercanos a 0.5 (marcados en rojo) para los píxeles mas difíciles.

En la figura 3.3 se muestran los diagramas de confianza y las distribuciones de probabilidad producidas por cada uno de los métodos en las dos bases de datos. En los diagramas de confianza, las curvas mas cercanas a la diagonal indican redes mejor calibradas. Se puede observar fácilmente que al maximizar la entropía de las predicciones erróneas mediante los regularizadores propuestos, se obtienen modelos mejor calibrados. Para un mayor análisis y resultados complementarios, refirirse a la sección 5 del Anexo D.

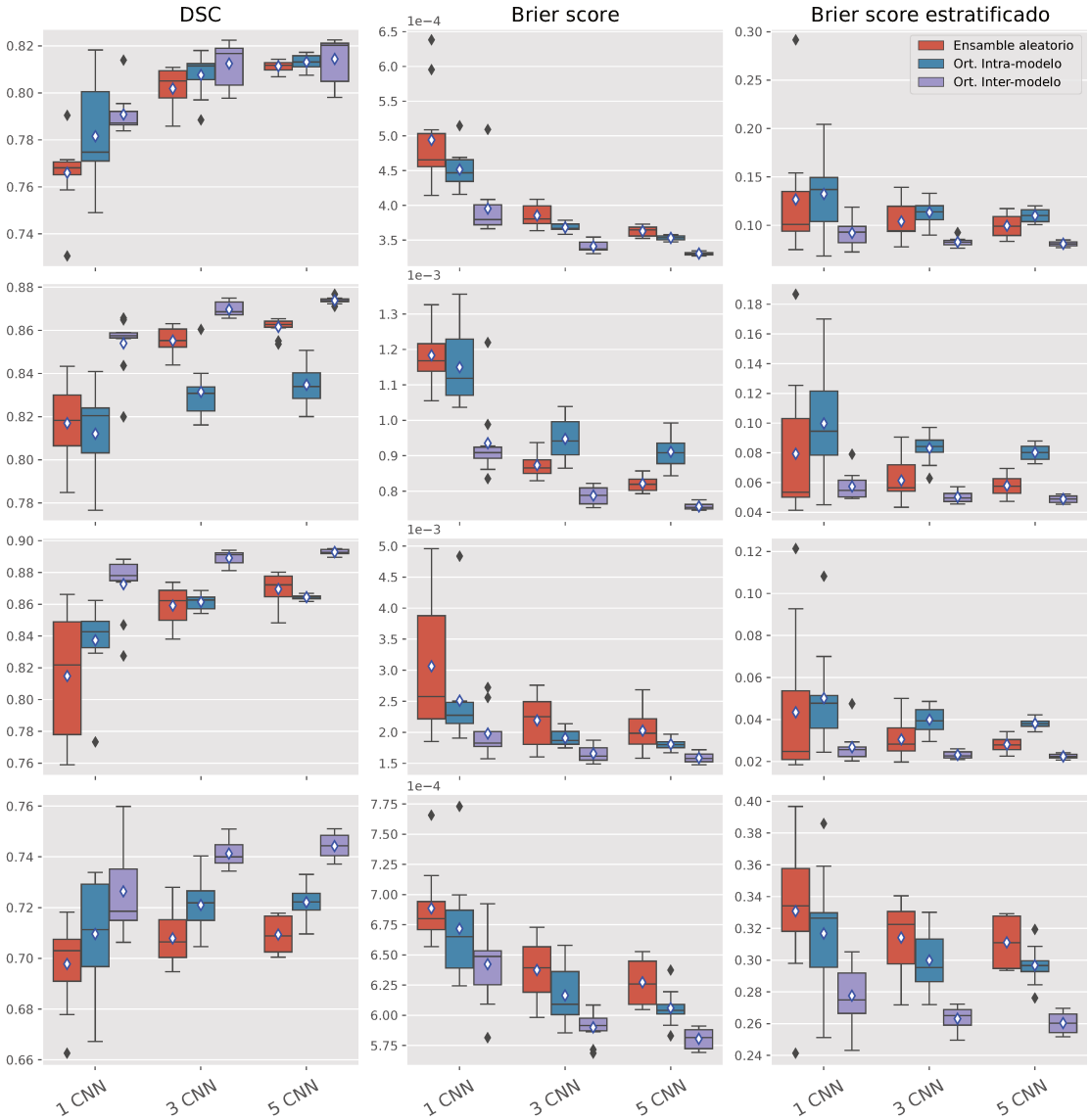


Figura 3.1: Evaluación cuantitativa de las distintas configuraciones de ensambles: Las filas, de arriba hacia abajo, muestran los resultados para: (i) enhanced tumor; (ii) tumor core; (iii) whole tumor; (iv) WMH. Los gráficos de cajas muestran el valor medio y la desviación estándar de las predicciones obtenidas con: las redes individuales sin ensamblar, el ensamble de 3 CNNs y el ensamble de 5 CNNs.

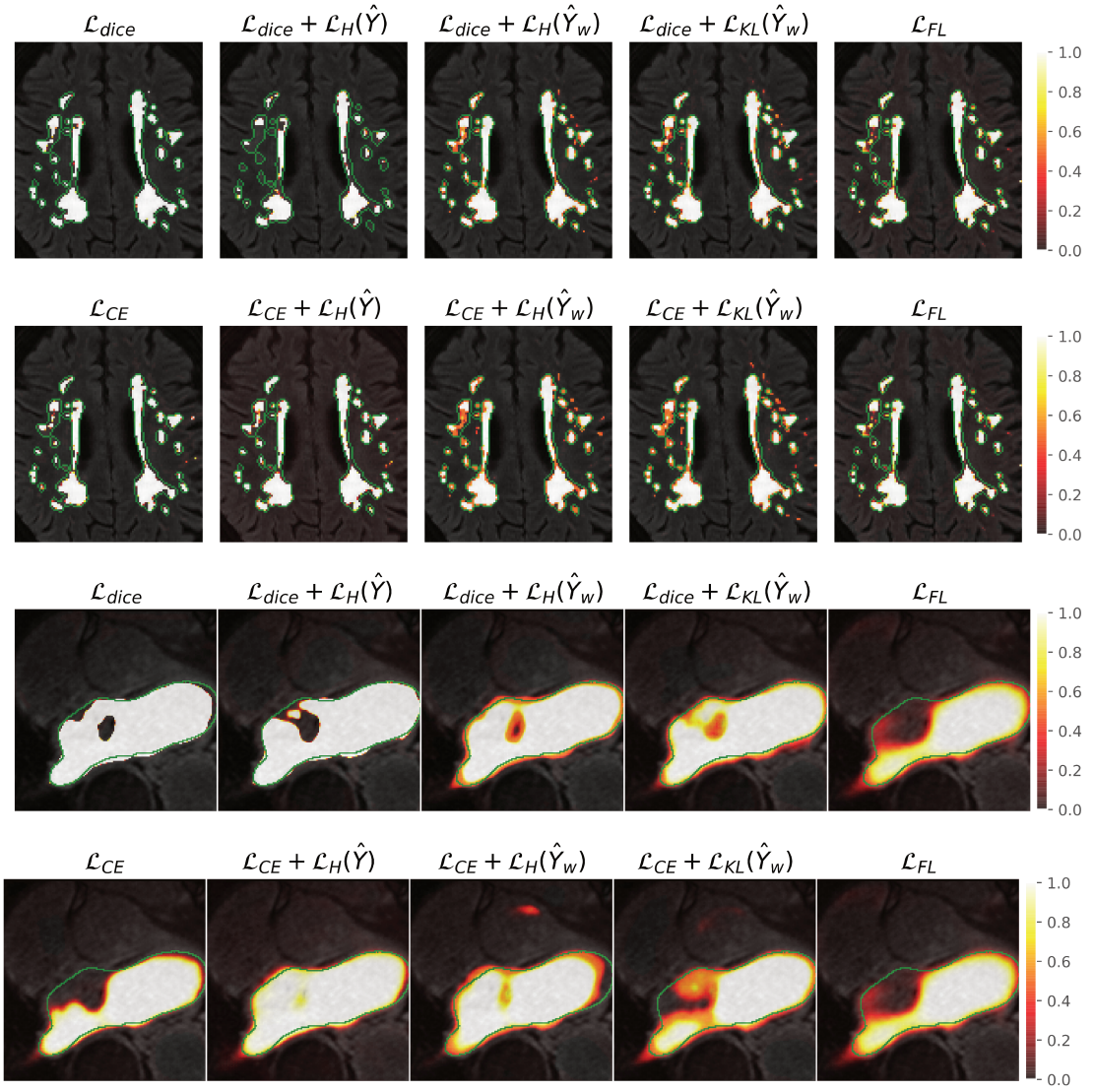
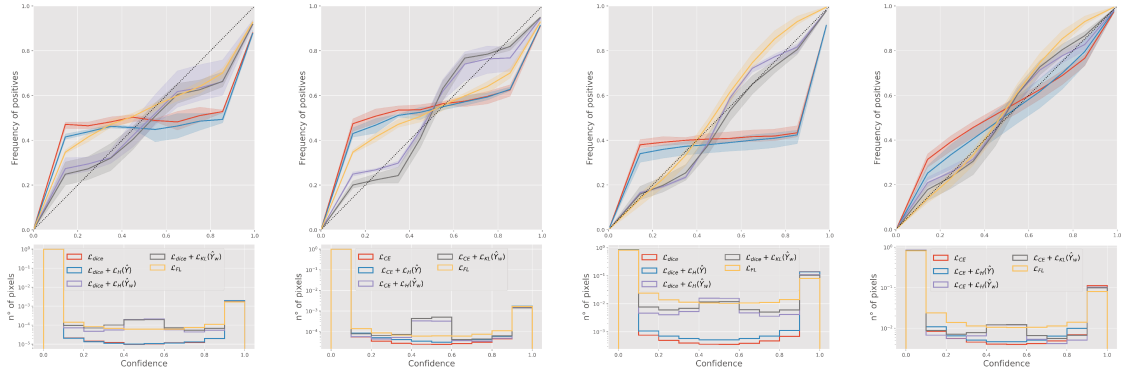


Figura 3.2: Evaluación cualitativa de los mapas de probabilidad generados por las distintas configuraciones de función objetivo para segmentación de WMH y LA.



(a) Segmentación de WMH (b) Segmentación de WMH (c) Segmentación de LA (d) Segmentación de LA

Figura 3.3: Diagramas de confianza (arriba) y distribuciones de probabilidad (abajo) calculados sobre todo el volumen de imágenes de test para las distintas configuraciones de función objetivo.

Capítulo 4

Seguimiento de ojos

4.1. Antecedentes

Se han desarrollado distintos algoritmos con el fin de identificar automáticamente la ubicación de puntos de referencia en imágenes o videos. La mayoría, se han diseñado para detectar puntos característicos de todo el rostro [53, 54] y por lo tanto utilizan imágenes de rostros completos. Otros trabajos, principalmente para aplicaciones de bajo costo, se han enfocado en detectar las esquinas de los ojos [55, 56] en regiones extraídas de la imagen a muy baja resolución. Por el contrario, en la aplicación estudiada en este trabajo, los videos contienen únicamente la región ocular, tomada a una resolución muy alta mediante zoom sobre la zona. Esto provoca que la apariencia del ojo y de sus contornos varíe significativamente entre individuos o ante pequeños movimientos de los párpados. De esta forma, el objetivo de estimar la posición exacta de la esquina del ojo a partir de su apariencia se vuelve desafiante.

Recientemente se observa una marcada tendencia a reemplazar el uso de métodos tradicionales de procesamiento de imágenes o de inteligencia artificial por métodos basados en aprendizaje profundo. Específicamente, por métodos basados en CNNs para la tarea de detección y seguimiento de puntos de referencia facial [57, 58, 59].

4.2. Métodos propuestos

4.2.1. Cascada de CNN

Se diseñó una cascada de CNNs para la estimación precisa de las coordenadas de las esquinas del ojo en imágenes de alta resolución.

Para esto, se implementó una red de estimación gruesa, la cual realiza una primera predicción a partir de la imagen completa. Posteriormente, una red de refinamiento, la cual recibe una región mas pequeña extraída a partir de la primera y realiza una estimación fina de la coordenada del punto.

4.2.2. Seguimiento de regiones

Se adicionó información temporal para incrementar la precisión del algoritmo y disminuir el tiempo computacional. Para esto se adoptó un enfoque mixto donde una vez detectada la esquina del ojo, la red de regresión gruesa es reemplazada por un algoritmo de seguimiento estándar.

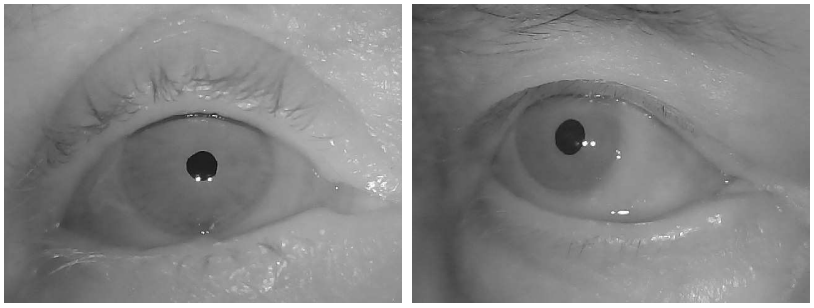


Figura 4.1: Imágenes oculares obtenidas con el dispositivo Oscann.

Se utilizó el filtro de correlación kernelizado (KFC) implementado en la librería OpenCV3 [60] como algoritmo de seguimiento. El mismo se inicializa con la predicción de la cascada de CNNs y luego, la primera CCN es reemplazada por la actualización del algoritmo de seguimiento a lo largo del video.

Para estabilizar el punto estimado a lo largo del tiempo, se agregó un filtro off-line a las predicciones de cada video [61].

4.2.3. Datos

Se creó una base de datos según las especificaciones requeridas para aplicaciones de diagnóstico de enfermedades neurológicas.

Específicamente, se grabaron 18 videos de pacientes diferentes utilizando el dispositivo Oscann [62]. El mismo cuenta con una cámara infraroja con resolución de imagen de 640x480 píxeles. Luego, se seleccionaron 15 cuadros por cada video como los que se muestran en la Figura 4.1.

Por último, las coordenadas (en píxeles) de la esquina del ojo fueron cuidadosamente etiquetadas a mano para cada imagen seleccionada.

4.2.4. Pre-procesamiento

Debido a la poca cantidad de datos disponibles, se realizó un complejo proceso de aumentación de datos, aplicando los siguientes métodos de forma aleatoria:

- Rotaciones ($\pm$ 40 grados)
- Suavizado Gaussiano.
- Variaciones de brillo.
- Transformaciones afin con matrices aleatorias.
- Deformaciones elásticas [63].

Para el entrenamiento de la CNN de refinamiento, se extrajeron múltiples regiones por cada imagen, aplicando desplazamientos aleatorios a la coordenada de referencia. Luego se aplicaron las mismas transformaciones previamente listadas. La Figura 4.2 muestra cuatro regiones diferentes generadas a partir de una única imagen. Se observa cómo la apariencia y la posición de la esquina dentro de la imagen varían considerablemente.

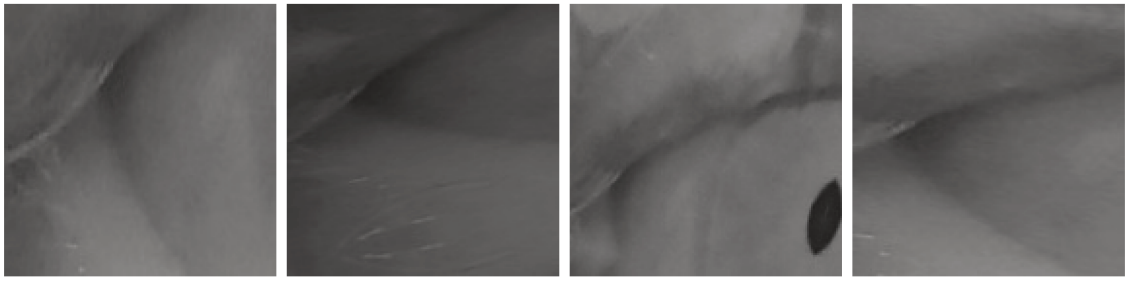


Figura 4.2: Ejemplos de regiones generadas mediante técnicas de aumentación.

Las mismas transformaciones se aplicaron a las coordenadas de referencia. Finalmente se normalizaron las coordenadas (lx,ly) entre -1 y 1 usando:

$$lxn = \frac{lx - 0{,}5 \cdot Ian}{0{,}5 \cdot Ian} \quad lyn = \frac{ly - 0{,}5 \cdot Ial}{0{,}5 \cdot Ial}, \tag{4.1}$$

donde Ian y Ial son el ancho y el alto respectivamente.

4.3. Experimentos y resultados

En la Figura 4.3 se muestra un diagrama del procedimiento de regresión aplicado a una imagen de ejemplo. Se observa la predicción inicial de la coordenada (en color azul), la extracción de la región de interés alrededor de la misma y el refinamiento efectuado por segunda red (en color rojo). En la sección IV del Anexo F se realiza una evaluación cuantitativa de estos resultados.

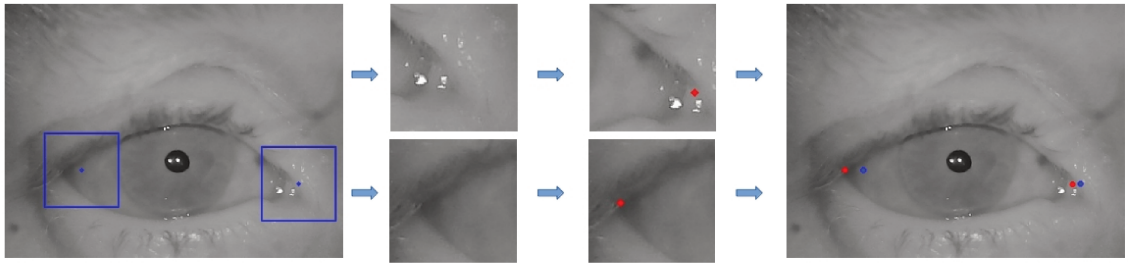


Figura 4.3: Estimación y refinamiento de las coordenadas de la esquina del ojo.

La estabilidad de la predicción en los videos se evaluó de manera subjetiva. Se observó que el filtro eliminó el efecto de temblor en la coordenada estimada. Esta mejora puede observarse en los videos disponibles a través del siguiente link: *https://agostinal.github.io/Corner-detector-project*.

Capítulo 5

Conclusiones

En esta tesis se desarrollaron métodos computacionales para el análisis de imágenes médicas basados en aprendizaje profundo, que resultaron en nuevos aportes metodológicos en tres contextos de aplicación diferentes.

En primer lugar, se demostró que los auto-codificadores para reducción de ruido pueden usarse para convertir segmentaciones erróneas de distintos órganos o estructuras en máscaras anatómicamente plausibles. El método desarrollado funciona como un paso independiente de post-procesamiento, que permite incorporar restricciones anatómicas a métodos arbitrarios de segmentación. Se realizaron experimentos para tareas de segmentación anatómica binaria y multi-clase de imágenes de rayos-X y RM indicando que el método propuesto funciona para distintas estructuras anatómicas y modalidades de imágenes. Además, Post-DAE no utiliza la información de intensidad. Por lo tanto, puede ser entrenado con máscaras independientes que hayan sido anotadas para otra modalidad de imagen, haciendo al método robusto a los corrimientos de dominio. Post-DAE puede implementarse fácilmente, es rápido al momento de la inferencia, es independiente de la modalidad de imagen y del método original de segmentación. Además se realizó un análisis sobre su comportamiento frente a estructuras con alguna patología, y pudo comprobarse que el método tiende a reconstruir la estructura anatómica original. Esto constituye, al mismo tiempo, una ventaja y una limitación del enfoque propuesto ya que Post-DAE transforma las máscaras de segmentación para que luzcan similares a las anatómicamente plausibles utilizadas en el entrenamiento. Estos son hechos importantes que deben ser considerados cuando se diseñan algoritmos de segmentación que incluyan Post-DAE como etapa de post-procesamiento. Lo mismo ocurre con problemas diferentes a la segmentación anatómica. En escenarios como la segmentación de lesiones o tumores cerebrales, donde la forma y la topología no son regulares, la aplicabilidad de Post-DAE puede ser limitada.

También se abordó el problema de estimación de incertidumbre aplicado a la segmentación de imágenes médicas. Para esto se estudiaron dos enfoques diferentes. En primer lugar se diseñó un método de ensamble ortogonal de redes. Este consiste en un esquema novedoso de entrenamiento de ensambles que aporta diversidad entre sus modelos al imponer explícitamente restricciones de ortogonalidad mediante la incorporación de un término de regularización. Se realizaron experimentos para tareas de segmentación binaria y multi-clase de imágenes de RM y se demostró que además de mejorar el desempeño en segmentación de los ensambles, el término propuesto reduce los errores de calibración, logrando modelos más precisos y confiables.

En segundo lugar, se siguió un enfoque alternativo y se abordó el problema de calibración promoviendo directamente la *auto-conciencia* de las DNNs referida a la incertidumbre de sus predicciones. Para esto se diseñó un término de regularización que penaliza la baja entropía de las salidas de la red únicamente para los píxeles mal clasificados, lo que acerca su estimación a posteriori hacia una distribución uniforme. Se realizaron experimentos con dos bases de datos populares, funciones de pérdida de segmentación y arquitecturas de redes demostrando que el método propuesto supera a la literatura en tareas de segmentación y calibración, mejorando la estimación de incerteza de los modelos entrenados con las funciones de pérdida clásicas de forma sencilla y efectiva.

Por último, con el objetivo de contribuir en la mejora de los algoritmos actuales de seguimiento de ojos para diagnóstico de enfermedades neurológicas, se diseñó e implementó un modelo para la regresión de las coordenadas de las esquina del ojo. El modelo fue diseñado para esta aplicación en particular por lo que fue necesario recolectar y etiquetar una base de datos completa. Se implementó una CNN en cascada para regresión con refinamiento de coordenadas de puntos de referencia en imágenes estáticas. Se agregó información temporal tanto utilizando un enfoque mixto de seguimiento de regiones a lo largo del video como implementando un filtro de suavizado temporal a la estimación final. Los resultados experimentales confirmaron la eficacia del método propuesto en diferentes videos. Con estos resultados prometedores, se espera aportar información valiosa para, en una segunda etapa, ser capaz de medir cambios en la posición de la cabeza durante experimentos clínicos en pacientes.

Capítulo 6

Publicaciones

Los resultados obtenidos durante la realización de la presente tesis fueron publicados en:

Revistas:

1. Larrazabal, A. J., Martínez, C. E., Glocker B. and Ferrante, E.. "Post-DAE: Anatomically Plausible Segmentation via Post-Processing with Denoising Autoencoders", *IEEE Transactions on Medical Imaging*, 2020, 39(12), 3813-3820.

2. Larrazabal, A. J., García Cena C.E., and Martínez C.E. "Video-oculography eye tracking towards clinical applications: A review."*Computers in biology and medicine*, 108 (2019): 57-66.

Congresos:

1. Larrazabal A. J., Martínez C., Dolz J*., Ferrante E*. "Maximum Entropy on Erroneous Predictions (MEEP): Improving model calibration for medical image segmentation". En revisión

2. Larrazabal A.J., Martínez C., Dolz J., Ferrante E. "Orthogonal Ensemble Networks for Biomedical Image Segmentation". En: de Bruijne M. et al. (eds) *International Conference on Medical Image Computing and Computer Assisted Intervention* - MICCAI 2021. Lecture Notes in Computer Science, vol 12903, pp. 594-603. Springer, Cham, 2021.

3. Larrazabal, A. J., Martinez, C., and Ferrante, E. "Anatomical priors for image segmentation via post-processing with denoising autoencoders". En *International Conference on Medical Image Computing and Computer-Assisted Intervention* - MICCAI 2019. Lecture Notes in Computer Science, vol 11769, pp. 585-593. Springer, Cham, 2019.

4. Larrazabal, A. J., García Cena C.E., and Martínez C.E. "Eye corners tracking for head movement estimation". En *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. IEEE, pp. 53-58, 2019.

Adicionalmente, en una línea de trabajo relacionada con la temática de esta tesis, se ha presentado un artículo que avanza sobre el estudio de desbalance de género en corpus de imágenes médicas y el impacto que podría tener en el entrenamiento de sistemas de asistencia al diagnóstico:

1. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. "Gender imbalance in medical imaging datasets produces biased classifiers for computer -aided diagnosis."*Proceedings of the National Academy of Sciences*, 117(23), 12592-12594.

Anexos

Referido a los artículos incluidos en los Anexos A, B,C D, E y F.

- Larrazabal, A. J., Martínez, C. E. and Ferrante, E.. "Anatomical Priors for Image Segmentation via Post-Processing with Denoising Autoencoders", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 585-593, 2019.
- Larrazabal, A. J., Martínez, C. E., Glocker B. and Ferrante, E.. "Post-DAE: Anatomically Plausible Segmentation via Post-Processing with Denoising Autoencoders", *IEEE Transactions on Medical Imaging*, 39(12), 3813-3820.
- Larrazabal, A. J., Martínez, C. E., Dolz J. and Ferrante, E.. "Orthogonal Ensemble Networks for Biomedical Image Segmentation", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.
- Larrazabal A.J., Martínez C. E., Dolz J*., Ferrante E*. "Maximum Entropy on Erroneous Predictions (MEEP): Improving model calibration for medical image segmentation". En revisión
- Larrazabal, A. J., García Cena, C. E. and Martínez, C. E.. "Video-oculography eye tracking towards clinical applications: A review", *Computers in biology and medicine*, 108 (2019), 57-66.
- Larrazabal, A. J., García Cena, C. E. and Martínez, C. E. "Eye corners tracking for head movement estimation", *IEEE International Work Conference on Bioinspired Intelligence*, 2019

la tesista declara haber contribuido en el diseño, implementación y evaluación de los algoritmos y experimentos realizados para obtener los resultados que allí se presentan, bajo la guía y supervisión del director Dr. C. Martínez, la co-directora Dra. Cecilia García Cena y el Dr. E. Ferrante. En cuanto a la escritura de los artículos, la tesista ha sido la autora principal, guiada por sugerencias y revisiones del director y coautores que en cada artículo se indican. La firma del Director avala esta declaración.

Dr. C. Martínez

Anexo A

Anatomical Priors for Image Segmentation via Post-Processing with Denoising Autoencoders

Anatomical Priors for Image Segmentation via Post-Processing with Denoising Autoencoders

Agostina J. Larrazabal, Cesar Martinez, Enzo Ferrante

Research institute for signals, systems and computational intelligence, sinc(i),
FICH-UNL / CONICET, Santa Fe, Argentina

Abstract. Deep convolutional neural networks (CNN) proved to be highly accurate to perform anatomical segmentation of medical images. However, some of the most popular CNN architectures for image segmentation still rely on post-processing strategies (e.g. Conditional Random Fields) to incorporate connectivity constraints into the resulting masks. These post-processing steps are based on the assumption that objects are usually continuous and therefore nearby pixels should be assigned the same object label. Even if it is a valid assumption in general, these methods do not offer a straightforward way to incorporate more complex priors like convexity or arbitrary shape restrictions.
In this work we propose Post-DAE, a post-processing method based on denoising autoencoders (DAE) trained using only segmentation masks. We learn a low-dimensional space of anatomically plausible segmentations, and use it as a post-processing step to impose shape constraints on the resulting masks obtained with arbitrary segmentation methods. Our approach is independent of image modality and intensity information since it employs only segmentation masks for training. This enables the use of anatomical segmentations that do not need to be paired with intensity images, making the approach very flexible. Our experimental results on anatomical segmentation of X-ray images show that Post-DAE can improve the quality of noisy and incorrect segmentation masks obtained with a variety of standard methods, by bringing them back to a feasible space, with almost no extra computational time.

Keywords: anatomical segmentation, autoencoders, convolutional neural networks, learning representations, post-processing

1 Introduction

Segmentation of anatomical structures is a fundamental task for biomedical image analysis. It constitutes the first step in several medical procedures such as shape analysis for population studies, computed assisted diagnosis and automatic radiotherapy planning, among many others. The accuracy and anatomical plausibility of these segmentations is therefore of paramount importance, since it will necessarily influence the overall quality of such procedures.

During the last years, convolutional neural networks (CNNs) proved to be highly accurate to perform segmentation in biomedical images [1–3]. One of the

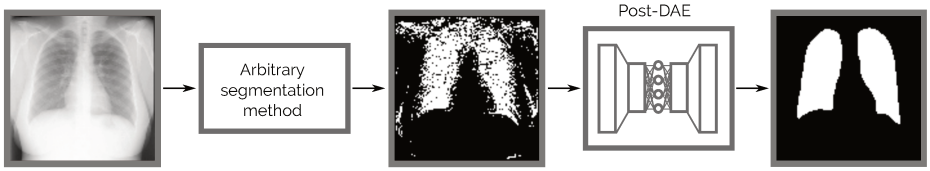


Fig. 1: Post-DAE works as a post-processing step and improves the anatomical plausability of segmentation masks obtained with arbitrary methods.

tricks that enables the use of CNNs in large images (by reducing the number of learned parameters) is known as parameter sharing scheme. The assumption behind this idea is that, at every layer, shared parameters are used to learn new representations of the input data along the whole image. These parameters (also referred as weights or kernels) are successively convoluted with the input data resulting in more abstract representations. This trick is especially useful for tasks like image classification, where invariance to translation is a desired property since objects may appear in any location. However, in case of anatomical structures in medical images where their location tend to be highly regular, this property leads to incorrect predictions in areas with similar intensities when enough contextual information is not considered. Shape and topology tend also to be preserved in anatomical images of the same type. However, as discussed in [4], the pixel-level predictions of most CNN architectures are not designed to account for higher-order topological properties.

Before the advent of CNNs, other classical learning based segmentation methods were popular for this task (e.g. Random Forest (RF) [5]), some of which are still being used specially when the amount of annotated data is not enough to train deep CNNs. The pixel-level predictions of these approaches are also influenced by image patches of fixed size. In these cases, handcrafted features are extracted from image patches and used to train a classifier, which predicts the class corresponding to the central pixel in that patch. These methods suffer from the same limitations related to the lack of shape and topological information discussed before.

In this work, we introduce Post-DAE (post-processing with denoising autoencoders), a post-processing method which produces anatomically plausible segmentations by improving pixel-level predictions coming from arbitrary classifiers (e.g. CNNs or RF), incorporating shape and topological priors. We employ Denoising Autoencoders (DAE) to learn compact and non-linear representations of anatomical structures, using only segmentation masks. This model is then applied as a post-processing method for image segmentation, bringing arbitrary and potentially erroneous segmentation masks into an anatomically plausible space (see Figure 1).

Contributions. Our contributions are 3-fold: (i) we show, for the first time, that DAE can be used as an independent post-processing step to correct problematic and non-anatomically plausible masks produced by arbitrary segmentation

methods; (ii) we design a method that can be trained using segmentation-only datasets or anatomical masks coming from arbitrary image modalities, since the DAE is trained using only segmentation masks, and no intensity information is required during learning; (iii) we validate Post-DAE in the context of lung segmentation in X-ray images, bench-marking with other classical post-processing method and showing its robustness by improving segmentation masks coming from both, CNN and RF-based classifiers.

Related works. One popular strategy to incorporate prior knowledge about shape and topology into medical image segmentation is to modify the loss used to train the model. The work of [4] incorporates high-order regularization through a topology aware loss function. The main disadvantage is that such loss function is constructed ad-hoc for every dataset, requiring the user to manually specify the topological relations between the semantic classes through a topological validity table. More similar to our work are those by [6, 7], where an autoencoder (AE) is used to learn lower dimensional representations of image anatomy. The AE is used to define a loss term that imposes anatomical constraints during training. The main disadvantage of these approaches is that they can only be used during training of CNN architectures. Other methods like RF-based segmentation can not be improved through this technique. On the contrary, our method post-processes arbitrary segmentation masks. Therefore, it can be used to improve results obtained with any segmentation method, even those methods which do not rely on an explicit training phase (e.g. level-sets methods).

Post-processing methods have also been considered in the literature. In [3], the output CNN scores are considered as unary potentials of a Markov random field (MRF) energy minimization problem, where spatial homogeneity is propagated through pairwise relations. Similarly, [2] uses a fully connected conditional random field (CRF) as post-processing step. However, as stated by [2], finding a global set of parameters for the graphical models which can consistently improve the segmentation of all classes remains a challenging problem. Moreover, these methods do not incorporate shape priors. Instead, they are based on the assumption that objects are usually continuous and therefore nearby pixels (or pixels with similar appearence) should be assigned the same object label. Conversely, our post-processing method makes use of a DAE to impose shape priors, transforming any segmentation mask into an anatomically plausible one.

2 Anatomical Priors for Image Segmentation via Post-Processing with DAE

Problem statement. Given a dataset of unpaired anatomical segmentation masks $\mathcal{D}_{\mathcal{A}} = \{S_i^A\}_{0 \leq i \leq |\mathcal{D}_{\mathcal{A}}|}$ (unpaired in the sense that no corresponding intensity image associated to the segmentation mask is required) we aim at learning a model that can bring segmentations $\mathcal{D}_{\mathcal{P}} = \{S_i^P\}_{0 \leq i \leq |\mathcal{D}_{\mathcal{P}}|}$ predicted by arbitrary classifiers P into an anatomically feasible space. We stress the fact that our method works as a post-processing step in the space of segmentations, mak-

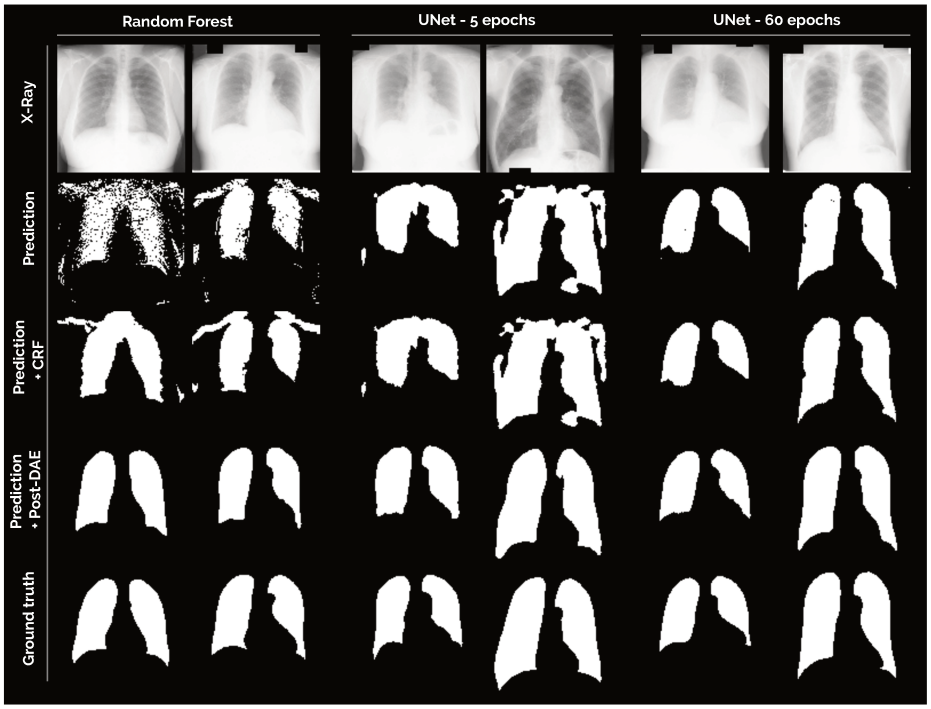


Fig. 2: Predictions obtained with three different methods: Random Forest, UNet trained for 5 epochs and until convergence. Rows from top to bottom: (i) X-Ray image; (ii) original mask obtained with the corresponding method; (iii) mask post-processed with a fully connected CRF; (iv) mask post-processed with the proposed Post-DAE method; and (v) ground-truth.

ing it independent of the predictor, image intensities and modality. We employ denoising autoencoders (DAE) to learn such model.

Denoising autoencoders. DAEs are neural networks designed to reconstruct a clean input from a corrupted version of it [8]. In our case, they will be used to reconstruct anatomically plausible segmentation masks from corrupted or erroneous ones. The standard architecture for an autoencoder follows an encoder-decoder scheme (see the Sup. Mat. for a detailed description of the architecture used in this work). The encoder $f_{enc}(S_i)$ is a mapping that transforms the input into a hidden representation h. In our case, it consists of successive non-linearities, pooling and convolutional layers, with a final fully connected layer that concentrates all information into a low dimensional code h. This code is then feed into the decoder $f_{dec}(h)$, which maps it back to the original input dimensions through a series of up-convolutions and non-linearities. The output of $f_{dec}(h)$ has the same size than the input S_i.

The model is called *denoising* autoenconder because a degradation function ϕ is used to degrade the ground-truth segmentation masks, producing noisy segmentations $\hat{S}_i = \phi(S_i)$ used for training. The model is trained to minimize the reconstruction error measured by a loss function based on the Dice coefficient (DSC), a metric used to compare the quality of predicted segmentations with respect to the ground-truth (we refer the reader to [9] for a complete description of the Dice loss):

$$\mathcal{L}_{DAE}(S_i) = DSC(S_i, f_{dec}(f_{enc}(\phi(S_i))). \tag{1}$$

The dimensionality of the learned representation $h = f_{enc}(S_i)$ is much lower than the input, producing a bottleneck effect which forces the code h to retain as much information as possible about the input. In that way, minimizing the reconstruction error amounts to maximizing a lower bound on the mutual information between input S_i and the learnt representation h [8].

Mask degradation strategy. The masks used to train the DAE were artificially degraded during training to simulate erroneous segmentations. To this end, we randomly apply the following degradation functions $\phi(S_i)$ to the ground truth masks S_i: (i) addition and removal of random geometric shapes (circles, ellipses, lines and rectangles) to simulate over and under segmentations; (ii) morphological operations (e.g. erosion, dilation, etc) with variable kernels to perform more subtle mask modifications and (iii) random swapping of foreground-background labels in the pixels close to the mask borders.
Note that, even if the proposed degradation strategy does not represent the distribution of masks generated by a particular classifier, it worked well in practise when post-processing both, random forest and U-Net masks. However, better results could be obtained if we know the classifier beforehand, by feeding the predicted segmentation masks while training the DAE.

Post-processing with DAEs. The proposed method is rooted in the so-called manifold assumption [10], which states that natural high dimensional data (like anatomical segmentation masks) concentrate close to a non-linear low-dimensional manifold. We learn such low-dimensional anatomically plausible manifold using the aforementioned DAE. Then, given a segmentation mask S_i^P obtained with an arbitrary predictor P (e.g. CNN or RF), we project it into that manifold using f_{enc} and reconstruct the corresponding anatomically feasible mask with f_{dec}. Unlike other methods like [6, 7] which incorporate the anatomical priors while training the segmentation network, we choose to make it a post-processing step. In that way, we achieve independence with respect to the initial predictor, and enable improvement for arbitrary segmentation methods.

Our hypothesis (empirically validated by the following experiments) is that those masks which are far from the anatomical space, will be mapped to a similar, but anatomically plausible segmentation. Meanwhile, masks which are anatomically correct, will be mapped to themselves, incurring in almost no modification.

3 Experiments and Discussion

Database description. We benchmark the proposed method in the context of lung segmentation in X-Ray images, using the Japanese Society of Radiological Technology (JSRT) database [11]. JSRT is a public database containing 247 PA chest X-ray images with expert segmentation masks, of 2048x2048 pixels and isotropic spacing of 0.175 mm/pixel, which are downsampled to 1024x1024 in our experiments. Lungs present high variability among subjects, making the representation learning task especially challenging. Note that we did not perform any pre-processing regarding image alignment. We divide the database in 3 folds considering 70% for training, 10% for validation and 20% for testing. The same folds were used to train the U-Net, Random Forest and Post-DAE methods.

Post-processing with DAE. Post-DAE receives a 1024x1024 binary segmentation as input. The network was trained to minimize the Dice loss function using Adam Optimizer. We employed learning rate 0.0001; batch size of 15 and 150 epochs.

Post-processing with CRF. We compare Post-DAE with the SOA post-processing method based on a fully connected CRF [12]. The CRF is used to impose connectivity constraints to a given segmentation, based on the assumption that objects are usually continuous and nearby pixels with similar appearance should be assigned the same object label. We use an efficient implementation of a dense CRF[1] that can handle large pixel neighbourhoods in reasonable inference times. Differently from our method which uses only binary segmentations for post-processing, the CRF incorporates intensity information from the original images. Therefore, it has to be re-adjusted depending on the image properties of every dataset. Instead, our method is trained once and can be used independently of the image propierties. Note that we do not compare Post-DAE with other methods like [6, 7] which incorporate anatomical priors while training the segmentation method itself, since these are not post-processing strategies.

Baseline segmentation methods. We train two different models which produce segmentation masks of various qualities to benchmark our post-processing method. The first model is a CNN based on UNet architecture [1] (see the Sup. Mat. for a detailed description of the architecture and the training parameters such as optimizer, learning rate, etc.). The UNet was implemented in Keras and trained in GPU using a Dice loss function. To evaluate the effect of Post-DAE in different masks, we save the UNet model every 5 epochs during training, and predict segmentation masks for the test fold using all these models. The second method is a RF classifier trained using intensity and texture features. We

[1] We used the public implementation available at `https://github.com/lucasb-eyer/pydensecrf` with Potts compatibility function and hand-tuned parameters $\theta_\alpha = 17$, $\theta_\beta = 3$, $\theta_\gamma = 3$ chosen using the validation fold. See the implementation website for more details about the aforementioned parameters.

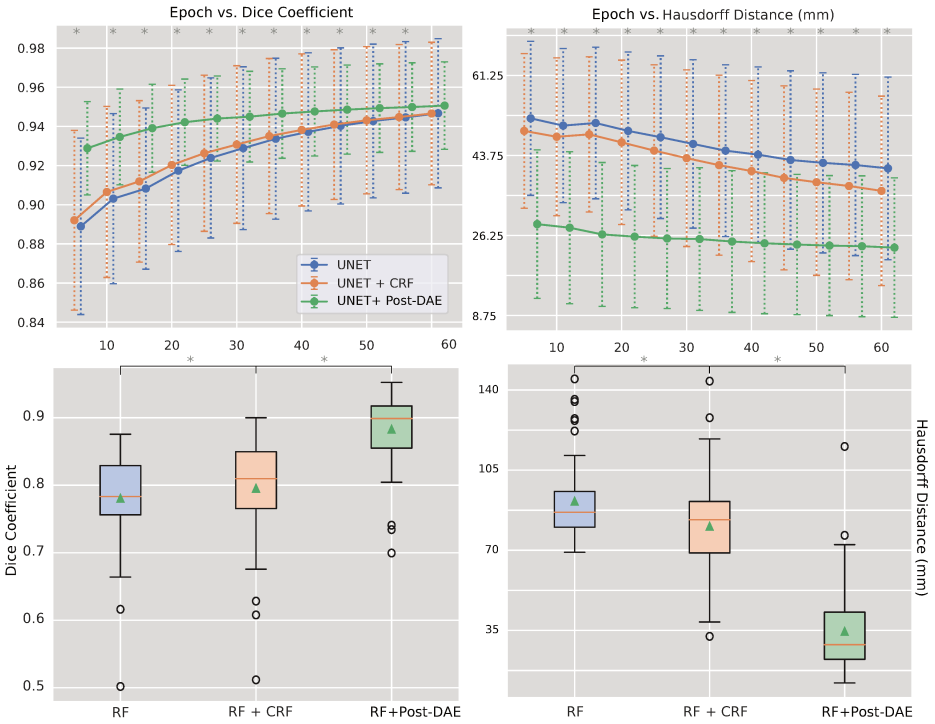


Fig. 3: Quantitative evaluation of the proposed method. We compare Post-DAE with the classic fully-connected CRF [12] adopted as post-processing step by many segmentation methods like [2]. Top row shows mean and standard deviation for post-processing UNet predictions on the test fold at different training stages (from 5 epochs to convergence). We use Dice coefficient and Hausdorff distance to measure the segmentation quality. Bottom row show results for post-processing the Random Forest predictions. The symbol $*$ indicates that Post-DAE outperforms the other methods (no post-processing and CRF) with statistical significance (p-value < 0.05 according to Wilcoxon test). The green triangle in the box indicates the mean value.

used Haralick [13] features which are based on gray level co-ocurrency in image patches. We adopted a public implementation available online with default parameters[2] which produces acceptable segmentation masks.

Results and discussion. Figure 2 shows some visual examples while Figure 3 summarizes the quantitative results (see the video in the Sup. Mat. for more visual results). Both figures show the consistent improvement that can be obtained using Post-DAE as a post-processing step, specially in low quality seg-

[2] The source code and a complete description of the method is publicly available online at: `https://github.com/dgriffiths3/ml_segmentation`

mentation masks like those obtained by the RF model and the UNet trained for only 5 epochs. In these cases, substantial improvements are obtained in terms of Dice coefficient and Hausdorff distance, by bringing the erroneous segmentation masks into an anatomically feasible space. In case of segmentations that are already of good quality (like the UNet trained until convergence), the post-processing significantly improves the Hausdorff distance, by erasing spurious segmentations (holes in the lung and small isolated blobs) that remain even in well trained models. When compared with CRF post-processing, Post-DAE significantly outperforms the baseline in the context of anatomical segmentation. In terms of running time, the CRF model takes 1.3 seconds in a Intel i7-7700 CPU, while Post-DAE takes 0.7 seconds in a Titan Xp GPU.

One of the limitations of Post-DAE is related to data regularity. In case of anatomical structures like lung, heart or liver, even if we found high inter-subject variability, the segmentation masks are somehow uniform in terms of shape and topology. Even pathological organs tend to have similar structure, which can be well-encoded by the DAE (specially if pathological cases are seen during training). However, in other cases like brain lesions or tumors where shape is not that regular, it is not clear how Post-DAE would perform. This case lies out of the scope of this paper, but will be explored as future work.

Conclusions and future works. In this work we have showed, for the first time in the MIC community, that autoencoders can be used as an independent post-processing step to incorporate anatomical priors into arbitrary segmentation methods. Post-DAE can be easily implemented, is fast at inference, can cope with arbitrary shape priors and is independent of the image modality and segmentation method. In the future, we plan to extend this method to muti-class and volumetric segmentation cases (like anatomical segmentation in brain images).

4 Acknowledgments

EF is beneficiary of an AXA Research Fund grant. The authors gratefully acknowledge NVIDIA Corporation with the donation of the Titan Xp GPU used for this research, and the support of UNL (CAID-PIC-50420150100098LI) and ANPCyT (PICT 2016-0651).

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. of MICCAI. (2015)
2. Kamnitsas, K., et al.: Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis **36** (2017) 61 – 78
3. Shakeri, M., et al.: Sub-cortical brain structure segmentation using F-CNN's. In: Proc. of ISBI. (2016)
4. BenTaieb, A., Hamarneh, G.: Topology aware fully convolutional networks for histology gland segmentation. In: Proc. of MICCAI. (2016)

5. Breiman, L.: Random forests. Machine learning **45**(1) (2001) 5–32
6. Oktay, O., et al.: Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE TMI **37**(2) (2018) 384–395
7. Ravishankar, H., et al.: Learning and incorporating shape models for semantic segmentation. In: Proc. of MICCAI. (2017)
8. Vincent, P., et al.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. JMLR **11** (2010) 3371–3408
9. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proc. of Fourth International Conference on 3D Vision (3DV). (2016)
10. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning. MIT Press (2009)
11. Shiraishi, J., et al.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am Jour of Roent **174**(1) (2000) 71–74
12. Krähenbühl, P., Koltun, V.: Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In: Proc. of Nips. (2011)
13. Haralick, R.M., Shanmugam, K., et al.: Textural features for image classification. IEEE Transactions on systems, man, and cybernetics (6) (1973) 610–621

Anexo B

Post-DAE: Anatomically Plausible Segmentation via Post-Processing with Denoising Autoencoders

Post-DAE: Anatomically Plausible Segmentation via Post-Processing with Denoising Autoencoders

Agostina J Larrazabal, César Martínez, Ben Glocker and Enzo Ferrante

***Abstract*—We introduce Post-DAE, a post-processing method based on denoising autoencoders (DAE) to improve the anatomical plausibility of arbitrary biomedical image segmentation algorithms. Some of the most popular segmentation methods (e.g. based on convolutional neural networks or random forest classifiers) incorporate additional post-processing steps to ensure that the resulting masks fulfill expected connectivity constraints. These methods operate under the hypothesis that contiguous pixels with similar aspect should belong to the same class. Even if valid in general, this assumption does not consider more complex priors like topological restrictions or convexity, which cannot be easily incorporated into these methods.**

Post-DAE leverages the latest developments in manifold learning via denoising autoencoders. First, we learn a compact and non-linear embedding that represents the space of anatomically plausible segmentations. Then, given a segmentation mask obtained with an arbitrary method, we reconstruct its anatomically plausible version by projecting it onto the learnt manifold. The proposed method is trained using unpaired segmentation mask, what makes it independent of intensity information and image modality. We performed experiments in binary and multi-label segmentation of chest X-ray and cardiac magnetic resonance images. We show how erroneous and noisy segmentation masks can be improved using Post-DAE. With almost no additional computation cost, our method brings erroneous segmentations back to a feasible space.

***Index Terms*—anatomical segmentation, autoencoders, convolutional neural networks, learning representations, post-processing**

I. Introduction

ANATOMICAL segmentation is a fundamental task in medical image computing, which consists in associating pixels of a medical image with a given organ or anatomical structure. It constitutes an essential step in many imaging pipelines such as computer assisted diagnosis, morphometric analysis for population studies and radiotherapy planning. The correctness and anatomical plausibility of these results is thus of paramount importance, since it will directly influence the overall quality of subsequent analyses.

Article accepted for publication in IEEE Transactions on Medical Imaging.

A.J. Larrazabal, C. Martínez and E. Ferrante are with the Institute for Signals, Systems and Computational Intelligence, sinc(i) CONICET-UNL, Santa Fe, Argentina. (e-mails: alarrazabal@sinc.unl.edu.ar - cmartinez@sinc.unl.edu.ar, eferrante@sinc.unl.edu.ar). B. Glocker is with the Biomedical Image Analysis Group, Imperial College London, London, UK. (e-mail: b.glocker@imperial.ac.uk)

E. Ferrante is beneficiary of an AXA Research Fund grant. The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPUs used for this research, and the support of UNL (CAID-PIC-50420150100098LI, CAID-PIC-50220140100084LI) and ANPCyT (PICT 2016-0651, PICT 2018-03907).

Convolutional neural networks (CNNs) proved to perform biomedical image segmentation in a highly accurate way [1]–[3]. CNNs constitute a particular type of neural network specially suited for regularly structured data, like 2D or 3D images, where hierarchical representations of the input are learned using stacked convolutional layers. At every layer, shared parameters (also referred as weights or kernel) are used to learn new representations of the input image. This sharing scheme reduces the number of parameters that should be learnt and allows the use of CNNs in large images. Thanks to the inherently regular structure of the images, these parameters are successively convoluted with the input data resulting in more abstract representations. This trick is particularly helpful for tasks in which invariance to translation is an expected property, such as image classification. However, in medical images, where the location of the anatomical structures is often highly regular, this property may lead to incorrect predictions in regions with similar intensities when insufficient contextual information is considered. The organs observed in anatomical images tend to preserve shape and topology across patients. Nonetheless, the pixel-level predictions of most CNN architectures do not account for such higher-order topological properties, as discussed in [4].

Before the emergence of CNNs, other learning-based methods were popular for biomedical image segmentation (e.g. Random Forest (RF) [5]). When the amount of annotated data is small and insufficient for training deep CNNs, some of these classical methods are still in use. A popular strategy is to adopt patch-based methods, where handcrafted features are generated from image patches and then used to train a classifier. Such classifier will then make pixel-level predictions considering only the image area around the central pixel of the patch. This results in methods which are also agnostic to the global shape and topology of the anatomical structures.

In this work we introduce Post-DAE, a post-processing method which improves pixel-level predictions generated with arbitrary classifiers by incorporating shape and topological priors. We employ denoising autoencoders (DAEs) to learn compact and non-linear representations of anatomical structures, using only segmentation masks. The DAE is then used to bring potentially erroneous segmentation masks into an anatomically plausible space (see Figure 1).

Contributions. A preliminary version of this work was published in MICCAI 2019 [6]. In this extended version we provide additional experiments in the context of multi-class lung and heart segmentation of X-ray images, and left ventricle delineation in cardiac magnetic resonance (CMR) images.

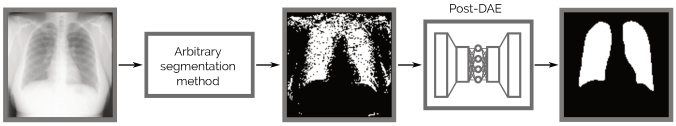


Fig. 1: Post-DAE worflow: the method is implemented as a post-processing step which maps arbitrary segmentation masks to anatomical plausible cases.

We also include a more complete and updated state-of-the-art section, a deeper analysis of how our method behaves in images with gross abnormalities and out-of-distribution cases, together with additional illustrations and extended discussion.

Our contributions can be summarized as follows: (i) we show that denoising autoencoders used as a post-processing step can improve the anatomical plausibility of unfeasible segmentation masks; (ii) we present results in the context of binary and multi-label segmentation of chest X-ray and CMR images, bench-marking with other classical post-processing method and showing the robustness of Post-DAE by improving segmentation masks coming from both, CNN and RF-based classifiers and (iii) we analyze the behaviour and limitations of our method when post-processing abnormal and out-of-distribution anatomical segmentation masks.

II. Related Work

Multiple alternatives have been proposed to incorporate prior knowledge in medical image segmentation (see [7] for a complete review). One popular strategy to integrate priors about shape and topology into learning based segmentation methods is to modify the loss function used to train the model. A topology aware loss function which incorporates high-order regularization was proposed in [4]. In this case, a manually defined topological validity table specifies the relation between the structures of interest. This constitutes a disadvantage since such loss function must be constructed ad-hoc for every dataset. More similar to our method are those by [8], [9], where compact anatomical represention are learnt by means of autoencoders. Such global representation is incorporated into the loss function and used to encourage anatomical plausibility into the predicted segmentation masks. Differently from our method designed to improve arbitrary segmentations, the main disadvantage of [8], [9] is that they are specifically tailored to be used when training a CNN model. Therefore, they cannot be used to improve results obtained with other segmentation approaches like RF or even level-sets methods, which do not rely on an explicit training phase.

An alternative simple but effective approach is to increase the receptive field of the network, i.e. the area of the input image that influences a single prediction. Even if this strategy does not incorporate an explicit shape prior, it allows the network to consider high-order interactions between distant image regions, learning to encode certain global features about shape and topology. In this regard, [2] proposed to increase the receptive field of the CNN by means of a dual path focusing on a wider low resolution area of the input image. This increases the contextual information provided to the network, but also augments the complexity of the segmentation model itself. Another approach to deal with the lack of spatial context in patch-based convolutional architectures is to augment the model including information about pixel location. In [10] the authors suggest that location information is a crucial discriminator in patch-based image segmentation, and show experimental results about the gain in performance when adding it explicitly to the description. In a more recent work [11], the authors propose the use of a spectral location parametrization specially adapted to brain volume coordinates to improve CNN-based image segmentation. However, albeit the fact that these strategies increase the accuracy of the resulting segmentation masks by incorporating contextual information, they do not incorporate explicit priors about shape and topology.

Alternative approaches implemented as post-processing methods have also been considered. Shakeri and co-workers [3] pose the problem as a discrete energy minimization problem, where CNN predictions are seen as unary potentials of a Markov random field (MRF) [12]. In this framework, pairwise relations are used to propagate spatial homogeneity. Following a similar idea, a fully connected conditional random field (CRF) is used in [2] as a post-processing step. However, as stated by [2], it is hard to find a global set of CRF parameters which can consistently improve the segmentation of all structures of interest. Moreover, there is no shape or topological prior incorporated in these models. Instead, these methods only operate under the hypothesis that pixels which are contiguous and exhibit similar aspect should belong to the same class. Even if valid in general, this assumption does not consider more complex priors like topological restrictions or convexity, which can be easily encoded in our post-processing methods.

Similar to the work of [13], our model can be trained using segmentation-only datasets which are not paired with image data (or data of the same image modality as the problem at hand). Our model is agnostic to image intensity (and thus insensitive to domain shift) and its formulation is much simpler than the one introduced in [13].

III. Anatomically Plausible Segmentation via Post-DAE

Given a dataset of anatomical segmentation masks (without paired intensity images) $\mathcal{D}_{\mathcal{A}} = \{S_i^A\}_{0 \leq i \leq |\mathcal{D}_{\mathcal{A}}|}$ we intend to learn a model that can bring segmentations $\mathcal{D}_{\mathcal{P}} = \{S_i^P\}_{0 \leq i \leq |\mathcal{D}_{\mathcal{P}}|}$ predicted by different classifiers P into an anatomically feasible space. We stress the fact that our method works as a post-processing step in the space of segmentations, making it independent of the predictor, image intensities and modality.

Denoising autoenconders (DAE) are neural networks designed to reconstruct a clean input from a corrupted version of it [14]. In this work, we propose to employ DAEs to recover anatomically plausible segmentation masks from corrupted or incorrect ones. The standard architecture for an autoencoder follows an encoder-decoder scheme (see the Supplementary Material for a detailed description of the architecture used in this work). The encoder $f_{enc}(S_i)$ is a mapping function

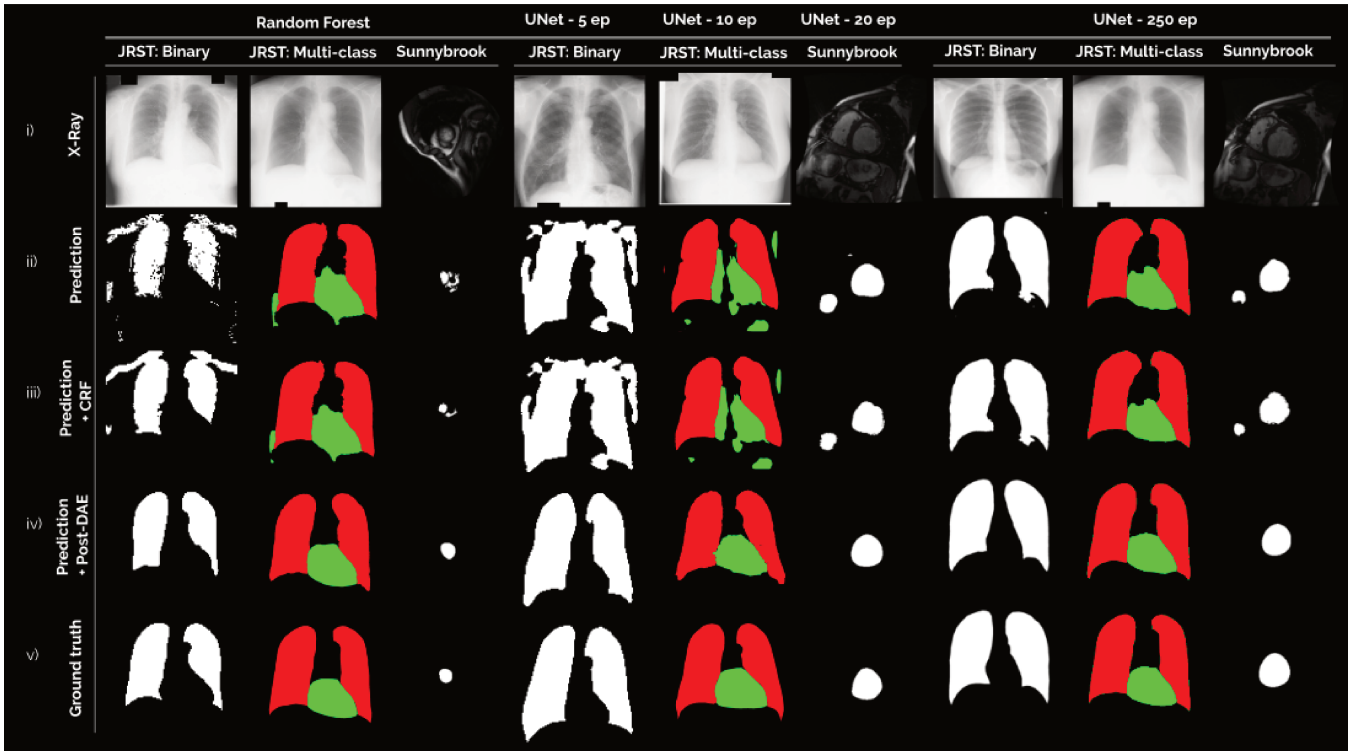


Fig. 2: Predictions obtained with segmentation methods of several qualities: random forest and UNet trained for different number of epochs. We include examples for both binary (white images) and multi-class (color images) segmentation. for: (i) X-Ray or CMR image; (ii) segmentation mask predicted by each baseline method; (iii) segmentation mask after post-processing with a CRF; (iv) segmentation mask after post-processing with our Post-DAE; and (v) ground-truth expert segmentations.

that turns the input into a lower dimensional hidden encoding h. In our implementation, $f_{enc}(S_i)$ is composed of stacked convolutions, non-linearities and pooling layers. At the end, a fully connected layer concentrates all information into a low dimensional code h. Then, the decoder $f_{dec}(h)$ maps this code back to the original input dimensions by means of successive non-linearities and up-convolutions.

The model is called *denosing* autoenconder due to the fact that it is trained with noisy segmentations $\hat{S}_i = \phi(S_i)$, which are obtained by degrading the ground-truth segmentation masks with a degradation function ϕ. To minimize the reconstruction error of the predicted segmentations with respect to the ground-truth, we train the DAE using a loss function based on the Dice coefficient (DSC) (for an exhaustive description of the Dice loss please see [15]):

$$\mathcal{L}_{DAE}(S_i) = DSC(S_i, f_{dec}(f_{enc}(\phi(S_i)))). \quad (1)$$

The learnt encoding $h = f_{enc}(S_i)$ is forced to retain as much information as possible about the input. This is due to the bottleneck effect produced by the reduced dimensionaliy of h. In this context, minimizing the reconstruction loss amounts to maximizing a lower bound on the mutual information between input S_i and the learnt representation h [14].

A. Mask degradation strategy.

We simulate corrupted segmentations to train the DAE by artificially degrading the ground truth segmentation masks S_i with the following random degradation functions $\phi(S_i)$[1]: (i) We simulate over and under segmentation by adding and removing random geometric shapes (including polygons, lines and ellipses); (ii) erosion, dilation and other morphological operations with variable kernels are applied to perform minor mask alterations; (iii) mask borders are modified by random swapping of foreground-background labels in the pixels close to the organ boundaries. In addition, data augmentation was performed by randomly resizing the original masks.

B. Post-processing with denoising autoencoders.

The proposed method is rooted in the so-called manifold assumption [17], which states that natural high dimensional data (like anatomical segmentation masks) concentrate close to a non-linear low-dimensional manifold. We use the DAE to learn such anatomically plausible manifold. Then, given a segmentation mask S_i^P produced with an arbitrary predictor P (e.g. CNN or RF), we project it onto that manifold using f_{enc} and reconstruct the corresponding anatomically feasible mask with f_{dec}. Different from other approaches like [8], [9] which incorporate the anatomical priors during the segmentation network training, our method is agnostic to the training process of the original predictor. Since it is conceived as a post-processing step, segmentation masks produced with

[1]Our code associated to Post-DAE, the degradation function and the UNet model is publicly available at the following Colab Python Notebook:https://colab.research.google.com/drive/1wLZLo81clNR_c-UJTpBV4fh-BMbByAU8?usp=sharing

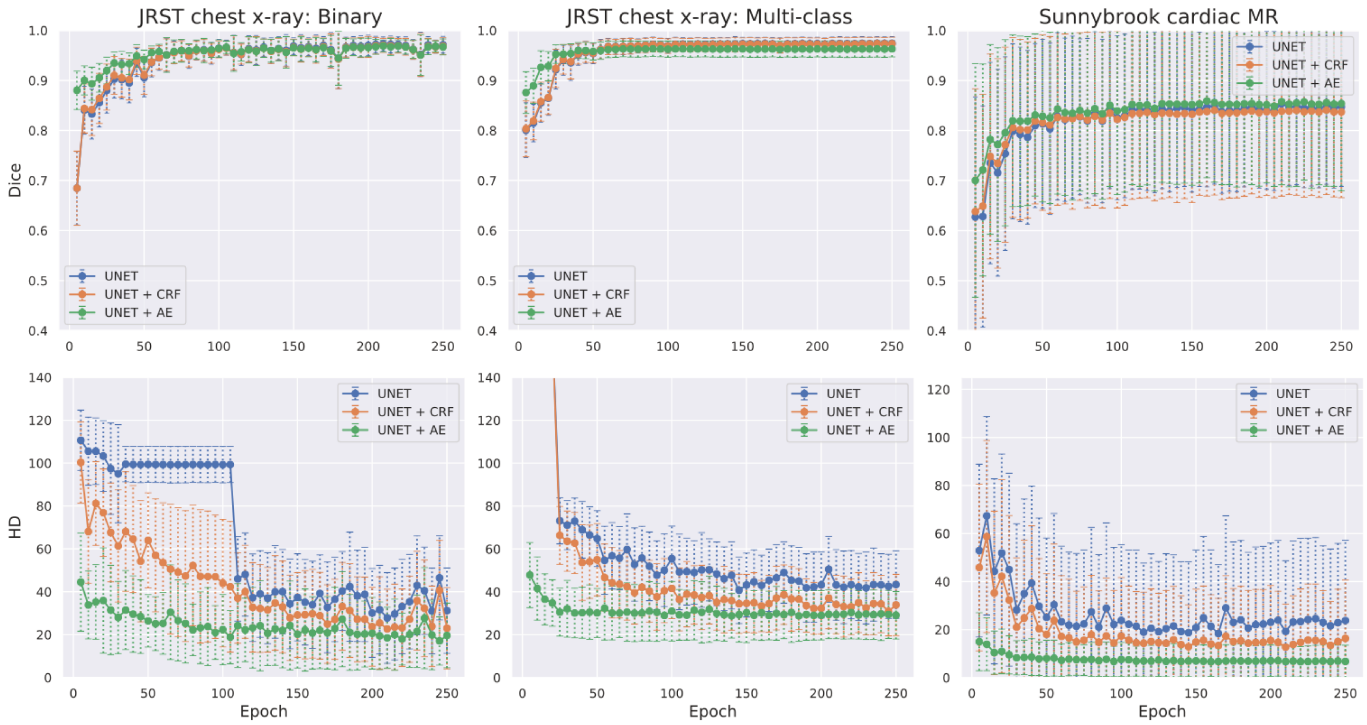


Fig. 3: Quantitative comparison between post-processing with Post-DAE and CRF [16]. We show mean and standard deviation for post-processing UNet predictions on the test fold at different training stages (from 5 epochs to convergence). First column shows the results for binary lung segmentation in JRST, the second one for multi-class (lung, heart) in JSRT and the third one for binary LV masks in Sunnybrook. We use Dice coefficient and Hausdorff distance (HD) to measure the segmentation quality. Note that initial low quality segmentations are improved both in terms of Dice and HD. For better initial segmentations, significant improvements in terms of HD are still obtained.

arbitrary methods can by improved using Post-DAE. Recent studies [18], [19] show that different autoencoders trained with healthy brain images (operating in the intensity domain) can be used to perform anomaly detection on pathological brain images, by just looking at the differences between the original pathological image and the one processed by the autoencoder. In the same spirit, our hypothesis (empirically validated with the experiments presented in the next section) is that those masks which are far from the anatomical space, will be mapped to a similar, but anatomically plausible segmentation. Meanwhile, masks which are anatomically correct, will suffer almost no modification, being mapped to themselves.

IV. Experimental setting

A. Database description.

We benchmark the proposed method in two different anatomical segmentation scenarios, including chest X-ray and cardiac magnetic resonance (CMR) images.

Chest X-ray dataset: in the context of lung and heart segmentation in X-Ray images, we used the Japanese Society of Radiological Technology (JSRT) database [20]. This is a public dataset with expert annotations composed of 247 PA thoracic X-ray images (2048x2048 pixels and spacing of 0.175mm x 0.175mm), which are downsampled to 1024x1024 in our experiments. Lungs and heart present high inter-subject variability, what makes the representation learning task especially challenging. We divide the database in 3 folds considering 70% for training, 10% for validation and 20% for testing. The same folds were used to train the U-Net, random forest and Post-DAE methods. We did not apply image alignment for pre-processing.

Cardiac MR dataset: We used images from a version of the Sunnybrook Cardiac Dataset (SCD) [21] publicly available at https://github.com/mshunshin/SegNetCMR. It includes 45 cine-MR images (every image composed of 6 to 12 short-axis (SAX) 2D slices) captured at end-systole (ES) and end-diastole (ED) time points, with corresponding segmentation masks of the left ventricle (LV). The image resolution is 256x256, covering a field of view of 320 mm x 320 mm. We partitioned the dataset using the originally suggested train/test partition scheme (taking 35 images from the training fold for validation).

B. Post-processing with CRF.

The proposed method is compared with a standard post-processing strategy based on a fully connected CRF [16]. This method operates under the hypothesis that pixels which are contiguous and exhibit similar aspect should belong to the same class. We use an efficient implementation of a

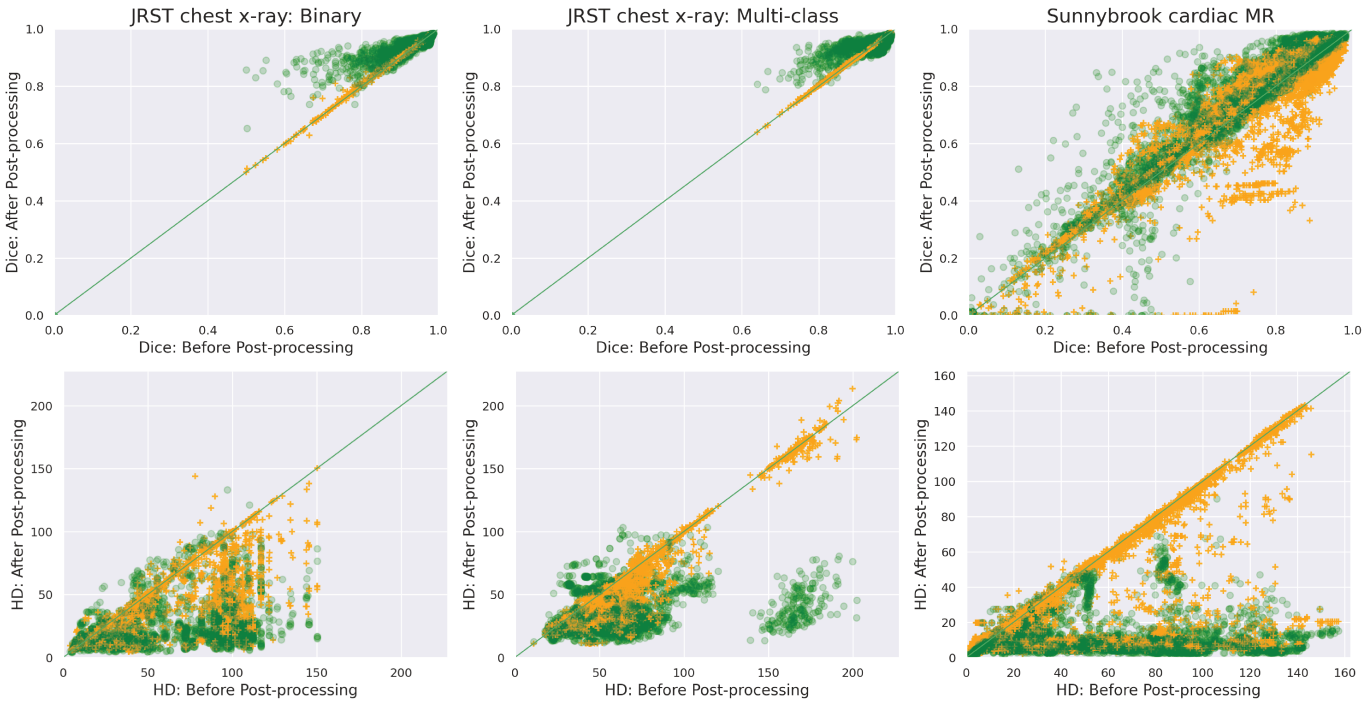


Fig. 4: Scatter plots comparing Dice (top) and HD (bottom) before and after post-processing with Post-DAE (green) and CRF (orange) for all samples in the previous study. We include segmentation masks generated with the UNet models trained from 5 to 250 epochs and random forest. First column shows the results for binary lung segmentation in JRST, the second one for multi-class (lung, heart) in JSRT and the third one for binary LV masks in Sunnybrook.

dense CRF. [2] Since the CRF formulation incorporates intensity information from the original images, the model parameters have to be re-adjusted whenever the image dataset is changed. In contrast, the proposed Post-DAE is agnostic to image intensity, and only needs to be trained once.

C. Training Post-DAE

Post-DAE is independent of the segmentation methods and it was trained separately from them. The model was implemented in Keras and trained for 150 epochs (the architecture and training details are included in the Supplementary Material). During training, we used the mask degradation strategy described in Section III-A. At test time, we took the segmentation masks generated by the baseline segmentation methods, and post-processed them by simply passing them through the DAE.

D. Baseline segmentation methods.

We trained binary and multi-class versions of two different segmentation models which produce masks of various qualities. For the X-ray images, we tackled binary (lungs vs background) and multi-class (lungs, heart, background) segmentation. For the CMR images, we focus on binary LV segmentation. The RF classifier was trained using intensity and texture features. For the binary segmentation, we adopted a public implementation available online with default parameters[3] which produces acceptable segmentation masks. It uses Haralick [22] features which are based on gray level co-ocurrency in image patches. For the multi-label segmentation, we apply a different implementation[4] which has a better performance for multi-label predictions and has been used in [23]. This RF variant leverages randomized offset boxes for calculating average intensity and intensity difference features efficiently via integral images. For the multi-label segmentation using RF we produced segmentations of various qualities by using different tree depths at test time. The second method is a CNN based on UNet architecture [1] (see the Supplementary Material for a detailed description of the architecture and the training parameters such as optimizer, learning rate, etc.). The UNet was implemented in Keras and trained in GPU using Dice loss function. To compare the effect of Post-DAE in different segmentation qualities, we save the UNet model every 5 epochs during training, and predict segmentation masks for the test fold using all these models. We compared post-processing with Post-DAE and CRF, reporting results for all these cases.

[2]We used the public implementation available at https://github.com/lucasb-eyer/pydensecrf with Potts compatibility function and hand-tuned parameters $\theta_\alpha = 17$, $\theta_\beta = 3$, $\theta_\gamma = 3$ for the X-ray images and $\theta_\alpha = 7$, $\theta_\beta = 3$, $\theta_\gamma = 3$ for the CMR images, chosen using the validation fold. See the website for more details about the aforementioned parameters.

[3]The source code and a complete description of the method is publicly available online at: https://github.com/dgriffiths3/ml_segmentation

[4]Publicly available at: https://github.com/biomedia-mira/oak2

TABLE I: Mean and standard deviation for post-processing random forest predictions. The numbers in bold indicates that Post-DAE outperforms the other methods (no post-processing and CRF) with statistical significance according to Wilcoxon test with Bonferroni correction.

Segmentations		JRST Chest X-ray						Sunnybrook Cardiac MR
		Multi-class					Binary	Binary
		Depth: 8	Depth: 12	Depth: 16	Depth: 20	Full model	Full model	Full model
Dice	RF	0.858 (0.042)	0.913 (0.033)	0.936 (0.029)	0.949 (0.029)	0.956 (0.028)	0.781 (0.070)	0.46 (0.24)
	RF + CRF	0.860 (0.042)	0.914 (0.032)	0.937 (0.029)	0.950 (0.029)	0.956 (0.027)	0.795 (0.074)	0.44 0.25
	RF + DAE	**0.922 (0.024)**	**0.943 (0.020)**	**0.948 (0.018)**	0.951 (0.018)	0.951 (0.019)	**0.865 (0.056)**	**0.47 (0.25)**
Hausdorff Distance (HD)	RF	102.26 (11.68)	96.00 (14.31)	88.70 (14.04)	77.45 (12.98)	72.28 (14.07)	91.41 (17.52)	27.73 (9.89)
	RF+CRF	101.17 (12.94)	92.97 (14.72)	81.26 (12.73)	74.51 (13.10)	67.29 (13.53))	80.45 (22.28)	26.87 (10.03)
	RF +DAE	**63.73 (11.85)**	**60.72 (12.20)**	**62.47 (15.16)**	**62.95 (16.84)**	**60.69 (14.12)**	**32.01 (18.44)**	**23.60 (9.88)**

V. Results and discussion

Figure 2 shows some visual examples while Table I and Figure 3 summarize the quantitative results obtained when post-processing segmentations produced by a RF classifier and a UNet. Our best results are in line with those obtained for other deep learning based state-of-the-art methods. For JSRT, recent works [24]–[27] report average Dice values for lung and heart ranging from 0.943 [27] to 0.965 [24]. For the Sunnybrook dataset, recent works [28]–[31] report average Dice ranging from 0.88 [28] to 0.93 [30] for LV segmentation. Both figures show the consistent improvement achieved when using Post-DAE as a post-processing step, specially in low quality segmentation masks like those obtained by the binary RF model, the multiclass RF considering incomplete tree depths and the UNet trained only for a few epochs. In these cases, substantial improvements are obtained in terms of Dice coefficient and Hausdorff distance (HD), by bringing the erroneous segmentation masks into an anatomically feasible space. In case of segmentations that are already of good quality (like multi-class RF or the UNet trained until convergence), Post-DAE significantly improves the HD, by erasing spurious segmentations that remain even in well trained models, like holes in the lung or small isolated blobs. When compared with CRF post-processing, Post-DAE significantly outperforms the baseline in the context of anatomical segmentation. In terms of running time, the CRF model takes 1.3 seconds while Post-DAE takes 0.76 seconds in a Intel i7-7700 CPU.

Scatter plots in Figure 4 show the change in terms of Dice (top) and HD (bottom) between initial segmentations before post-processing (x-axis) and after post-processing (y-axis), when comparing them with the ground-truth masks. In the Dice plots we observe how the green points tend to concentrate in the upper part of the diagonal, while orange crosses stick to it, indicating that Post-DAE improves the segmentations more than CRF. HD scatter plots should be read in the opposite way, i.e. the lower the better. Green points, corresponding to Post-DAE, concentrate in the bottom part while the orange crosses (CRF) tend to be over them, indicating that Post-DAE outperforms CRF also in terms of HD.

We include additional experiments in the Supplementary Material showing that Post-DAE can be trained with unpaired segmentation masks. These segmentation masks could be annotated on different image modalities, come from a different dataset with the same image modality or even from segmentation-only datasets. This experiment highlights the fact that our method does not require image intensity information for training/testing, making it robust to domain shift.

A. Out-of-distribution segmentation masks and limitations

In this section we analyze the behaviour and limitations of Post-DAE when post-processing masks which are out-of-distribution. In this context, out-of-distribution cases could appear mainly due to two reasons: erroneous segmentations or pathological images.

In the first case, erroneous masks may be generated by a segmentation method with low performance. See for example the masks in Figure 2, obtained with the RF model or the UNet trained for only 5 or 10 epochs. These cases are represented in the scatter plots depicted in Figure 4 by the points with low Dice or high Hausdorff before post-processing. Post-DAE clearly improves erroneous segmentations in this scenario, increasing the Dice after post processing and/or reducing its Hausdorff distance. This improvement is explained by the way Post-DAE was trained: we degraded the ground truth masks by introducing similar errors (see Section III-A) and force the DAE to reconstruct anatomically plausible segmentation masks. Since the test images are anatomically plausible as well, mapping erroneous segmentations to realistic ones improves the results.

The second scenario is related to abnormal cases. Big occlusions or deformed organs, possibly due to manifestations

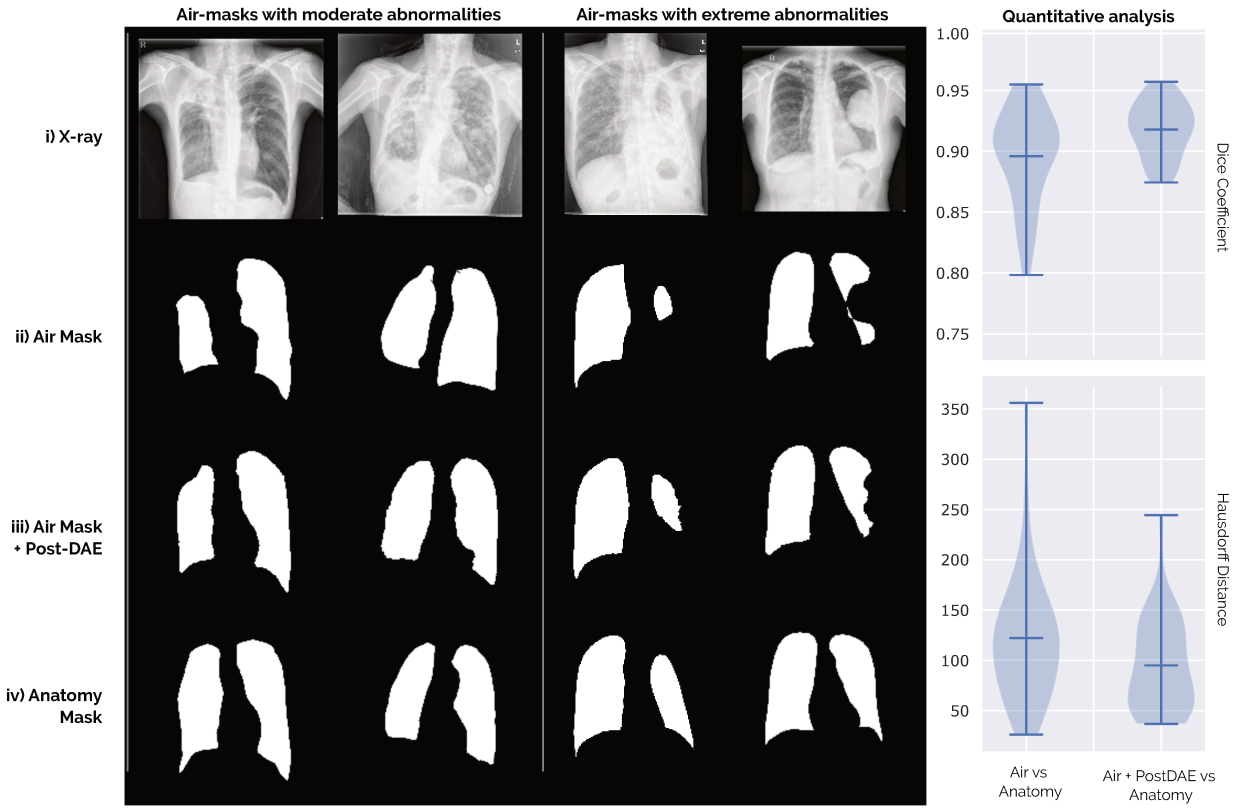


Fig. 5: Analysis for out-of-distribution segmentation masks presenting moderate and extreme abnormalities for tuberculosis patients from the Shenzen database. See Section V.A for a complete discussion about this experiment.

of a particular disease or radiological occlusions, could make these masks look entirely different from the anatomically plausible cases. To analyze the behaviour of our model in this scenario, we employed a different chest X-ray dataset containing patients diagnosed with tuberculosis. This dataset is a subset of the original Shenzhen database [32], [33] formed by 38 X-ray images exhibiting tuberculosis manifestations. Every image was annotated by two expert radiologists following different approaches to delineate the lungs as discussed in [34]. The first approach was to segment only the air cavity part of the lung field, i.e. segmenting only the dark part and ignoring lighter areas covered with fluid. We call these the *air masks*. In the second approach, the annotator delineated the expected anatomy of the lungs, including occluded areas following a comparative approach by mirroring the normal lung field onto the abnormal one [34]. We call these *anatomy masks*. Figure 5 shows examples for both types of segmentation masks (rows (ii) and (iv)). Note that those corresponding to the air approach might present moderate or extreme abnormalities (e.g. missing complete parts of the lung). We applied the Post-DAE model to the *air masks* and analyzed its effect.

We used a Post-DAE model trained with the JSRT dataset, where the lung masks are mostly anatomically plausible since there are no big abnormalities or occlusions in the images. As expected, our method tends to map the air to the anatomy masks. However, note that when abnormalities are too extreme (see columns 3 and 4 in Figure 5) the real anatomy can not be completely reconstructed. We quantified this experiment by measuring the Dice coefficient between the air and anatomy masks before and after post-processing the air masks with Post-DAE. The violin plots included in Figure 5 show that the post-processed air masks are significantly closer to the anatomy masks than the original ones, both in terms of Dice and Hausdorff metrics. This constitutes, at the same time, an advantage and a limitation of our approach: Post-DAE will transform the segmentation masks so that they look closer to the anatomically plausible ones used at training. These are important facts that must be considered when designing segmentation workflows which include Post-DAE. The same holds for problems different from anatomical segmentation. In scenarios like brain lesion or tumor segmentation, where shape and topology is not regular, the applicability of Post-DAE may be limited.

VI. Conclusions

In this work we have shown that denoising autoencoders can be used to render erroneous segmentations of different organs into anatomically plausible masks. Our method works as an independent post-processing step, allowing to incorporate anatomical priors into arbitrary segmentation methods. The provided experimental evaluation in the context of binary and multi-class anatomical segmentation of X-ray and CMR images indicates that our method can deal with a variety of anatomical structures in different image modalities. Moreover, Post-DAE does not use intensity information. Therefore, it can be trained with unpaired segmentation masks annotated on

different image modalities or coming from segmentation-only datasets, making the method robust to domain shift. Post-DAE can be easily implemented, is fast at inference, can cope with arbitrary shape priors and is independent of the image modality and segmentation method. In the future, we plan to explore the use of Post-DAE in the context of lesion segmentation [35], where the regions of interest are not as regular as anatomical structures.

ACKNOWLEDGMENTS

We thank Alexandros Karargyris, Sema Candemir and Stefan Jaeger for sharing the segmentation masks used in the out-of-distribution experiments.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*, 2015.

[2] K. Kamnitsas *et al.*, "Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61 – 78, 2017.

[3] M. Shakeri *et al.*, "Sub-cortical brain structure segmentation using F-CNN's," in *Proc. of ISBI*, 2016.

[4] A. BenTaieb and G. Hamarneh, "Topology aware fully convolutional networks for histology gland segmentation," in *Proc. of MICCAI*. Springer, 2016, pp. 460–468.

[5] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] A. Larrazabal, C. Martinez, and E. Ferrante, "Anatomical priors for image segmentation via post-processing with denoising autoencoders," in *Proc. of MICCAI*, 2019.

[7] M. S. Nosrati and G. Hamarneh, "Incorporating prior knowledge in medical image segmentation: a survey," *arXiv preprint arXiv:1607.01092*, 2016.

[8] O. Oktay *et al.*, "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation," *IEEE TMI*, vol. 37, no. 2, pp. 384–395, 2018.

[9] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *Proc. of MICCAI*, 2017.

[10] C. Wachinger, M. Brennan, G. Sharp, and P. Golland, "On the importance of location and features for the patch-based segmentation of parotid glands," in *MICCAI Workshop on Image-Guided Adaptive Radiation Therapy*, 2014.

[11] C. Wachinger, M. Reuter, and T. Klein, "Deepnat: Deep convolutional neural network for segmenting neuroanatomy," *NeuroImage*, vol. 170, pp. 434–445, 2018.

[12] N. Paragios, E. Ferrante, B. Glocker, N. Komodakis, S. Parisot, and E. I. Zacharaki, "(Hyper)-graphical models in biomedical image analysis," *Medical Image Analysis*, jun 2016.

[13] A. V. Dalca, J. Guttag, and M. R. Sabuncu, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9290–9299.

[14] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *JMLR*, vol. 11, pp. 3371–3408, 2010.

[15] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. of Fourth International Conference on 3D Vision (3DV)*, 2016.

[16] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," in *Proc. of Nips*, 2011.

[17] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-supervised learning*. MIT Press, 2009.

[18] N. Pawlowski *et al.*, "Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders," in *MIDL 2018*, 2018.

[19] H. Uzunova, S. Schultz, H. Handels, and J. Ehrhardt, "Unsupervised pathology detection in medical images using conditional variational autoencoders," *International journal of computer assisted radiology and surgery*, vol. 14, no. 3, pp. 451–461, 2019.

[20] J. Shiraishi *et al.*, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *Am Jour of Roent*, vol. 174, no. 1, pp. 71–74, 2000.

[21] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright, "Evaluation framework for algorithms segmenting short axis cardiac mri," *The MIDAS Journal*, vol. 49, 2009.

[22] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[23] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine ct via dense classification from sparse annotations," in *MICCAI*. Springer, 2013, pp. 262–270.

[24] M. Frid-Adar, A. Ben-Cohen, R. Amer, and H. Greenspan, "Improving the segmentation of anatomical structures in chest radiographs using u-net with an imagenet pre-trained encoder," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, 2018, pp. 159–168.

[25] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, and E. P. Xing, "Scan: Structure correcting adversarial network for organ segmentation in chest x-rays," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 263–273.

[26] A. A. Novikov, D. Lenis, D. Major, J. Hladuvka, M. Wimmer, and K. Buhler, "Fully convolutional architectures for multiclass segmentation in chest radiographs," *IEEE transactions on medical imaging*, vol. 37, no. 8, pp. 1865–1876, 2018.

[27] L. Mansilla, D. H. Milone, and E. Ferrante, "Learning deformable registration of medical images with anatomical constraints," *Neural Networks*, vol. 124, pp. 269–279, 2020.

[28] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache, "3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation," *IEEE transactions on medical imaging*, vol. 37, no. 9, pp. 2137–2148, 2018.

[29] A. H. Curiale, F. D. Colavecchia, and G. Mato, "Automatic quantification of the lv function and mass: a deep learning approach for cardiovascular mri," *Elsevier CMPB*, vol. 169, pp. 37–50, 2019.

[30] M. Chen, L. Fang, and H. Liu, "Fr-net: Focal loss constrained deep residual networks for segmentation of cardiac mri," in *ISBI 2019*. IEEE, 2019, pp. 764–767.

[31] J. V. Stough, J. DiPalma, Z. Ma, B. K. Fornwalt, and C. M. Haggerty, "Ventricular segmentation and quantitative assessment in cardiac mr using convolutional neural networks," in *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10578. International Society for Optics and Photonics, 2018, p. 1057826.

[32] S. Candemir *et al.*, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE TMI*, vol. 33, no. 2, pp. 577–590, 2013.

[33] S. Jaeger *et al.*, "Automatic tuberculosis screening using chest radiographs," *IEEE TMI*, vol. 33, no. 2, pp. 233–245, 2013.

[34] A. Karargyris *et al.*, "Combination of texture and shape features to detect pulmonary abnormalities in digital chest x-rays," *International journal of computer assisted radiology and surgery*, vol. 11, no. 1, pp. 99–106, 2016.

[35] N. Roulet, D. F. Slezak, and E. Ferrante, "Joint learning of brain lesion and anatomy segmentation from heterogeneous datasets," in *MIDL*, 2019, pp. 401–413.

[36] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

SUPPLEMENTARY MATERIAL

A. UNet details

The UNet model (see Table II) receives a 1024x1024 gray image as input and was trained using the soft Dice loss [15], batch size of 4, Adam optimizer with learning rate 1e-5 and the other parameters as by Keras default. We used data augmentation including random rotations, shifts, zoom and shear. We also used dropout for regularization, including a dropout layer after layer L_5 with keep probability p=0.5. For the multi-class UNet we used categorical cross-entropy loss and changed the initial learning rate to to 1e-4.

TABLE II: Detailed description of the UNet architecture used as baseline model segmentation

		Kernel	**Stride**	**#Kernels**		**NonLin**	
				Binary	MC	Binary	MC
L1	Conv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
	Max Pooling	(f:2,2)	(s:2,2)				
L2	Conv	(f:3,3)	(s:1,1)	(N:32)	(N:32)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:32)	(N:32)	ReLu	ReLu
	Max Pooling	(f:2,2)	(s:2,2)				
L3	Conv	(f:3,3)	(s:1,1)	(N:64)	(N:64)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:64)	(N:64)	ReLu	ReLu
	Max Pooling	(f:2,2)	(s:2,2)				
L4	Conv	(f:3,3)	(s:1,1)	(N:128)	(N:128)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:128)	(N:128)	ReLu	ReLu
	Max Pooling	(f:2,2)	(s:2,2)				
L5	Conv	(f:3,3)	(s:1,1)	(N:256)	(N:256)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:256)	(N:256)	ReLu	ReLu
L6	UpConv	(f:3,3)	(s:1,1)	(N:128)	(N:128)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:128)	(N:128)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:128)	(N:128)	ReLu	ReLu
L7	UpConv	(f:3,3)	(s:1,1)	(N:64)	(N:64)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:64)	(N:64)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:64)	(N:64)	ReLu	ReLu
L8	UpConv	(f:3,3)	(s:1,1)	(N:32)	(N:32)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:32)	(N:32)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:32)	(N:32)	ReLu	ReLu
L9	UpConv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
L10	Conv	(f:3,3)	(s:1,1)	(N:2)	(N:3)	ReLu	ReLu
	Conv	(f:1,1)	(s:1,1)	(N:1)	(N:3)	Sigmoid	SoftMax

B. Post-DAE

Post-DAE (see Table III) receives a 1024x1024 segmentation as input. The network was also trained to minimize the Dice loss function using Adam Optimizer. We used learning rate of 0.0001, batch size of 15 and 150 epochs. The multi-class Post-DAE implementation receives a one-hot encoded segmentation of size 1024x1024x3 segmentation as input. Because of memory restrictions, in this case we reduced the batch size to 8.

C. Additional experiments for segmentation-only datasets

We performed an extra experiment aiming to show that it is possible to use annotations from a different dataset to train Post-DAE. We used the DAE trained with JSRT database to post-process results obtained with the binary UNet for the Montgomery County X-ray Set [36], a different chest X-ray dataset with manual lung annotations. X-ray images in this dataset were acquired from the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA. This set contains 138 posterior-anterior x-rays, which were divided in 100 for training, 14 for validation and 24 for testing. The size of the X-rays is either 4020x4892 or 4892x4020 pixels. All images were downsampled to 1024x1024 in our experiments, padded with 0's to obtain a 1:1 aspect ratio and rigidly aligned. With this dataset, we trained two segmentation models: a Random Forest and a UNet architecture (saving its output every 5 epochs during training), predicting segmentation masks for the test fold using all these models. These masks were then post-processed using Post-DAE. Figure 6 shows the results for this experiment, where unpaired annotations coming from a different dataset are used to train Post-DAE. It can be observed how our method improves the segmentation quality even when the annotations used to train Post-DAE are coming from a different dataset.

TABLE III: Detailed architecture of the simple denoising auto encoder model used to implement the proposed Post-DAE.

		Kernel	**Stride**	**#Kernels**		**NonLin**	
				Binary	MC	Binary	MC
L_1	Conv	(f:3,3)	(s:2,2)	(N:16)	(N:16)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:16)	(N:16	ReLu	ReLu
L2	Conv	(f:3,3)	(s:2,2)	(N:32)	(N:32)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:32)	(N:32)	ReLu	ReLu
L3	Conv	(f:3,3)	(s:2,2)	(N:32)	(N:32)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:32)	(N:32)	ReLu	ReLu
L4	Conv	(f:3,3)	(s:2,2)	(N:32)	(N:32)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:32)	(N:32)	ReLu	ReLu
L5	Conv	(f:3,3)	(s:2,2)	(N:32)	(N:32)	ReLu	ReLu
L6	FC	-	-	(N:512)	(N:1024)	None	None
L6	FC	-	-	(N:1024)	(N:4096)	Relu	Relu
L8	UpConv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
L9	UpConv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
L10	UpConv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
L11	UpConv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
L12	UpConv	(f:3,3)	(s:1,1)	(N:16)	(N:16)	ReLu	ReLu
	Conv	(f:3,3)	(s:1,1)	(N:1)	(N:3)	Sigmoid	SoftMax

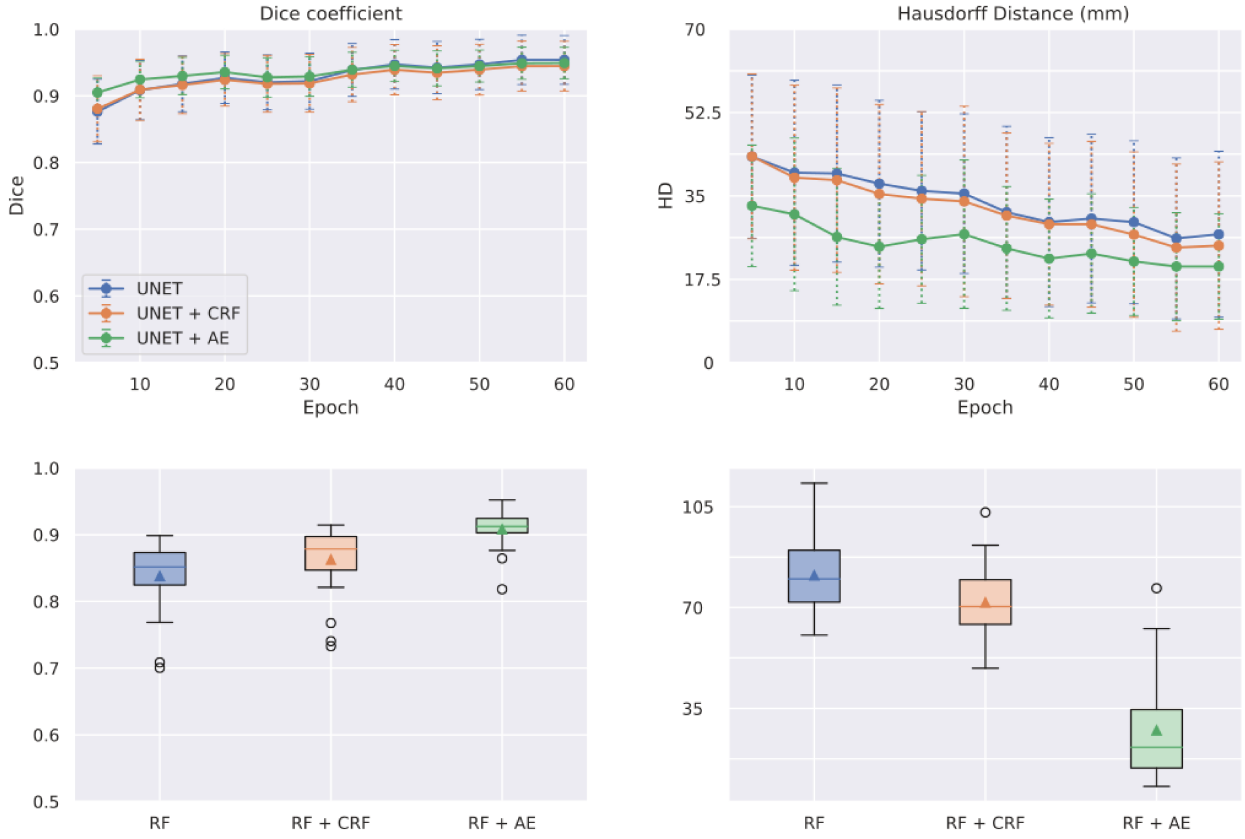


Fig. 6: Quantitative evaluation of the proposed method in a new data-set, which was not seen by the Post-DAE. Top row shows mean and standard deviation for post-processing UNet predictions on the test fold at different training stages (from 5 epochs to convergence). We use Dice coefficient and Hausdorff distance to measure the segmentation quality. Bottom row show results for post-processing the random forest predictions. The triangle in the box indicates the mean value.

Anexo C

Orthogonal Ensemble Networks for Biomedical Image Segmentation

Orthogonal Ensemble Networks for Biomedical Image Segmentation

Agostina J. Larrazabal[1], César Martínez[1], Jose Dolz[2], Enzo Ferrante[1]

[1]Research institute for signals, systems and computational intelligence, sinc(i), FICH-UNL / CONICET, Santa Fe, Argentina
[2]Laboratory for Imagery, Vision and Artificial Intelligence, École de Technologie Supérieure, Montreal, Canada

Abstract. Despite the astonishing performance of deep-learning based approaches for visual tasks such as semantic segmentation, they are known to produce miscalibrated predictions, which could be harmful for critical decision-making processes. Ensemble learning has shown to not only boost the performance of individual models but also reduce their miscalibration by averaging independent predictions. In this scenario, model diversity has become a key factor, which facilitates individual models converging to different functional solutions. In this work, we introduce Orthogonal Ensemble Networks (OEN), a novel framework to explicitly enforce model diversity by means of orthogonal constraints. The proposed method is based on the hypothesis that inducing orthogonality among the constituents of the ensemble will increase the overall model diversity. We resort to a new pairwise orthogonality constraint which can be used to regularize a sequential ensemble training process, resulting on improved predictive performance and better calibrated model outputs. We benchmark the proposed framework in two challenging brain lesion segmentation tasks –brain tumor and white matter hyper-intensity segmentation in MR images. The experimental results show that our approach produces more robust and well-calibrated ensemble models and can deal with challenging tasks in the context of biomedical image segmentation.

Keywords: image segmentation, ensemble networks, orthogonal constraints

1 Introduction

In the past few years, deep learning-based methods have become the *de facto* solution for many computer vision and medical imaging tasks. Nevertheless, despite their success and great ability to learn highly discriminative features, they are shown to be poorly calibrated [1], often resulting in over-confident predictions. This results in a major problem, which can have catastrophic consequences in critical decision-making systems, such as medical diagnosis, where the downstream decision depends on predicted probabilities.

Ensemble learning is a simple strategy to improve both the robustness and calibration performance of predictive models [2, 3]. In this scenario, a common approach is to train the same model under different conditions, which can foster the model convergence to different functional solutions. Techniques to produce ensembles include dataset shift [4], Monte-Carlo Dropout [5], batch-ensemble [6] or different model hyperparameters [7], among others. Then, by averaging the predictions, individual mistakes can be dismissed leading to a reduced miscalibration. In this context, ensuring *diversity* across models is a key factor to build a robust ensemble. To promote model diversity in ensembles many mechanisms have been proposed. These include using latent variables [8], integrating attention in the embeddings to enforce different learners to attend to different parts of the object [9] or isolating the adversarial vulnerability in sub-models by distilling non-robust features to induce diverse outputs against a transfer attack [10].

Nevertheless, despite the relevance of obtaining well-calibrated models in clinical applications, relatively few works have studied this problem. Particularly, in the context of medical image segmentation, it was suggested that models trained with the well-known soft Dice loss [11] produce miscalibrated models [12], which tend to be highly overconfident. Furthermore, the recent work in [13] proposed the use of ensembles to improve confidence calibration. However, the importance of model diversity was not assessed in this work. Thus, given the negative impact of miscalibrated models in health-related tasks, and the current practices in medical image segmentation of systematically employing the Dice loss as an objective function, we believe it is of paramount importance to investigate the effect of ensemble learning in image segmentation, and how to enforce model diversity to generate high-performing and well-calibrated models.

Contributions. In this work, we propose a novel learning strategy to boost model diversity in deep convolutional neural networks (DCNN) ensembles, which improves both segmentation accuracy and model calibration in two challenging brain lesion segmentation scenarios. The main hypothesis is that inducing orthogonality among the constituents of the ensemble will increase the overall model diversity. We resort to a novel pairwise orthogonality constraint which can be used to regularize a sequential ensemble training process, resulting on improved predictive performance and better calibrated model outputs. In this context, our contributions are 3-fold: (1) we propose a novel filter orthogonality constraint for ensemble diversification, (2) we show that diversified ensembles improve not only segmentation accuracy but also confidence calibration and (3) we showcase the proposed framework in two challenging brain lesion segmentation tasks, including tumor and white-matter hyperintensity (WMH) segmentation on magnetic resonance images.

2 Related works

Diversifying ensembles has been used to improve classification and segmentation performance of DCNNs in several contexts. In [14] authors propose an explicit

way to construct diverse ensembles bringing together multiple CNN models and architectures. Although they obtain successful results, this approach requires to manually design and train various architectures. An ensemble of 3D U-Nets with different hyper-parameters for brain tumor segmentation is proposed in [15], where authors point out that using different hyper-parameters reduces the correlations of random errors with respect to homogeneous configurations. However, no study on the diversity of the models and its influence on performance is presented. In [16] authors present a different view, highlighting that many automatic segmentation algorithms tend to exhibit asymmetric errors, typically producing more false positives than false negatives. By modifying the loss function, they train a diverse ensemble of models with very high recall, while sacrificing their precision, with a sufficiently high threshold to remove all false positives. While the authors achieve a significant increase in performance no study on the final calibration of the ensemble is carried out.

Following the success of ensemble methods at improving discriminative performance, its capability to improve confidence calibration has begun to be explored. [2] uses a combination of independent models to reduce confidence uncertainty by averaging predictions over multiple models. In [13] authors achieve an improvement in both segmentation quality and uncertainty estimation by training ensembles of CNNs with random initialization of parameters and random shuffling of training data. While these results are promising, we believe that confidence calibration can be further improved by directly enforcing diversity into the models instead of randomly initializing the weights.

As pointed out in [17] over-sized DNNs often result in a high level of overfitting and many redundant features. However, when filters are learned to be as orthogonal as possible, they become decorrelated and their filter responses are no longer redundant, thereby fully utilizing the model capacity. [18] follows a very similar approach but they regularize both negatively and positively correlated features according to their differentiation and based on their relative cosine distances. Differently from these works where ortogonality constraints are used to decorrelated the filters within a single model, here we propose to enforce filter orthogonality among the constituents of the ensemble to boost model diversity.

3 Orthogonal Ensemble Networks for Image Segmentation

Given a dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})_i\}_{0 \leq i \leq |\mathcal{D}|}$ composed of images $\mathbf{x}$ and corresponding segmentation masks $\mathbf{y}$, we aim at training a model which approximates the underlying conditional distribution $p(\mathbf{y}|\mathbf{x})$, mapping input images $\mathbf{x}$ into segmentation maps $\mathbf{y}$. Thus, $p(y_j = k|\mathbf{x})$ will indicate the probability that a given pixel (or voxel) j is assigned class $k \in \mathcal{C}$ from a set of possible classes $\mathcal{C}$. The distribution is commonly approximated by a neural network $f_{\mathbf{w}}$ parameterized by weights $\mathbf{w}$. In other words, $f_{\mathbf{w}}(\mathbf{x}) = p(\mathbf{y}|\mathbf{x}; \mathbf{w})$. Parameters $\mathbf{w}$ are learnt so that they minimize a particular loss function over the training dataset. Given

a set of segmentation networks $\{f_{\mathbf{w}^1}, f_{\mathbf{w}^2} ... f_{\mathbf{w}^N}\}$, a simple strategy to build an ensemble network $f_{\mathbf{E}}$ is to average their predictions as:

$$f_{\mathbf{E}}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} f_{\mathbf{w}^i}(\mathbf{x}). \tag{1}$$

Under the hypothesis that diversifying the set of models $f^i_{\mathbf{w}}$ will lead to more accurate and calibrated ensemble predictions, we propose to boost its overall performance by incorporating pairwise orthogonality constraints during training.

Inducing model diversity via orthogonal constraints. Modern deep neural networks are parameterized by millons of learnable weights, resulting in redundant features that can be either a shifted version of each other or be very similar with almost no variation [18]. Inducing orthogonality between convolutional filters from the same layer of a given network has shown to be a good way to reduce filter redundancy [17]. Here we exploit this principle not only to avoid redundancy within a single neural model, but among the constituents of a neural ensemble.

Given two vectors $\mathbf{x}$ and $\mathbf{y}$, cosine similarity quantifies orthogonality (or decorrelation), ranging from -1 (i.e., exactly opposite) to 1 (i.e., exactly the same), with 0 indicating orthogonality. It can be defined as:

$$\mathrm{SIM}_C(\mathbf{x}, \mathbf{y}) = \frac{< \mathbf{x}, \mathbf{y} >}{||\mathbf{x}||\ ||\mathbf{y}||}. \tag{2}$$

Following [18], we consider the squared cosine similarity to induce orthogonality between filters through a new regularization term in the loss function. An advantage of this measure is that it takes into account both negative and positive correlations.

In order to enforce diversity within and between the ensemble models, we propose to include two regularization terms into the overall learning objective. The first one, referred to as self-orthogonality loss ($\mathcal{L}_{\mathrm{SelfOrth}}$), aims at penalizing the correlation between filters in the same layer, for a given model. Thus, for a given convolutional layer l, this term is calculated as follows:

$$\mathcal{L}_{\mathrm{SelfOrth}}(\mathbf{w}_l) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathrm{SIM}_C(\mathbf{w}_{l,i}, \mathbf{w}_{l,j})^2, \tag{3}$$

where $\mathbf{w}_{l,i}$ and $\mathbf{w}_{l,j}$ are vectorized versions of each of the n convolutional kernels from layer l. We also define an inter-orthogonality loss term ($\mathcal{L}_{\mathrm{InterOrth}}$) which penalizes correlation between filters from different models in the ensemble. To this end, following a sequential training scheme, the inter-orthogonality loss for layer l of model N_e is estimated as follows:

$$\mathcal{L}_{\mathrm{InterOrth}}(\mathbf{w}_l; \{\mathbf{w}^e_l\}_{0 \leq e < N_e}) = \frac{1}{N_e} \sum_{e=0}^{N_e - 1} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{SIM}_C(\mathbf{w}_{l,i}, \mathbf{w}^e_{l,j})^2, \tag{4}$$

where $\{\mathbf{w}_l^e\}_{0 \leq e < N_e}$ are the parameters of the previous $N_e - 1$ models trained during the sequential ensemble construction.

Thus, the learning objective to train the proposed OEN amounts to:

$$\mathcal{L} = \mathcal{L}_{Seg} + \lambda \sum_l \Big(\mathcal{L}_{\text{SelfOrth}}(\mathbf{w}_l) + \mathcal{L}_{\text{InterOrth}}(\mathbf{w}_l; \{\mathbf{w}_l^e\}) \Big), \tag{5}$$

where $\mathcal{L}_{Seg}$ is the segmentation loss (e.g. soft Dice loss or cross entropy) and λ is a hyperparamether controlling the influence of the orthogonality terms. [1]

4 Experimental framework

Database description. We benchmark the proposed method in the context of brain tumor and WMH segmentation in MR images. For brain tumor we use the BraTS 2020 dataset [19–21] which contains 369 images with expert segmentation masks (including GD-enhancing tumor, peritumoral edema, and the necrotic and non-enhancing tumor core). Each patient was scanned with FLAIR, T1ce, T1, and T2. The images were re-sampled to an isotropic $1.0mm$ voxel spacing, skull-striped and co-registered by the challenge organizers. The provided training set, we divide the database in training (315), validation (17) and test (37). The second dataset [22] consists of 60 MR images with binary masks indicating the presence of WMH lesions. For each subject, co-registered 3D T1-weighted and a 2D multi-slice FLAIR images were provided. We split the dataset in training (42), validation (3) and test (15). All images have $3mm$ spacing in the z dimension, and approximately $1mm \times 1mm$ in the axial plane.

Segmentation network. For all the experiments, the backbone segmentation network was a state-of-the-art ResUNet architecture [23] implemented in Keras 2.3 with TensorFlow as backend, with soft Dice [11] as segmentation loss $\mathcal{L}_{Seg}$. For the BraTS dataset, the input was a four-channel tensor (FLAIR, T1ce, T1, and T2) and a softmax activation was used as output, whereas a two-channel input (T1, FLAIR) was employed in the WMH, with a sigmoid activation function in the output. During training, patches of size $64 \times 64 \times 64$ were extracted from each volume, and networks were trained until convergence by sampling the patches randomly, with equal probability for each class in the case of tumour segmentation, and 0.9 probability in the case of WMH. We used Adam optimizer with a batch size of 64. The initial learning rate was set to 0.001 for BraTS and 0.0001 for WMH, and it was reduced by a factor of 0.85 every 10 epochs. Hyperparameters were chosen using the validation split, and results reported on the hold-out test set.

Baselines and ensemble training. We trained two different baselines to benchmark the proposed method. In the first one (*random* ensemble) each model

[1] Our code associated to the orthogonal ensemble networks training is publicly available at: `https://github.com/agosl/Orthogonal_Ensemble_Networks`

was randomly initialized and trained to reduce only the segmentation error $\mathcal{L}_{Seg}$. Therefore, its main source of diversity comes from the initialization of the weights. The second approach (*self-orthogonal* ensemble) includes the $\mathcal{L}_{\text{SelfOrth}}$ term in the learning objective, creating and ensemble of models individually trained with the self-orthogonality constraint. Thus, while each model learns orthogonal filters, orthogonality between different models in the ensemble was not imposed. We compared these two models with the proposed orthogonal ensemble network which also encourages inter-model diversity by minimizing the full objective defined in Eq. 5 (referred as *inter-orthogonal*). Note that in our approach models are trained sequentially. For each of the proposed settings we trained 10 models. During evaluation, we assembled groups of 1, 3 and 5 models from each setting by averaging the individual probability outputs. To provide better statistics, we repeated this process 10 times, each with different model selection. We empirically observed that beyond 5 models, the performance of the ensemble did not improve. Furthermore, λ was set to 0.1 and 1 for the WMH and brain tumour segmentation task, respectively.

Measuring calibration for image segmentation. Given a segmentation network $f_{\mathbf{w}}$, if the model is well-calibrated its output for a single pixel j can be interpreted as the probability $p(y_j = k|\mathbf{x}; \mathbf{w})$ for a given class $k \in \mathcal{C}$. In this case, the class probability can be seen as the model confidence or probability of correctness, and can be used as a measure for predictive uncertainty at the pixel level [13]. A common metric used to measure calibration performance is the Brier score [24], a proper scoring rule whose optimal value corresponds to a perfect prediction. In other words, a system that is both perfectly calibrated and perfectly discriminative will have a Brier score of zero. In the context of image segmentation, for an image with N pixels (voxels), the Brier score can be defined as:

$$Br = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{C}|}\sum_{k=1}^{|\mathcal{C}|}\Big(p(y_i = k|\mathbf{x}; \mathbf{w}) - \mathbb{1}[\bar{y}_i = k]\Big)^2, \tag{6}$$

where $\mathbb{1}[\bar{y}_i = k]$ is the indicator function whose value is 1 when $\bar{y}_i$ (the ground truth class for pixel i) is equal to k, and 0 otherwise.

Stratified Brier Score. In problems with highly imbalanced classes (such as brain lesion segmentation where most of the pixels are background), calibration may be good overall but poor for the minority class. In this case, the majority class will dominate and miscalibration in the class of interest will not be reflected in the standard Brier score. In [25], the authors proposed the stratified Brier score to measure calibration in binary classification problems with high imbalance. Here, we extend this concept to the segmentation task and propose to measure the stratified Brier score individually per-class, treating every structure of interest as a binary segmentation problem, to account for mis-calibration in the minority classes. For a given image with ground truth segmentation $\bar{\mathbf{y}}$, we construct the *stratified* Brier score for the class k, Br^k, by computing it only

in the subset of pixels $\mathcal{P}_k = \{p : \bar{y}_p = k\}$, i.e. pixels whose ground truth label is k. The problem is therefore binarized considering all the other classes within a single background class. The formulation of the stratified Brier score Br^k is given by:

$$Br^k = \frac{1}{|\mathcal{P}_k|} \sum_{i \in \mathcal{P}_k} \Big(p(y_i = k|\mathbf{x}; \mathbf{w}) - \mathbb{1}[\bar{y}_i = k] \Big)^2. \tag{7}$$

Segmentation evaluation. In addition to the metrics presented to measure the model miscalibration, we resort to the common Dice Similarity coefficient (DSC) to assess the quality of the segmentations.

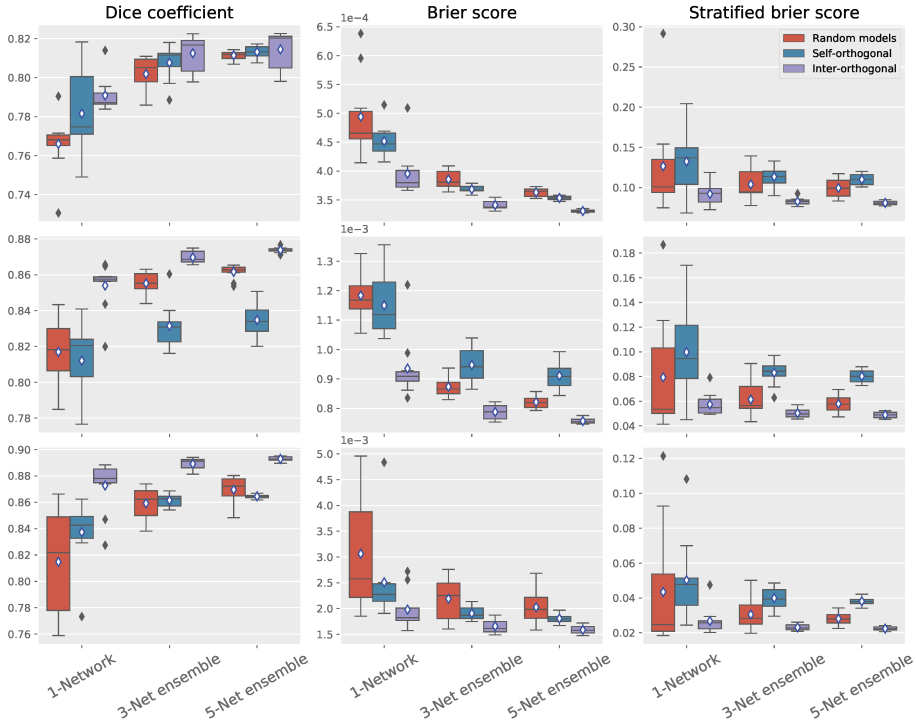


Fig. 1. Quantitative evaluation of the proposed method on BraTS: Rows from top to bottom show results for: (i) enhanced tumor; (ii) tumor core; (iii) whole tumor. Boxplots show mean and standard deviation for predictions obtained with individual models, 3-networks ensembles and 5-networks ensembles.

5 Results and discussion

We present quantitative results for brain tumor and WMH segmentation in Fig. 1 and Fig. 2, respectively. We can observe that the model just integrating *self-orthogonality* outperforms the baseline model across groups and metrics. This

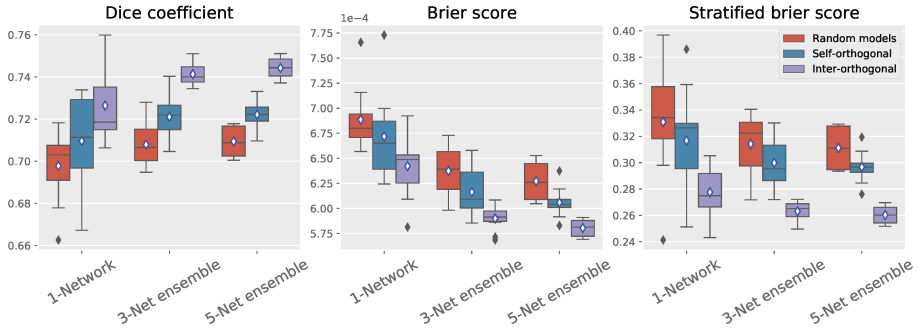


Fig. 2. Quantitative evaluation of the proposed method for WMH segmentation. Boxplots show mean and standard deviation for predictions obtained with individual models, 3-networks ensembles and 5-networks ensembles.

improvement is further stressed when explicitly enforcing model diversity by incorporating the $inter - orthogonality$ term computed between pairs of models during sequential training. In particular, our proposed learning strategy consistently leads to improvement on both model calibration and segmentation performance and across the two different segmentation tasks. This demonstrates the benefits of the proposed learning strategy to generate well-calibrated and highly performing segmentation models.

Another important observation is related to differences between Brier and stratified Brier scores. Given the small Brier value reported for all the models (less than 1^{-3}), one could think that these models are well calibrated. However, when having a closer look at the stratified Brier score, the higher value (more than 0.1 in most of the cases) reflects calibration issues. This results from the majority class dominating the traditional Brier score. Thus, studying the stratified Brier score allows us to better appreciate the improvements obtained by the inter-orthogonal ensemble with respect to the other models.

In addition, we depict in Fig. 3 the variance in the predictions across the components of the ensemble trained with and without the orthogonal losses, demonstrating that the orthogonal constraints bring diversity to the ensemble. As expected, we found that integrating the inter-orthogonal objective term leads to an increase in the variance of the predictions compared to the baseline models.

Last but not least, it is surprising to see that the inter-orthogonal regularization term boosts the performance even when considering the individual models. We believe that this is due to a regularization effect of the inter-orthogonal term, which implicitly reduces the complexity of the model by adding orthogonality constraints with respect to specific points in the parameter space, i.e. the weights of the previously trained models.

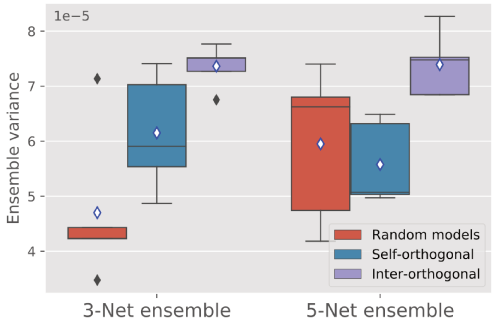


Fig. 3. Quantitative evaluation of the ensembles diversity. Boxplots depict the mean and standard deviation of the variance in the predictions when training the ensemble with and without the proposed orthogonal losses.

6 Conclusions

In this work we introduced Orthogonal Ensemble Networks (OEN), a novel training framework that produces more diverse ensembles. Our formulation explicitly imposes orthogonal constraints during training by integrating a regularization term that enhances the inter-model diversity. Experiments across two different segmentation tasks have demonstrated that, in addition to improved segmentation performance, the proposed inter-model orthogonality constraints reduce miscalibration, leading to more reliable predictions.

Acknowledgments

The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPUs used for this research, and the support of UNL (CAID-0620190100145LI, CAID-50220140100084LI) and ANPCyT (PICT 2018-03907). This research was enabled in part by support provided by Calcul Québec and Compute Canada.

References

1. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. (2017) 1321–1330
2. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS. (2017)
3. Stickland, A.C., Murray, I.: Diverse ensembles improve calibration. In: ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. (2020)
4. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: NeurIPS. (2019)

5. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. (2016) 1050–1059
6. Wen, Y., Tran, D., Ba, J.: Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In: ICLR. (2020)
7. Wenzel, F., Snoek, J., Tran, D., Jenatton, R.: Hyperparameter ensembles for robustness and uncertainty quantification. In: NeurIPS. (2020)
8. Sinha, S., Bharadhwaj, H., Goyal, A., Larochelle, H., Garg, A., Shkurti, F.: Dibs: Diversity inducing information bottleneck in model ensembles. In: AAAI. (2020)
9. Kim, W., Goyal, B., Chawla, K., Lee, J., Kwon, K.: Attention-based ensemble for deep metric learning. In: ECCV. (2018) 736–751
10. Yang, H., Zhang, J., Dong, H., Inkawhich, N., Gardner, A., Touchet, A., Wilkes, W., Berry, H., Li, H.: Dverge: diversifying vulnerabilities for enhanced robust generation of ensembles. arXiv preprint arXiv:2009.14720 (2020)
11. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV), IEEE (2016) 565–571
12. Sander, J., de Vos, B.D., Wolterink, J.M., Išgum, I.: Towards increased trustworthiness of deep learning segmentation methods on cardiac mri. In: Medical Imaging 2019: Image Processing. Volume 10949., International Society for Optics and Photonics (2019) 1094919
13. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging **39**(12) (2020) 3868–3878
14. Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: International MICCAI brainlesion workshop, Springer (2017) 450–462
15. Feng, X., Tustison, N.J., Patel, S.H., Meyer, C.H.: Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. Frontiers in computational neuroscience **14** (2020) 25
16. Ma, T., Zhang, H., Ong, H., Vora, A., Nguyen, T.D., Gupta, A., Wang, Y., Sabuncu, M.R.: Ensembling low precision models for binary biomedical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (2021) 325–334
17. Wang, J., Chen, Y., Chakraborty, R., Yu, S.X.: Orthogonal convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 11505–11515
18. Ayinde, B.O., Inanc, T., Zurada, J.M.: Regularizing deep neural networks by enhancing diversity in feature extraction. IEEE transactions on neural networks and learning systems **30**(9) (2019) 2650–2661
19. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4** (2017) 170117
20. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)

21. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10) (2014) 1993–2024
22. Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE transactions on medical imaging **38**(11) (2019) 2556–2568
23. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5) (2018) 749–753
24. Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly weather review **78**(1) (1950) 1–3
25. Wallace, B.C., Dahabreh, I.J.: Improving class probability estimates for imbalanced data. Knowledge and information systems **41**(1) (2014) 33–52

Anexo D

Maximum Entropy on Erroneous Predictions (MEEP): Improving model calibration for medical image segmentation

Maximum Entropy on Erroneous Predictions (MEEP): Improving model calibration for medical image segmentation

Agostina Larrazabal[1] , Cesar Martínez[1], José Dolz[2]*, Enzo Ferrante[1]*
[1] *sinc(i), CONICET - Universidad Nacional del Litoral, Santa Fe, Argentina*
[2] *ÉTS, Montreal, Canada*
{alarrazabal, cmartinez, eferrante} @sinc.unl.edu.ar, jose.dolz@etsmtl.ca

Abstract

Modern deep neural networks have achieved remarkable progress in medical image segmentation tasks. However, it has recently been observed that they tend to produce overconfident estimates, even in situations of high uncertainty, leading to unreliable models. In this work we introduce Maximum Entropy on Erroneous Predictions (MEEP), a training strategy for segmentation networks which selectively penalizes overconfident predictions, focusing only on misclassified pixels. In particular, we design a regularization term that encourages high entropy posteriors for wrong predictions, increasing the network uncertainty in complex scenarios. Our method is agnostic to the neural architecture, does not increase model complexity and can be coupled with multiple segmentation loss functions. We benchmark the proposed strategy in two challenging 3D medical image segmentation tasks: white matter hyperintensity lesions in magnetic resonance images (MRI) of the brain, and atrial segmentation in cardiac MRI. The experimental results obtained when training two state-of-the-art segmentation architectures (UNet and ResUNet) demonstrate that coupling MEEP with standard segmentation losses leads to improvements not only in terms of model calibration, but also in segmentation quality.

1. Introduction

Deep learning-based methods have become the *de facto* solution for many computer vision and medical imaging tasks, dominating the literature in image segmentation. Nevertheless, despite their success and great ability to learn highly discriminative features, they are shown to be poorly calibrated, often resulting in over-confident predictions, even when they are wrong [1]. Thus, when a model is miscalibrated, there is little correlation between the confidence of its predictions and how accurate such predictions actually are [2]. This results in a major problem, which can have catastrophic consequences in critical decision-making systems, such as medical diagnosis, where the downstream decision depends on predicted probabilities.

As shown by [4], the uncertainty estimates inferred from segmentation models can provide insights into the confidence of any particular segmentation mask, and highlight areas of likely segmentation errors for the practitioner. In order to improve the accuracy and reliability of CNN-based medical image segmentation models, it is crucial to develop both accurate and well-calibrated systems, which allow the user to distinguish between reliable and unreliable predictions. Even though this is an important aspect that should be considered when comprehensively evaluating model performance in image segmentation (particularly important in biomedical imaging), most of the metrics used to quantify segmentation quality focus only on the maximum a posteriori (MAP) estimates. Since these metrics (like Sørensen–Dice (DSC) coefficient and Jaccard index) operate on the space of hard segmentation masks instead of soft probabilities, they are good at evaluating the discriminative performance of the model (i.e. the capacity to distinguish between different classes of interest) but they overlook the quality of the estimated posteriors. Differently from DSC and Jaccard, proper scoring rules [5] like the negative log likelihood or Brier score [6] operate directly on the estimated posteriors and are not only affected by the discrimination performance, but also by model calibration. The more visually appealing reliability diagrams [7], and associated quantitative metrics like the Expected Calibration Error (ECE) [8], are also affected by calibration and have been adopted to comprehensively evaluate model performance for deep networks. Even though several studies dealing with calibration on image classification have been published [1, 9], assessing the effect of miscalibrated networks in the context of image segmentation, and particularly biomedical images, has been mostly overlooked.

Contribution. In this work, we propose a new method based on entropy maximization to enhance the quality of the posteriors estimated by segmentation networks, and as-

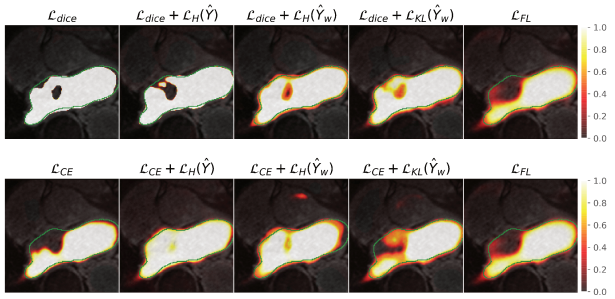


Figure 1. Segmentation results obtained on the left atrium (LA) segmentation task in MR images (ground truth is shown in green). From left to right: **1st column** corresponds to UNet trained with standard segmentation losses, e.g., Dice ($\mathcal{L}_{dice}$) and cross entropy ($\mathcal{L}_{CE}$). **2nd column:** we incorporate a regularization $\mathcal{L}_H(\hat{Y})$ term which penalizes low entropy in all the predictions following [10]. **3rd and 4th columns:** results obtained with variations of the proposed MEEP method which penalizes low entropy predictions only on misclassified pixels. We can clearly observe how the proposed MEEP models push the predicted probabilities towards 0.5 in highly uncertain areas. **5th column:** results with focal loss.

sess its effectiveness using a variety of metrics to reflect the improvement in terms of both discrimination and calibration performance. Our hypothesis is that penalizing low entropy on the probability estimates for erroneous pixel predictions during training should help to avoid overconfident estimates in situations of high uncertainty. The underlying idea is that, if a pixel is difficult to classify, it is better assigning uniformly distributed (i.e. high entropy) probabilities to all classes, rather than being overconfident on the wrong class. To this end, we design two simple regularization terms which push the estimated posteriors for misclassified pixels towards a uniform distribution by penalizing low entropy predictions. We benchmark the proposed method in two challenging medical image segmentation tasks, and compare our approach with previous works also based on low entropy penalization. Our experiments show that the proposed model encourages the network to avoid overconfident wrong predictions resulting on significant improvements. Moreover, it can be coupled with arbitrary segmentation losses and neural architectures, is easy to implement and does not increase model complexity. Last, we further show that assessing the performance of segmentation models only from a discriminative perspective does not provide a complete overview of the model performance, and argue that including calibration metrics should be preferred. This will allow to not only evaluate the segmentation power of a given model, but also its reliability, which is of pivotal importance in a medical environment.

2. Related work

Calibration of deep learning models. Obtaining well-calibrated probability estimates of supervised machine learning approaches has attracted the attention of the research community even before the deep learning era. Initial attempts to address this problem include histogram binning [11] or Bayesian binning [8], among others, which were not initially formulated for deep learning models. Nevertheless, with the increase of popularity of deep neural networks, several works to directly address the calibration of these models have recently emerged.

For instance, Bayesian neural networks learn a posterior distribution over parameters that quantifies parameter uncertainty –a type of *epistemic uncertainty*–, providing a natural approach to quantify model uncertainty. Among others, well-known Bayesian methods include variational inference [12], dropout-based variational inference [13] or stochastic expectation propagation [14]. A popular non-Bayesian method is ensemble learning, a simple strategy that improves both the robustness and calibration performance of predictive models [15, 16]. Techniques to produce ensembles include dataset shift [17], Monte-Carlo Dropout [13], batch-ensemble [18], different model hyperparameters [19], or orthogonal constraints [20], among others. By averaging the predictions of multiple models, individual mistakes can be dismissed leading to a reduced miscalibration. However, even though this technique tends to improve the networks calibration, it does not directly promote uncertainty awareness. Furthermore, ensembling typically requires retraining several models from scratch, incurring into computationally expensive steps for large datasets and complex models. Guo et *al.* [1] empirically evaluated several post training ad-hoc calibration strategies, finding that a simple temperature scaling of logits yielded the best results. A drawback of this simple strategy, though, is that calibration performance largely degrades under data distribution shift [17].

Another alternative is to address the calibration problem during training, for example by clamping over-confident predictions. In [10], authors proposed to regularize the neural network output by penalizing low entropy output distributions, which was achieved by integrating an entropy regularized term into the main learning objective. We want to emphasize that even though the main motivation in [10] was to achieve better generalization by avoiding overfitting, recent observations [21] highlight that these techniques have a favorable effect on model calibration. In particular, Muller et *al.* [21] advocate that this is due to label smoothing encouraging logits of the correct class to be equidistant –by a constant depending on α– to the logits of the other classes. This prevents logits differences to be very large, which occurs when the standard cross-entropy loss is used, resulting in smoother softmax distributions. In a similar line of work, [9] empirically justified the excellent performance of

focal loss to learn well-calibrated models. More concretely, authors observed that focal loss [22] minimizes a Kullback-Leibler (KL) divergence between the predicted softmax distribution and the target distribution, while increasing the entropy of the predicted distribution.

Uncertainty on image segmentation. Bayesian SegNet [24] was among the first attempts to include a measure of model uncertainty along with pixel-wise class predictions. To achieve this, authors resorted to Monte Carlo sampling with dropout (MCDO) [13] at test time to generate a posterior distribution of pixel class labels. This idea has been further exploited in the medical domain, for example in the context of brain lesions segmentation [25], [26]. Nevertheless, a major drawback of MCDO-based approaches is that they focus on epistemic uncertainty, which can be reduced by using additional training samples. Furthermore, they require multiple forward passes at test time, which can incur in a larger computational burden. To alleviate the high complexity of MCDO strategies, a Bayesian Network was proposed for fast uncertainty estimation with a single forward pass [27] that treats the output of the DNN as a per-voxel factored distribution. However, recent research [28] has found that current state-of-the-art Bayesian neural networks are prone to make very similar bad predictions, whereas ensemble deep neural networks tend to be more diverse in making predictions, and therefore obtain better results in computing uncertainty. In this regard, Jungo et *al.* [29] analyse different uncertainty estimation methods for brain tumor segmentation using U-Net-like architectures and show that, even though most approaches perform similarly, ensemble methods bring a slight advantage. Motivated by these observations, the use of ensembles has gained an increasing popularity [15, 20, 23].

An in-depth analysis of the calibration quality obtained by training segmentation networks with the two most commonly used loss functions, Dice coefficient and cross entropy, was conducted in [23]. In line with [30], [31], authors showed that loss functions directly impact calibration quality and segmentation performance, noting that models trained with soft dice loss tend to be poorly calibrated and overconfident. Authors also highlight the need to explore new loss functions to improve both segmentation and calibration quality. More recently, [4] show that the model architecture does not have a major influence on the quality of the uncertainty estimates, but they do not discuss how uncertainty estimation can be enhanced by training these models with more suitable loss functions. Label smoothing (LS) has also been proposed to improve calibration in segmentation models. However, in contrast to classification networks, the benefit of LS in semantic segmentation remains unclear. Islam et al [32] claim that standard LS flattens the training labels with a uniform distribution without considering the spatial consistency aspect, crucial for image segmentation. Thus, they propose a label smoothing strategy for image segmentation by designing a weight matrix with a Gaussian kernel which is applied across the one-hot encoded expert labels to obtain soft class probabilities. Although considering spatial awareness is a very relevant aspect in segmentation tasks, this technique is leaving aside situations in which uncertainty occurs due to alterations in image quality or other non-structural factors. They stress that the resulting label probabilities for each class are similar to one-hot within homogeneous areas and thus preserve high confidence in non-ambiguous regions while uncertainty is captured near object boundaries. Our proposed method achieves the same effect but generalized to different sources of uncertainty by selectively maximizing the entropy only for ambiguous or difficult to classify pixels.

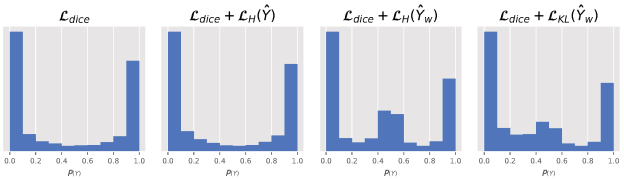


Figure 2. Distribution of the magnitude of softmax probabilities on Left Atrium dataset obtained when training with different loss functions. These plots motivate the proposed *selective* confidence penalty over prior work [10], as the resulting probability distributions produced by our method are smoother, leading to better calibrated networks.

3. Methods

3.1. Preliminaries

Let us have a training dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})_n\}_{1 \leq n \leq |\mathcal{D}|}$, where $\mathbf{x}_n \in \mathbb{R}^{\Omega_n}$ denotes an input image and $\mathbf{y}_n \in \{0,1\}^{\Omega_n \times K}$ its corresponding pixel-wise one-hot label. Ω_n denotes the spatial image domain and K the number of segmentation classes. We aim at training a model, parameterized by θ, which approximates the underlying conditional distribution $p(\mathbf{y}|\mathbf{x}, \theta)$, where θ is chosen to optimize a given loss function. The output of our model, at a given pixel i, is given as $\hat{y}_i$, whose associated class probability is $p(\mathbf{y}|\mathbf{x}, \theta)$. Thus, $p(\hat{y}_{i,k} = k|x_i, \theta)$ will indicate the probability that a given pixel (or voxel) i is assigned to the class $k \in K$. For simplicity, we will denote this probability as $\hat{p}_{i,k}$.

3.2. Maximum Entropy on Erroneous Predictions

Since confident predictions correspond to low entropy output distributions, a network is overconfident when it places all the predicted probability on a single class for each training example, which is often a symptom of overfitting [33]. Therefore, maximizing the entropy of the output probability distribution encourages high uncertainty (or

low confidence) in the network predictions. In contrast to prior work [10], which penalizes low entropy in the entire output distributions, we propose to selectively penalize overconfidence exclusively for those pixels which are misclassified, i.e. the more challenging ones. To motivate our strategy, we plot the distribution of the magnitude of softmax probabilities in Figure 2. It can be observed that for models trained with standard $\mathcal{L}_{dice}$ loss [30], most of the predictions lie in the first or last bin of the histogram, indicating that the network assigns either high or low probability to the pixels, avoiding the use of intermediate values, i.e., high-entropy. We hypothesize that encouraging the network to assign high entropy values to erroneous predictions (i.e. uniformly distributed probabilities) will help to penalize overconfidence in complex scenarios. To this end, for every iteration of stochastic gradient descent during training, given the pixel-level class predictions $\hat{y}_i$ and their associated ground truth class y_i, we define the set of misclassified pixels as $\hat{\mathbf{y}}_w = \{y_i | \hat{y}_i \neq y_i\}$. We can then compute the entropy for this set as:

$$\mathcal{H}(\hat{\mathbf{y}}_w) = -\frac{1}{|\hat{\mathbf{y}}_w|} \sum_{k, i \in \hat{\mathbf{y}}_w} \hat{p}_{i,k} \log \hat{p}_{i,k}, \tag{1}$$

where $|\cdot|$ is used to denote the cardinality of the set. As we aim at maximizing the entropy of the output probabilities $\hat{\mathbf{y}}_w$ (eq. (1)), this equals to minimizing the negative entropy, i.e., $\min_\theta -\mathcal{H}(\hat{\mathbf{y}}_w)$. Note that given a uniform distribution $\mathbf{q}$, maximizing the entropy of $\mathbf{y}_w$ boils down to minimizing the *Kullback-Leibler* (KL) divergence between $\mathbf{y}_w$ and $\mathbf{q}$. From now, we will use $\mathcal{L}_{\mathcal{H}}(\hat{\mathbf{y}}_w) = \mathcal{H}(\hat{\mathbf{y}}_w)$ to refer to the additional loss term which computes the entropy for the missclassified pixels follwing equation 1.

Proxy for entropy maximization: In addition to explicitly maximizing the entropy of predictions (or to minimizing the negative entropy) as proposed in Equation 1, we resort to an alternative regularizer, which is a variant of the KL divergence, as presented in [34]. The idea is to encourage the output probabilities in $\mathbf{y}_w$ (the misclassified pixels) to be close to the uniform distribution (i.e. all elements in the probability simplex vector q are equal to $\frac{1}{K}$), resulting in max-uncertainty. This term can be expressed as:

$$\mathcal{D}_{KL}(\mathbf{q} || \hat{\mathbf{y}}_w) \overset{\text{K}}{=} \mathcal{H}(\mathbf{q}, \hat{\mathbf{y}}_w) \tag{2}$$

with q being the uniform distriubtion and the symbol $\overset{\text{K}}{=}$ representing equality up to an additive or multiplicative constant asociated with the number of classes. We refer the reader to the Appendix I in [34] for the Proof of this KL divergence variant, as well as its gradients. It is important to note that despite both terms, (1) and (2), push $\hat{\mathbf{y}}_w$ towards a uniform distribution, their gradient dynamics are different, and thus the effect on the weight updates differs. Here we perform an experimental analysis to assess which term leads to better performance. We will use $\mathcal{L}_{KL}(\hat{\mathbf{y}}_w) = \mathcal{D}_{KL}(\mathbf{q} || \hat{\mathbf{y}}_w)$ to refer to the additional loss term which employs the KL divergence as a proxy for entropy maximitation, as defined in eq. 2.

Global learning objective: Our final loss function takes the following form:

$$\mathcal{L} = \mathcal{L}_{Seg}(\mathbf{y}, \hat{\mathbf{y}}) - \lambda \mathcal{L}_{me}(\hat{\mathbf{y}}_w) \tag{3}$$

where $\hat{\mathbf{y}}$ is the entire set of pixel predictions, $\mathcal{L}_{Seg}$ the segmentation loss[1], $\mathcal{L}_{me}$ is one of the proposed maximum entropy regularization terms and λ balances the importance of each objective. Note that $\mathcal{L}_{me}$ can be defined in two ways, depending on whether we consider the standard entropy definition (i.e. we define $\mathcal{L}_{me}(\hat{\mathbf{y}}_w) = \mathcal{L}_H(\hat{\mathbf{y}}_w)$ based on eq. (1)) or the proxy for entropy maximization using the KL divergence (i.e. we set $\mathcal{L}_{me}(\hat{\mathbf{y}}_w) = \mathcal{L}_{KL}(\hat{\mathbf{y}}_w)$ While the first term on the right side of eq. 3 will account for producing good quality segmentations at the pixel level, the intuition behind the second term is that penalizing overconfident predictions only for challenging pixels should increase the awareness of the model about the more uncertain image regions, maintaining high confidence in regions that are actually identified correctly.

3.3. Baseline models

We trained different models to benchmark our proposed approach. As baseline, we trained the segmentation networks using a simple loss composed of a single segmentation error term $\mathcal{L}_{Seg}$, without adding any regularization term. For this purpose, we used two popular segmentation losses which are dominating the literature on semantic segmentation: cross-entropy loss ($\mathcal{L}_{CE}$) and the negative soft Dice coefficient known as soft Dice loss ($\mathcal{L}_{dice}$) as defined by [30]. Furthermore, we also compare our method to state-of-the-art approaches that have proven to provide better calibrated deep neural networks.

First, due to its similarity with our work, we include the confidence penalty loss proposed in [10], which discourages *all* the neural network predictions from being overconfident by penalizing low-entropy distributions. This is achieved by adding a low-entropy penalty term over all the pixels (in contrast with our method which only penalizes the missclassified pixels), which can be defined as:

$$\mathcal{L}_H(\hat{\mathbf{y}}) = -\frac{1}{|\hat{\mathbf{y}}|} \sum_{k, i \in \hat{\mathbf{y}}} \hat{p}_{i,k} \log \hat{p}_{i,k}. \tag{4}$$

We train two baseline models using the aforementioned regularizer $\mathcal{L}_H(\hat{\mathbf{y}})$: one considering cross-entropy ($\mathcal{L}_{CE}$) as the segmentation loss and another one using the Dice loss

[1] $\mathcal{L}_{Seg}$ can take the form of any segmentation loss (e.g., CE or Dice)

($\mathcal{L}_{dice}$). We also assess the performance of focal-loss [22], since recent findings [9] demonstrated the benefits of using this objective to train well-calibrated networks. Interestingly, authors observed that focal loss minimizes a KL divergence between the predicted softmax distribution and the target distribution, while increasing the entropy of the predicted distribution. Thus, the form of the learning objective, and in particular the regularizer, has the same spirit than our proposed term:

$$\mathcal{L}_{\mathcal{FL}}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k,i \in \mathbf{y}} \mathbb{1}[y_i = k]\Big(1 - \hat{p}_{i,k}\Big)^{\gamma} \log \hat{p}_{i,k} \quad (5)$$

where γ is the focusing hyper-parameter[2], and $\mathbb{1}[y_i = k]$ the indicator function, whose value is 1 when y_i (the ground truth class for pixel i) is equal to k, and 0 otherwise The intuition behind the modulating factor is to reduce the contribution of easy samples into the total loss.

4. Experiments and results

Dataset details. We benchmark the proposed method in the context of Left Atrial (LA) cavity and White Matter Hyperintensities (WMH) segmentation in MR images. For LA, we used the Atrial Segmentation Challenge dataset [35], which provides 100 3D gadolinium-enhanced MR imaging scans (GE-MRIs) and LA segmentation masks for training and validation. These scans have an isotropic resolution of $0.625 \times 0.625 \times 0.625\text{mm}^3$. We used the splits and preprocessed data from [36] where the scans were split into 80 for training and 20 for evaluation (5% of traininig images were used for validation). The WMH dataset [37] consists of 60 MR images with binary masks indicating the presence of WMH lesions. For each subject, co-registered 3D T1-weighted and a 2D multi-slice FLAIR images were provided, with images having a resolution of 1mm $\times$ 1mm $\times$ 3mm. We split the dataset into independent training (42), validation (3) and test (15) sets.

Segmentation network details. We benchmark our proposed method with two state-of-the-art DNN architectures (UNet [38] and ResUNet [39]) which were implemented using Tensorflow 2.3. During training, for the WMH dataset we extract patches of size $64 \times 64 \times 64$ from each volume, and we train the networks until convergence by randomly sampling patches so that the central pixel corresponds to foreground label with 0.9 probability. This patch based sampling strategy is used in highly imbalanced scenarios like WMH, where most of the pixels correspond to the background class while just a few of them are associated to foreground (lesion). For LA dataset all the scans were cropped to size $144 \times 144 \times 80$ during training, and centered at the heart region during test for better comparison of the segmentation performance. We used Adam optimizer with a batch size of 64 for WMH and 2 for LA. The initial learning rate was set to 0.0001, and it was reduced by a factor of 0.85 every 10 epochs. Hyper-parameters were chosen using the validation split, and results reported on the hold-out test set.

Training the baseline models. To benchmark our proposed method we trained the networks introduced in the previous section with different configurations in their learning objective. As baselines, we use the standard cross-entropy loss ($\mathcal{L}_{CE}$), and the negative of Dice coefficient [30] ($\mathcal{L}_{dice}$), which have been widely employed in medical image segmentation. To compare with more sophisticated methods we implemented the confidence penalty-based method [10] detailed in section 3.3 by adding the entropy penalizer $\mathcal{L}_H(\hat{\mathbf{y}})$ to the segmentation loss function, and using the hyper-parameter $\beta = 0.2$ suggested by the authors. We also include the focal-loss (see eq. 5) ($\mathcal{L}_{FL}$) with $\gamma = 2$, following the authors' findings. We compare these loss functions with the proposed regularizers. In particular, both $\mathcal{L}_{CE}$ and $\mathcal{L}_{dice}$ are employed as segmentation loss functions, and combined with the two proposed terms to penalize low entropy in wrongly classified pixels: $\mathcal{L}_H(\hat{\mathbf{y}}_w)$ (eq. 1) and $\mathcal{L}_{KL}(\hat{\mathbf{y}}_w)$ (eq. 2). We performed a grid search using different values of λ, and we found empirically for the WMH that 0.3 works best for models trained with $\mathcal{L}_{CE}$ and 1.0 for models trained with $\mathcal{L}_{dice}$. For the LA dataset, we chose 0.1 when combined with $\mathcal{L}_{CE}$ and 0.5 for models trained using $\mathcal{L}_{dice}$. For each setting we trained 3 models and report the average results.

4.1. Evaluation metrics

To assess segmentation performance we resort to the standard Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD), widely used in medical image segmentation. We also included metrics which are affected by model calibration, namely the Brier score, Stratified Brier score and Expected Calibration Error.

Brier score. Brier score [6] is a proper scoring rule whose optimal value corresponds to a perfect prediction. In other words, a system that is both perfectly calibrated and perfectly discriminative will have a Brier score of zero. In the context of image segmentation, for an image with N pixels (voxels), the Brier score can be defined as:

$$Br = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|K|}\sum_{k=1}^{K}\Big(\hat{p}_{i,k} - \mathbb{1}[y_i = k]\Big)^2. \quad (6)$$

Stratified Brier Score. In problems with highly imbalanced classes, e.g., brain lesion segmentation, calibration may be good overall but poor for the minority class. In this

[2]Note that when $\gamma = 0$, $\mathcal{L}_{\mathcal{FL}}$ is equivalent to $\mathcal{L}_{\mathcal{CE}}$.

Training loss		Segmentation performance				Calibration performance					
		dice coefficient		HD		Brier (1^{-4})		Brier^{+}		ECE (1^{-3})	
		WMH	LA	WMH	LA	WMH	LA	WMH	LA	WMH	LA
$\mathcal{L}_{dice}$	–	**0.770 (0.100)**	**0.886 (0.060)**	24.041 (10.845)	**28.282 (11.316)**	6.717 (4.184)	29.182(15.068)	0.257 (0.125)	0.107 (0.090)	0.667 (0.414)	28.861 (15.009)
	$+\mathcal{L}_H(\hat{Y})$ [10]	0.769 (0.099)	0.885 (0.050)	21.608 (8.830)	29.811 (11.168)	6.751 (4.194)	29.019(12.709)	0.249 (0.125)	0.109 (0.077)	0.670 (0.415)	28.458 (12.514)
	$+\mathcal{L}_H(\hat{Y}_w)$	0.758 (0.108)	0.873 (0.069)	21.243 (8.755)	29.374 (10.965)	5.874 (3.875)	24.709(13.774)	0.244 (0.124)	0.103 (0.086)	0.510 (0.350)	18.796 (15.005)
	$+\mathcal{L}_{KL}(\hat{Y}_w)$	**0.770 (0.098)**	0.881 (0.064)	**20.804 (8.122)**	28.415 (12.860)	**5.564 (3.586)**	**23.182(12.464)**	**0.231 (0.114)**	**0.095 (0.077)**	**0.471 (0.318)**	**15.587 (13.391)**
$\mathcal{L}_{CE}$	–	0.755 (0.111)	0.878 (0.070)	21.236 (7.735)	**27.163 (11.967)**	6.462 (4.141)	24.447 (14.876)	0.280 (0.140)	0.108 (0.092)	0.620 (0.400)	18.383 (16.700)
	$+\mathcal{L}_H(\hat{Y})$ [10]	0.760 (0.109)	0.881 (0.070)	23.124 (9.523)	29.464 (14.389)	6.369 (4.018)	23.539 (11.903)	0.242 (0.125)	0.096 (0.070)	4.100 (0.582)	15.590 (14.002)
	$+\mathcal{L}_H(\hat{Y}_w)$	0.770 (0.095)	**0.883 (0.058)**	**19.544 (7.254)**	28.560(13.352)	5.417 (3.547)	**22.506 (11.903)**	0.217 (0.104)	**0.093 (0.071)**	0.436 (0.301)	**15.242 (13.730)**
	$+\mathcal{L}_{KL}(\hat{Y}_w)$	**0.777 (0.093)**	0.876 (0.070)	22.298 (9.566)	28.736 (11.972)	**5.331 (3.478)**	24.085 (13.330)	**0.213 (0.099)**	0.105 (0.090)	**0.422 (0.289)**	17.348 (14.786)
	$\mathcal{L}_{FL}$	0.753 (0.113)	0.881 (0.064)	21.931 (8.167)	28.599 (11.968)	5.760 (3.732)	23.928 (11.626)	0.243 (0.130)	0.095 (0.066)	0.438 (0.310)	25.998 (12.740)

Table 1. Mean accuracy and standard deviation for both WMH and LA segmentation tasks with UNet as backbone. Our models are gray-shadowed and best results are highlighted in bold.

case, the majority class will dominate and miscalibration in the class of interest will not be reflected in the standard Brier score. In [40], the authors proposed the stratified Brier score to measure calibration in binary classification problems with high imbalance. In this work, we follow [20], where stratified Brier score was adopted for image segmentation. We compute the stratified Brier score individually per-class, treating every structure of interest as a binary segmentation problem. For a given image with ground truth segmentation $\mathbf{y}$, we construct the *stratified* Brier score for the class k, Br^k, by computing it only in the subset of pixels $\mathcal{S}_k = \{s : y_s = k\}$, i.e. pixels whose ground truth label is k. The problem is therefore binarized considering all the other classes within a single background class. The formulation of the stratified Brier score Br^k is given by:

$$Br^k = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \left(\hat{p}_{i,k} - \mathbb{1}[y_i = k] \right)^2. \tag{7}$$

Expected Calibration Error and reliability diagrams. Another measure widely adopted to quantify model calibration, which can be used as a complement to Brier score is the Expected Calibration Error (ECE) [8]. This metric directly assess the predictive power of the models by analyzing test examples confidence against the observed frequency of positive examples. To this end, the interval $[0, 1]$ is divided into M equispaced bins, where the i^{th} bin is the interval $\left(\frac{i-1}{M}, \frac{i}{M}\right]$, and B_i denote the set of samples with confidence belonging to the i^{th} bin. Then, for each bin, the frequency of positive examples of B_i is computed as $freq(B_i) = \frac{1}{|B_i|} \sum_{i \in B_i} \mathbb{1}[y_i = k]$while the confidence $C(B_i)$ of the i^{th} bin is computed as $C(B_i) = \frac{1}{|B_i|} \sum_{i \in B_i} \hat{p}_{i,k}$. To summarize these statistics we calculate the ECE as:

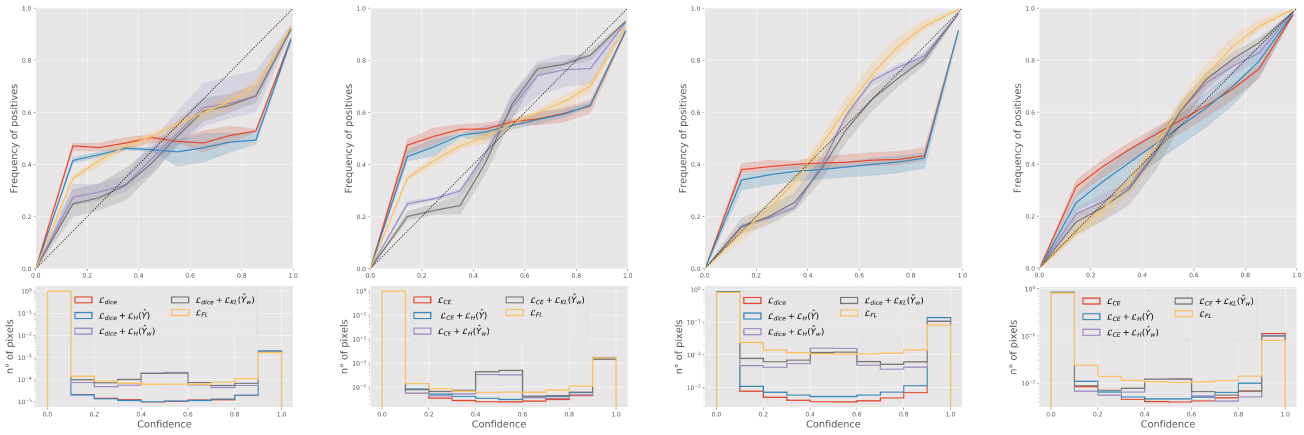
$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| freq(B_m) - C(B_m) \right|$$

We also generate *reliability diagrams* by plotting the observed frequency as a function of the class probability. In a perfectly calibrated model, the frequency on each bin matches the confidence, and hence all of the bars lie on the diagonal.

5. Results

Segmentation performance. The main goal of the proposed methodology is to improve the estimated uncertainty of the predictions, while retaining the segmentation power of original losses. Thus, we first assess whether integrating our regularization terms leads to a performance degradation. Table 1 reports the results across the different datasets with UNet as backbone architecture. First, we can observe that adding the proposed regularizers does not result in a remarkable loss of segmentation performance. Indeed, in some cases, e.g., $\mathcal{L}_{dice} + \mathcal{L}_{KL}(\hat{Y}_w)$ in WMH, the proposed model outperforms the baseline by more than 3% in terms of HD. Furthermore, this behaviour holds when the standard CE loss is used in conjunction with the proposed terms, suggesting that the overall segmentation performance is not negatively impacted by adding our regularizers into the main learning objective. Last, it is noteworthy to mention that even though focal loss sometimes outperforms the baselines, it typically falls behind the proposed two-terms losses.

Calibration performance. Our main focus in this section is to evaluate the calibration quality of a network that is trained with different losses. Recent empirical evidence [23] shows that, despite leading to strong predictive models, CE and specially Dice losses result in highly-confident predictions. The results obtained for calibration metrics (i.e. Brier and ECE, reported in the right-side of Table 1), are in line with these observations. These results evidence that regardless of the dataset, networks trained with any of these losses as a single objective, lead to the worst calibrated models. Explicitly penalizing low-entropy predictions over all the pixels, as in [10], typically improves the calibration metrics. Nevertheless, despite the gains observed with [10],



(a) WMH segmentation (b) WMH segmentation (c) LA segmentation (d) LA segmentation

Figure 3. The first row shows the reliability plot calculated on the entire volume of test images for each of the models while the bottom row shows the probability distributions produced by each method.

Training loss		Segmentation performance				Calibration performance					
		dice coefficient		HD		Brier (1^{-4})		Brier^{+}		ECE (1^{-3})	
		WMH	LA	WMH	LA	WMH	LA	WMH	LA	WMH	LA
$\mathcal{L}_{dice}$	–	0.768 (0.108)	**0.905 (0.024)**	20.499 (8.468)	24.311 (7.733)	6.742 (4.346)	24.940 (7.692)	0.258 (0.136)	0.092 (0.043)	0.667 (0.429)	24.698 (7.606)
	$+\mathcal{L}_H(\hat{Y})$	0.754 (0.122)	0.903 (0.029)	21.089 (7.475)	23.811 (8.902)	7.013 (4.643)	24.952 (8.964)	0.267 (0.157)	0.086 (0.051)	0.696 (0.461)	24.457 (8.765)
	$+\mathcal{L}_H(\hat{Y}_w)$	**0.786 (0.089)**	0.903 (0.025)	20.033 (7.566)	24.095 (8.357)	5.451 (3.492)	**19.565 (6.493)**	0.183 (0.095)	0.083 (0.036)	0.451 (0.287)	13.006 (5.160)
	$+\mathcal{L}_{KL}(\hat{Y}_w)$	**0.786 (0.093)**	0.900 (0.028)	**18.848 (6.513)**	**23.600 (6.496)**	**5.379 (3.502)**	20.106 (6.971)	**0.174 (0.094)**	**0.079 (0.042)**	**0.434 (0.285)**	**12.561 (6.222)**
$\mathcal{L}_{CE}$	–	0.770 (0.104)	0.890 (0.035)	18.928 (7.175)	26.596 (8.121)	6.256 (4.044)	22.458 (8.417)	0.259 (0.128)	0.113 (0.055)	0.602 (0.390)	16.700 (8.763)
	$+\mathcal{L}_H(\hat{Y})$	0.778 (0.092)	**0.896 (0.030)**	**18.554 (7.214)**	**25.137 (8.291)**	6.208 (3.842)	20.933 (7.552)	0.227 (0.103)	0.098 (0.044)	5.251 (0.504)	**11.784 (6.481)**
	$+\mathcal{L}_H(\hat{Y}_w)$	**0.779 (0.096)**	0.890 (0.036)	18.789 (7.205)	25.349 (6.265)	**5.169 (3.347)**	21.721 (7.639)	0.191 (0.093)	0.106 (0.053)	0.396 (0.257)	14.607 (7.872)
	$+\mathcal{L}_{KL}(\hat{Y}_w)$	0.775 (0.095)	0.895 (0.032)	20.949 (9.769)	25.576 (7.426)	5.269 (3.383)	**20.609 (7.260)**	**0.187 (0.091)**	**0.097 (0.047)**	**0.390 (0.248)**	12.811 (7.250)
	$\mathcal{L}_{FL}$	0.780 (0.090)	0.891 (0.031)	19.759 (7.372)	26.447 (7.442)	5.621 (3.703)	21.269 (6.681)	0.216 (0.103)	0.102 (0.041)	0.472 (0.327)	14.249 (6.140)

Table 2. Mean accuracy and standard deviation for both WMH and LS segmentation tasks with ResUNet as backbone. Our models are gray-shadowed and best results are highlighted in bold.

empirical results demonstrate that penalizing low-entropy values *only over misclassified pixels* brings the largest improvements, regardless of the main segmentation loss used. In particular, the proposed MEEP regularization terms outperform the two baselines in all the three calibration metrics and in both datasets, with improvements ranging from 1% to 13%.

When evaluating the proposed MEEP regularizers ($\mathcal{L}_{KL}(\hat{Y}_w)$ and $\mathcal{L}_H(\hat{Y}_w)$) combined with the segmentation losses based on DSC and CE, we observe that DSC with $\mathcal{L}_{KL}(\hat{Y}_w)$ consistently achieves better performance in most of the cases. However, for CE, both regularizers alternate best results, which depends on the dataset used. We hypothesize that this might be due to the different gradient dynamics shown by the two regularizers[3]. In the interval [0.1-0.9], the gradient of the entropy term is almost linear, and therefore penalize similarly all the predictions, regardless of how far they are from the uniform distribution. On the other hand, the gradient of the cross-entropy has a smaller scale on predictions close to 0.5, but it increases as these predictions move away from this value. This suggests that the CE term, and therefore our KL-based regularizer in (2), pushes pixels with very high confidence harder than the entropy regularization term, whereas it penalizes to a less extent predictions close to the uniform distribution.

Regarding the focal loss, even though it improves model calibration when compared with the vanilla models, we observe that the proposed regularizers achieve better calibration metrics. Mukhoti et *al.*, [9] demonstrated that the focal loss has an implicit weight regularizing effect, which is governed by the scale of the gradients. In particular, authors propose that the gradients derived from the focal and standard cross-entropy loss are proportional, with $\frac{\partial \mathcal{L}_F}{\partial \theta} = \frac{\partial \mathcal{L}_{CE}}{\partial \theta} f(\hat{p}_{i_{y_i}}, \gamma)$, where $\hat{p}_{i_{y_i}}$ is the predicted probability of the winner class on a given pixel i and $f(p, \gamma) = (1-p)^\gamma - \gamma p(1-p)^{\gamma-1} - \log(p), \forall \gamma \in \mathbb{R}^+$. This results in highly confidence predictions increasing faster at the beginning of the training due to larger weight norms of

[3]We refer to Fig 3 and App I in [34] for a detailed explanation regarding the different energies for binary classification and their derivatives.

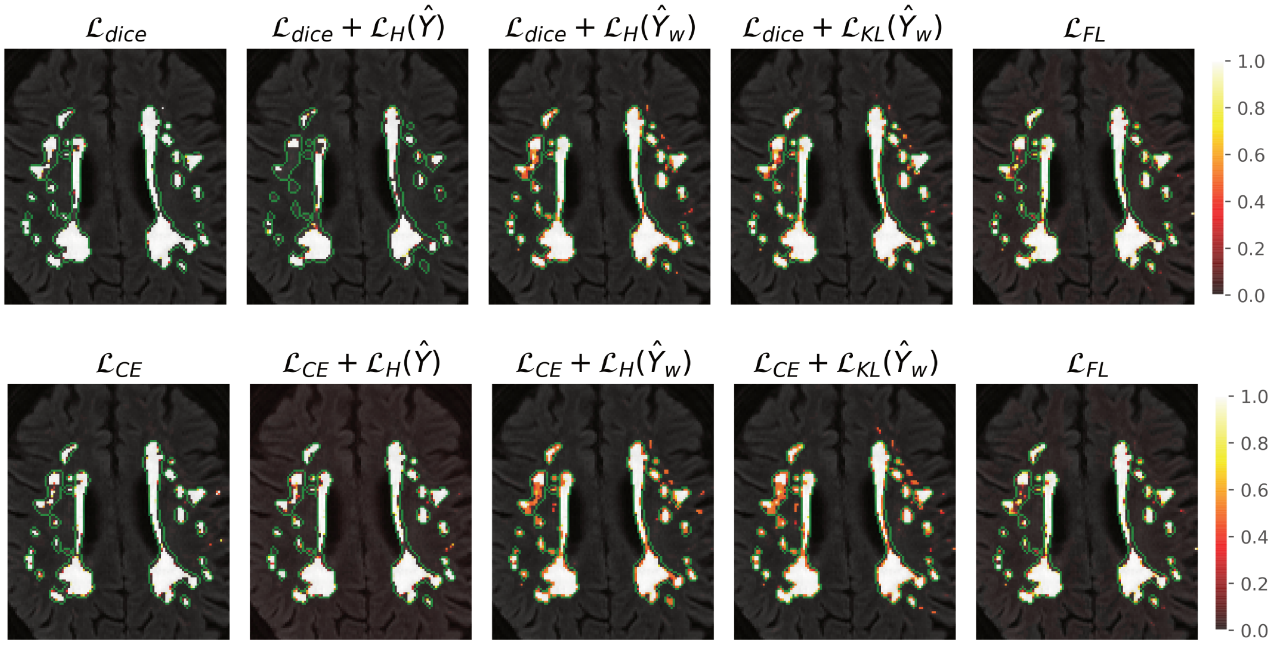


Figure 4. Qualitative evaluation of the proposed method for WMH segmentation.

the model trained with focal loss. As the training evolves and $f(\hat{p}_{i_{y_i}}, \gamma)$ becomes smaller than 1, the scale of the gradient decreases, decreasing also the norm of the weights. We believe, however, that the strong gradients observed at the beginning of the training with focal loss might have indeed an undesirable effect. Indeed, if the probability estimates of certain pixels are pushed towards the vertex of the simplex, and the weight given to the implicit low-entropy penalty term (i.e., γ in focal loss) is not strong enough, these high confident predictions might never recover. On the other hand, if the weight of the regularization term in the whole objective is relatively large, this could lead to trivial solutions. Thus, these results demonstrate empirically that the proposed method presents an efficient alternative to current losses designed for calibration purposes.

Figure 3 depicts the reliability diagrams and the probability distributions produced by each method in both datasets. In the reliability diagrams, curves closer to the diagonal indicates better calibrated networks. Thus, we can easily observe that both of our regularizers, which penalize high-entropy predictions over only the misclassified pixels result in models better calibrated.

Qualitative results. Figures 1 and 4 depict exemplar cases of the probability maps obtained for each loss function configuration. Here, it is worth highlighting the significant improvement achieved in the cases where the DSC is used as the learning objective. As it can be observed, the predictions made by the vanilla network trained with DSC loss tend to be highly overconfident, either assigning probabilities equal to 0 or 1. However, when employing the proposed regularizers, the models tend to use the full range of possible values, assigning scores around 0.5 (marked in red) to the more challenging pixels.

Is our method backbone-agnostic? We repeated the experiments using a ResUNet architecture instead of the standard UNet. Results shown in Table 2 demonstrate that the proposed terms typically lead to improvement on both model calibration and segmentation performance across the two different backbones, main segmentation losses and datasets. This improvement is further stressed for the calibration metrics, where the four variants of our method outperform the rest by a wide margin.

6. Conclusions

In this paper, we have presented a simple yet effective approach to improve the uncertainty estimates inferred from segmentation models when trained with popular segmentation losses. In contrast to prior literature, the proposed regularization terms aim at penalizing high-confident predictions only on misclassified pixels, which increases the network uncertainty in complex scenarios. In addition to directly maximizing the entropy on the set of erroneous pixels, we present a proxy for this term, which is formulated with a KL regularizer modeling high uncertainty over those pixels. Comprehensive results on two popular datasets, segmentation losses, and well-known architectures demonstrate that the proposed models outperform current literature in both discriminative and calibration tasks. Furthermore, even though both terms enforce

the same objective, we hypothesize that differences in performance might stem from the different gradient dynamics. Last, quantitative results suggest that models that perform satisfactorily in segmentation are not necessarily well calibrated, which is of pivotal importance in critical decision-making systems. Thus, we argue that both segmentation and calibration metrics should be preferred when assessing the performance of segmentation models for medical images.

References

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330. 1, 2

[2] D. Karimi and A. Gholipour, "Improving calibration and out-of-distribution detection in medical image segmentation with convolutional neural networks," *arXiv preprint arXiv:2004.06569*, 2020. 1

[3] R. Robinson, V. V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak *et al.*, "Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study," *Journal of Cardiovascular Magnetic Resonance*, vol. 21, no. 1, pp. 1–14, 2019.

[4] S. Czolbe, K. Arnavaz, O. Krause, and A. Feragen, "Is segmentation uncertainty useful?" in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 715–726. 1, 3

[5] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association*, vol. 102, no. 477, pp. 359–378, 2007. 1

[6] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950. 1, 5

[7] J. Bröcker and L. A. Smith, "Increasing the reliability of reliability diagrams," *Weather and forecasting*, vol. 22, no. 3, pp. 651–661, 2007. 1

[8] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 1, 2, 6

[9] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. Torr, and P. K. Dokania, "Calibrating deep neural networks using focal loss," *arXiv preprint arXiv:2002.09437*, 2020. 1, 2, 5, 7

[10] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017. 2, 3, 4, 5, 6

[11] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers," in *ICML*, vol. 1, 2001, pp. 609–616. 2

[12] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*, 2015, pp. 1613–1622. 2

[13] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059. 2, 3

[14] J. M. Hernández-Lobato and R. Adams, "Probabilistic backpropagation for scalable learning of bayesian neural networks," in *International conference on machine learning*, 2015, pp. 1861–1869. 2

[15] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *arXiv preprint arXiv:1612.01474*, 2016. 2, 3

[16] A. C. Stickland and I. Murray, "Diverse ensembles improve calibration," in *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*, 2020. 2

[17] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *NeurIPS*, 2019. 2

[18] Y. Wen, D. Tran, and J. Ba, "Batchensemble: an alternative approach to efficient ensemble and lifelong learning," in *ICLR*, 2020. 2

[19] F. Wenzel, J. Snoek, D. Tran, and R. Jenatton, "Hyperparameter ensembles for robustness and uncertainty quantification," in *NeurIPS*, 2020. 2

[20] A. J. Larrazabal, C. Martínez, J. Dolz, and E. Ferrante, "Orthogonal ensemble networks for biomedical image segmentation," *arXiv preprint arXiv:2105.10827*, 2021. 2, 3, 6

[21] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" *arXiv preprint arXiv:1906.02629*, 2019. 2

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988. 3, 5

[23] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 3868–3878, 2020. 3, 6

[24] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *British Machine Vision Conference 2017, BMVC*, 2017. 3

[25] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, "Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation," in *Medical Imaging with Deep Learning*, 2018. 3

[26] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical image analysis*, vol. 59, p. 101557, 2020. 3

[27] R. Jena and S. P. Awate, "A bayesian neural net to segment images with uncertainty estimates and good calibration," in *International Conference on Information Processing in Medical Imaging*, 2019, pp. 3–15. 3

[28] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep ensembles: A loss landscape perspective," *arXiv preprint arXiv:1912.02757*, 2019. 3

[29] A. Jungo, F. Balsiger, and M. Reyes, "Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation," *Frontiers in neuroscience*, vol. 14, p. 282, 2020. 3

[30] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571. 3, 4, 5

[31] J. Sander, B. D. de Vos, J. M. Wolterink, and I. Išgum, "Towards increased trustworthiness of deep learning segmentation methods on cardiac mri," in *Medical Imaging 2019: Image Processing*, vol. 10949. International Society for Optics and Photonics, 2019, p. 1094919. 3

[32] M. Islam and B. Glocker, "Spatially varying label smoothing: Capturing uncertainty from expert annotations," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 677–688. 3

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826. 3

[34] S. Belharbi, J. Rony, J. Dolz, I. B. Ayed, L. McCaffrey, and E. Granger, "Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty," *arXiv preprint arXiv:2011.07221*, 2020. 4, 7

[35] Z. Xiong, Q. Xia, Z. Hu, N. Huang, S. Vesal, N. Ravikumar, A. Maier, C. Li, Q. Tong, W. Si *et al.*, "A global benchmark of algorithms for segmenting late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical Image Analysis*, 2020. 5

[36] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *MICCAI*, 2019. 5

[37] H. J. Kuijf, J. M. Biesbroek, J. De Bresser, R. Heinen, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, M. J. Cardoso, A. Casamitjana *et al.*, "Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge," *IEEE transactions on medical imaging*, vol. 38, no. 11, pp. 2556–2568, 2019. 5

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of MICCAI*, 2015. 5

[39] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018. 5

[40] B. C. Wallace and I. J. Dahabreh, "Improving class probability estimates for imbalanced data," *Knowledge and information systems*, vol. 41, no. 1, pp. 33–52, 2014. 6

Anexo E

Video-oculography eye tracking towards clinical applications: A review

Computers in Biology and Medicine 108 (2019) 57–66

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Video-oculography eye tracking towards clinical applications: A review

A.J. Larrazabal[a,*], C.E. García Cena[b], C.E. Martínez[a]

[a] Research Institute for Signals, Systems and Computational Intelligence, Sinc(i), FICH-UNL/CONICET, Ruta Nac. No 168, Km 472.4 (3000), Santa Fe, Argentina
[b] Centre for Robotics and Automation, UPM-CSIC, José Gutiérrez Abascal Street, 28006 Madrid, Spain

ARTICLE INFO

Keywords:
Eye tracking
Eye gazing
Disease diagnoses
Head movements
Saccadic movements

ABSTRACT

Most neurological diseases are usually accompanied by a broad spectrum of oculomotor alterations. Being able to record and analyze these different types of eye movements would be a valuable tool to understand the functional integrity of brain structures. Nowadays, video-oculography is the most widely used eye-movements assessing method. This paper presents a study of the existing eye tracking video-oculography techniques and also analyzes the importance of measuring slight head movements for diseases diagnosis. In particular, two types of methods are reviewed and compared, including appearance-based and feature-based methods which are further subdivided into 2D-mapping and 3D model-based approaches. In order to demonstrate the advantages and disadvantages of these different eye tracking methods for disease diagnosis, a series of comparisons are conducted between them, addressing the complexity of the system, the accuracy achieved, the ability to measure head movements and the external conditions for which they have been designed. Lastly, it also highlights the open challenges in this research field and discusses possible future directions.

1. Introduction

Neurodegenerative motor disorders are usually accompanied by a broad spectrum of oculomotor alterations. Studies have shown that activity related to eye movements is observed in cortical and subcortical areas, which are directly and indirectly connected with several neural systems which interact among each other to control the suitable performance of the ocular and ocular-cephalic movements. For instance, brain basal ganglia, brainstem nuclei and the vestibular system are organized to produce different eye movements [1,2].

Therefore, an accurate and detailed eye movements analysis would be a key experimental tool for understanding the functional integrity of brain structures involved in motor and cognitive processing. In addition, it becomes a unique opportunity to detect the presence of different injuries in the nervous center thanks to the existing connection with different oculomotor control abnormalities. These injuries involve a wide variety of neurological diseases including parkinsonian syndromes [3,4] amyotrophic lateral sclerosis [5], Huntingtons disease [6], Alzheimers disease [7], minimal hepatic encephalopathy [8], among others.

Various types of eye movements are performed in people's daily lives, usually without being aware of them. These movements can be divided into two functional classes [2]. The first class is the one in charge of making the images remain fixed in the retina and comprises the next movements: vestibulo-ocular reflexes (VORs) which make the direction of the eyes remain constant when the head is moved; fixation system which make the gaze resting on a small predefined area; smooth pursuit which describe the eye following a moving object, and optokinetic nystagmus, which stabilize the fovea in relation to objects in the surrounding environment. Some variables related to fixation are commonly measured in different studies including total and mean fixation duration, fixation sequences and fixation rate.

In addition to this, humans use a second class of eye movement, named saccades, in order to rapidly shift the fovea in a stepwise manner onto a new target and bring its superior visual accuracy to bear on objects of interest during fixation. Several of these movements are made each second and they are an intrinsic part of the constant cycle of perception, action and cognition. Measurable saccade related parameters include saccade number, amplitude, fixation-saccade ratio and velocity peak detection. Being able to measure them is also an important contribution.

Besides, there is another kind of saccadic movement which is used alongside fixations. If the eyes remained completely fixed at one point for a long time, the retina would adapt to the constant input and induce the visual image to slowly fades away. In order to avoid the neural adaptation, short saccadic movements, known as microsaccades, shift the image on the retina back and forth in an involuntary manner in small magnitudes ranging from 3 min of arc to 1° [9].

* Corresponding author.
E-mail address: alarrazabal@sinc.unl.edu.ar (A.J. Larrazabal).

https://doi.org/10.1016/j.compbiomed.2019.03.025
Received 19 December 2018; Received in revised form 20 February 2019; Accepted 26 March 2019

Saccadic and microsaccadic eye movements are likely to be affected by cognitive impairments, as well as by dysfunctions related purely to oculomotor execution. Because of that, they have been extensively studied for a wide range of applications including the drowsiness detection [10], neurological disease diagnosing and sleep disorders studies [11]. On the other hand, within the field of clinical research, fixations are often analyzed in neuroscience, autism alteration studies [12], and psychological studies to determine a persons focus and level of attention [13].

Over the past 20 years, some of these measured variables of eye movements have been selected as possible markers for differential diagnosis including in pre-symptomatic individuals. Being furthermore eye tracking a less invasive and low cost clinical test compared with other diagnosis methods, this technique has become an invaluable tool for clinicians in diagnosis of neurodegenerative disorders.

Some eye movements abnormalities can be clinically assessed by trained doctors using, for example, Frenzel glasses or ophthalmoscopes [14]. Instead, in some cases it is necessary to make really accurate measurements, like metrics of saccadic accuracy, latencies with respect to stimulus onset, or eye velocity peak estimations which are not possible to be done by simple examination. Furthermore, for some tests with cognitive impairment, it is desirable to present stimulus under specific conditions such as defined target positions, which is also made difficult without the assistance of a synchronized device.

However, eye movements can be easily measured and recorded in the laboratory, covering the main necessities of accurate information. These recordings are highly useful for objective and precise identification of disease status and monitoring of disease progression. For example, increased error rate of antisaccades, which indicates cognitive dysfunction, can only be detected in the laboratory [2]. A number of different eye movements assessing methods have been developed and some of them are currently used. Some examples are electro-oculography [15], scleral search coil system [16], and video-oculography (VOG) [11]. In recent years, the latter has become the most widely used, as it is the only one considered non-invasive and allows for easy coordination of test design and stimuli provision that make it possible to automatically analyze the data.

VOG real-time eye detection and eye tracking is an active area of research in computer vision community since a long time. Presently, there are numerous of gaze trackers devices and software in the market, and there are more and more applications in which these techniques make an important contribution. Human-computer interaction [17,18], driving safety applications [19], pilot training [20], market and marketing research [21], studies of perception, attention and learning disorders [22] are some examples of these. But, while the idea of eye tracking exists for a long time, recent technological advances enable more precise quantification and automated evaluation making this technology broadly available to disease diagnosis [23,24].

Naturally, since there are so many possible application regarding eye movements, a lot of different approaches have been proposed for VOG gaze tracking. Although there already exist some papers that review the current gaze tracker systems or methods [11,25,26], as far as we know, our review is the only one specifically aimed at analyzing and comparing the VOG gaze estimation methods for application to neurological disease diagnosis and research, taking into account their own requirements and use conditions. In addition, the importance of measuring slight head movement present in some neurological diseases is also analyzed.

2. Eye tracking video oculography techniques

Technically, gaze tracking is the procedure of determining the point-of-gaze (POG) -where one is looking-on some monitor or screen or the visual axis of the eye in the 3D space. For this, video oculography devices are equipped at least with one or more video cameras that send images to a personal computer for image processing.

Recently, several approaches to gaze estimation have been reported in the literature. Although they have been studied focusing on different use conditions, in general, they can be classified in two categories: appearance-based and feature-based methods. Clearly, their choice depends on the application for which they have been designed. The quality of the camera and hardware, the required accuracy, the environmental conditions, the desired cost and the freedom of head movements are some factors that determine the approach to be followed. In the following, these two methods are going to be introduced, in order to compare their specifications with those required for medical applications.

2.1. Appearance-based methods

In recent years more and more applications have been developed to analyze human behavior in everyday situations. To do this, it is necessary to monitor the eye movements in uncontrolled scenarios where it is impossible to adjust the lighting conditions, perform a calibration or request the user's assistance. In this context, it is critical to design very robust methods for the different types and qualities of the images.

Fortunately, for most applications very high accuracy is usually not required. For instance, in environmental control or eye typing, where only a few buttons need to be activated, it may be more important to reduce costs by using web cameras, allowing easy and flexible hardware configurations, and avoiding the use of lighting systems and feature detection algorithms.

On the basis of this new approach, several papers [27–30] have presented methods that work with low-resolution images in different environmental conditions in which appearance-based methods seem to be a promising option. These methods address the gaze estimation problem by learning a mapping function directly from eye images to gaze directions. As input, they use all eye regions pixel values as high-dimensional feature vectors for estimating gaze directions [31]. The output, gaze direction, can be represented as the coordinates (x, y) on the screen where the gaze falls or the rotation angles of the eye with respect to the head position. For low quality images, this is a great advantage compared to the techniques employed by feature-based methods, which have to segment and analyze geometrically derived eye features from high-resolution observations as will be seen in the next section.

The mapping function allows to relate the raw input image with the coordinates of the gaze direction. These functions do not address any particular model, but are designed ad-hoc and are trained with eye images of known gaze direction using various regression techniques, including neural networks [32–34], local interpolation [35,36], or Gaussian process [37,38]. Its formulation depends on the regression technique followed. Fig. 1 shows an example where a convolutional neural network (CNN) is used as a mapping function.

These approaches make the system less restrictive, and even though the precision is not good enough for certain applications, they are very robust even when they are applied to relatively low-resolution cameras or under natural illumination, such as with a phone or computer applications or human computer interaction.

The main problem of these methods is that the appearance of an eye

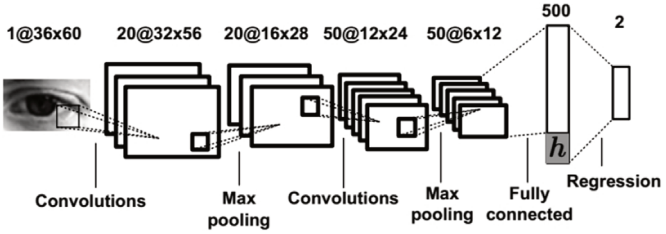


Fig. 1. Architecture of the CNN used as mapping function to predict gaze direction. (From Park et al. [27]).

depends not only upon gaze direction but also upon the head poses, imaging conditions and even on the identities of subjects, making it necessary to generate a person-specific training. In addition, due to the high dimensional feature vectors that must be mapped into the gaze directions, thousands of individual training samples are required to calculate the mapping coefficients.

To overcome these limitations, Sugano et al. [39] propose a learning-by-synthesis approach to appearance-based gaze estimation using a large dataset that contains diverse people, head poses, and gaze directions. Also to avoid the need of a person-specific training, Lu et al. [40] extract more advanced eye features, which help to learn a person-independent relationship between eye gaze change and eye appearance variation. On the other hand, Schneider et al. [41] perform embedding for each person in the training set and then learn a linear transformation that maps out the individual, subject-dependent manifolds avoiding the need of individual calibration.

Despite the recent research progress in the field of computer vision, estimating human gaze directions from only eye appearance is still an open challenge. The performance of appearance-based methods generally depends on the quality and diversity of the training data and generalization ability of the regression algorithm. Moreover, their accuracy is not high enough for clinical uses. For these reasons, appearance-based methods can be ruled out for devices designed for this purpose.

2.2. Feature-based methods

Methods using extracted local features such as contours, eye corners, and eye reflections, called feature-based methods, are the most popular approach for gaze estimation. These methods use geometrically derived eye features from high-resolution eye-images captured by zooming in the user's eyes (See Fig. 2). Once the features are extracted, the connection between the gaze directions and them can be modeled in various ways. Besides, depending on whether they are based on eye geometry or not, these methods can be divided into two main groups: 2D mapping-based gaze estimation methods and 3D model-based gaze estimation methods.

The 3D model-based methods [42,43], directly compute the 3D gaze direction vector from the eye features based on a geometric model of the eye. Then, the point of gaze is estimated by intersecting the gaze direction with the object being viewed, i.e a computer monitor. In order to calculate the center of the cornea and the eye vector, these models require accurate estimation of many user-dependent parameters such as cornea radii, angles between visual and optical axes, the distance between the cornea center and pupil center, among others. To understand why these parameters should be estimated, and which complex hardware calibration should be made during initial setup, the model proposed by Guestrin et al. [44] will be developed. This example is also a good basis for understanding model-based methods. The model and their parameters are shown in Fig. 3.

Considering a ray that comes from the light source I_i, reflects at a point $\mathbf{q}_{i,j}$ on the corneal surface, which is modeled as a convex spherical mirror of radius R, passes through the nodal point of the camera $\mathbf{o}_j$, and intersects the camera image plane at a point $\mathbf{u}_{i,j}$, the next two equations can be formulated:

$$\mathrm{q}_{ij} = \mathbf{o}_j + k_{q,ij}(\mathbf{o}_j - \mathbf{u}_{ij}) \text{ for some } k_{q.ij} \tag{1}$$

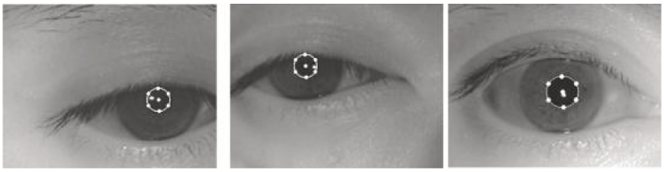


Fig. 2. Features from high-resolution eye-images (from Park et al. [43]).

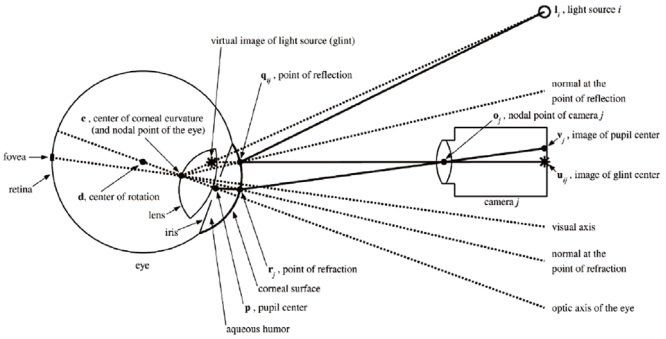


Fig. 3. Schematic representations of the eye, a camera, and a light source (from Guestrin et al. [44]).

$$||\mathbf{q}_{ij} - \mathbf{c}|| = R \tag{2}$$

In addition, based on the beam reflection laws, two more equations can be raised for these points.

$$(\mathbf{i}_i - \mathbf{o}_j) \times (\mathbf{q}_{ij} - \mathbf{o}_j) \bullet (\mathbf{c} - \mathbf{o}_j) = 0 \tag{3}$$

$$(\mathbf{i}_i - \mathbf{q}_{ij}) \bullet (\mathbf{q}_{ij} - \mathbf{c}) \cdot ||\mathbf{o}_j - \mathbf{q}_{ij}|| = (\mathbf{o}_j - \mathbf{q}_{ij}) \bullet (\mathbf{q}_{ij} - \mathbf{c}) \cdot ||\mathbf{i}_i - \mathbf{q}_{ij}|| \tag{4}$$

In the same way, considering a ray that comes from pupil center $\mathbf{p}$, refract at the point $\mathbf{r}_j$ on the corneal surface, passes through the nodal point of camera $\mathbf{o}_j$, and intersects the camera image plane at a point $\mathbf{v}_{ij}$, two more equations can be obtained.

$$\mathbf{r}_{ij} = \mathbf{o}_j + k_{r,j}(\mathbf{o}_j - \mathbf{v}_{ij}) \text{ for some } k_{r,j} \tag{5}$$

$$||\mathbf{r}_{ij} - \mathbf{c}|| = R \tag{6}$$

Then, applying beam refraction laws, the following equations are derived where n_1 and n_2 are the refraction index of the aqueous humor and cornea combined and of air respectively.

$$(\mathbf{r}_j - \mathbf{o}_j) \times (\mathbf{c} - \mathbf{o}_j) \bullet (\mathbf{p} - \mathbf{o}_j) = 0 \tag{7}$$

$$n_1||(\mathbf{r}_j - \mathbf{c}) \times (\mathbf{p} - \mathbf{r}_j)|| \cdot ||\mathbf{o}_j - \mathbf{r}_j|| = n_2||(\mathbf{r}_j - \mathbf{c}) \times (\mathbf{o}_j - \mathbf{r}_j)|| \bullet ||\mathbf{p} - \mathbf{r}_j|| \tag{8}$$

Finally, considering K as the distance between the pupil center and the center of corneal curvature leads to:

$$||\mathbf{p} - \mathbf{c}|| = K \tag{9}$$

By means of solving the proposed system of equations for $\mathbf{c}$ and $\mathbf{p}$, the optic axis of the eye in the space can be reconstructed as the line defined by these two points. It is important to note that to solve these equations, all the subject-specific parameters (R, K and n_1) have to be known. In general, if only one camera is available, they are obtained by the calibration process -detailed below-. Also, the angle between the optic axis and visual axis must be calculated and is usually done during the calibration procedure.

This parameters also rely on metric information requiring camera calibration and exact knowledge of the light sources and monitor position. These values may be directly measured once during the first setup but, to achieve a high accuracy, the eye parameters need to be estimated independently for each individual, making that a previous calibration step cannot be omitted.

The 3D model-based approaches can handle head movements in a robust manner with high accuracy but involving this relatively complex initial setup. They need to use at least a single camera with multiple calibrated light sources [44] or stereo cameras [45–47]. Even so, for some clinical diagnoses, it is important to be able to differentiate between oculocephalic and pure eye movements, so calculating the absolute position of the gaze is not always useful. Furthermore, regardless of the model complexity, the calibration might be only simplified, but

not avoided at all. In some works, to avoid the calibration process, a very simplified eye model is used. While it reduces calibration times and complexity, the accuracy obtained also greatly decreases.

On the other hand, the 2D mapping approaches [23,48,49] are based in finding a mapping function from 2D feature space like Pupil-Center-Corneal-Reflections (PCCR), contours, etc. to gaze point such the computer screen coordinates. That function avoids the need for the direct measurement or estimation of the eye model parameters throughout the system setup. Instead, they are implicitly included in the learning of the mapping function simplifying the setup process itself. The same happens with the camera calibration process and the system geometry determination.

Different features are used as inputs to the mapping function depending on the application and the image conditions. Mostly, they can be further divided into active light techniques such as PCCR or passive light techniques such as shape-based methods, depending on whether they require external light sources to detect eye features.

In the recent years, eye tracking applications using webcams under natural illumination have gained highly relevance in the community. In particular, passive image-based algorithms for eye localizing and tracking in the visible spectrum have been researched over the last years [27,50,51]. These algorithms propose the search for some features like iris or pupil center. For the purpose of iris tracking, the limbus, which is the boundary between the sclera (normally white) and iris (comparatively dark) is optically detected and tracked. Pupil tracking is similar to iris tracking except that a smaller boundary between iris and pupil is used for relative measurement.

Although without active illumination it is easier to segment the limbus due to the higher contrast between the iris and the sclera compared to the contrast between the pupil and the iris, pupil tracking has a lot of advantages. The pupil, which is much less covered by the eyelids than the limbus, enables vertical tracking. In addition, the sharper edge between the pupil and the iris provides a higher resolution.

Various iris and pupil center localization methods have been reported in the literature [52,53]. Several treat iris or pupil center localization as a circle detection or ellipse fitting problem [54–57]. Depending on the viewing angle, both iris and pupil appear elliptical and consequently can be modeled by different shape parameters. Simple ellipse models consist of voting-based methods [58,59] and model fitting methods [60,61]. Once the iris center has been successfully localized, regression-based methods can be used for finding the corresponding gaze points on the screen.

Since these methods directly map the eyes iris center or pupil center location to a target plane such as the monitor screen, the accuracy and robustness of the center localization significantly affect the performance of gaze tracking. For example, detection has some problems when the iris moves toward the corners or when the upper and lower boundaries of the iris are occluded by the eyelids and eyelashes, leading to gaze estimation errors.

On the other hand, for applications like clinical research, where experiments are performed in a doctor's office, it is not a problem to have infrared lighting, and thus active methods would be a better option. PCCR is the most common approach for feature-based gaze estimation methods.

When a light source (usually infrared) illuminates the eyes at different layers, the boundaries between the lens and the cornea act as convex mirrors and produces some reflections or virtual images, which are called corneal reflections or Purkinje images. In particular, the Purkinje image formed by the reflection of the outer surface of the cornea, called the first Purkinje image, is known as glint. The glint is the brightest and easiest reflection to detect and track. The PCCR technique uses the vector formed by the subtraction between the estimated center of the pupil and one or more near infrared (NIR) corneal reflections to estimate the gaze direction [49,62].

To compute the pupil-glint vector, the pupil center must also be extracted from the image. As it was already mentioned, different techniques are avaible for doing this but, with active illumination, the bright pupil-dark pupil method (BP-DP) is one of the most widely used for determining the accurate location of the pupil. When a light source is placed collinearly to the optical axis of the camera, most of the light is reflected back to the camera and the eye image shows a bright pupil. Conversely, when a light source is located away from the camera's optical axis, the image shows a dark pupil. Therefore, eye trackers with active IR illumination can use the difference between dark and bright pupil images by synchronously switching between the two light sources. This technique is very simple and robust in controlled conditions [11].

Besides that, the detection of the corneal reflections requires a narrow field of view (FOV) camera (long focal length) since the reflections are in general very small. Therefore, these systems work with high-resolution eye images captured by zooming in on movement-restricted users. Under these conditions, eye features can be easily and robustly extracted, this being an advantage over other methods.

These reported techniques are widely used and achieve really good results, but they have two major issues. First, because the mapping function is different for each person and for each system configuration, it is necessary to perform a tedious calibration procedure before each test to obtain the necessary parameters. In a typical calibration procedure, a set of visual targets such as those shown in Fig. 4, is presented to the user who normally has to stare to the computer for a period while the corresponding measurement is being done. Afterwords, from these correspondences, a mapping function is calculated.

The second drawback is that once the calibration has been performed, the person's head must remain motionless. Otherwise, there will be large errors between the actual and estimated directions. To avoid these errors head restraint systems are often used, and the calibration process is repeated every time movements are observed in the patient. Fig. 5 shows an active feature-based system, and the restraint system it uses to prevent head movement. With this device, Hernandez et al. [23] achieve an accuracy of less than 0.4°, reported as one of the minimum reaches in the literature.

A number of efforts are being made to minimize these shortcomings. 2D mapping methods assume that the mapping function have a particular parametric form such as a polynomial or a non-parametric form such as a neural networks whose coefficients have no physiological or physical meaning.

Polynomial interpolation is one of the main tools for parametric mapping functions, mainly due to its simplicity of execution and the good quality of the result obtained from it. In this case, the x and y gaze coordinates are estimated by means of a polynomial function. For example, a second order polynomial transformation is defined as:

$$\begin{aligned} x_c &= a_1 + a_2 x_e + a_3 y_e + a_4 x_e y_e + a_5 x_e^2 + a_6 y_e^2 \\ y_c &= b_1 + b_2 x_e + b_3 y_e + b_4 x_e y_e + b_5 x_e^2 + b_6 y_e^2, \end{aligned} \tag{10}$$

where (x_c, y_c) is the coordinate of the point on the screen where the gaze

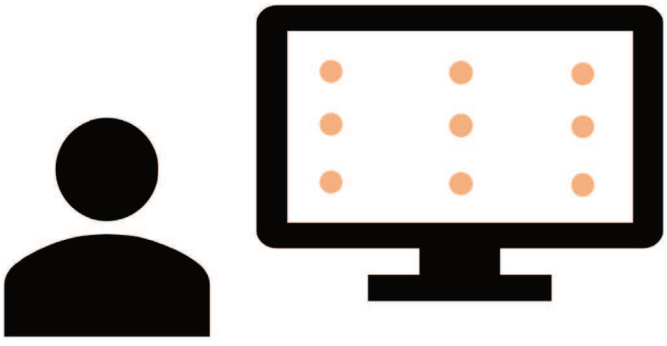


Fig. 4. Example of the calibration points.

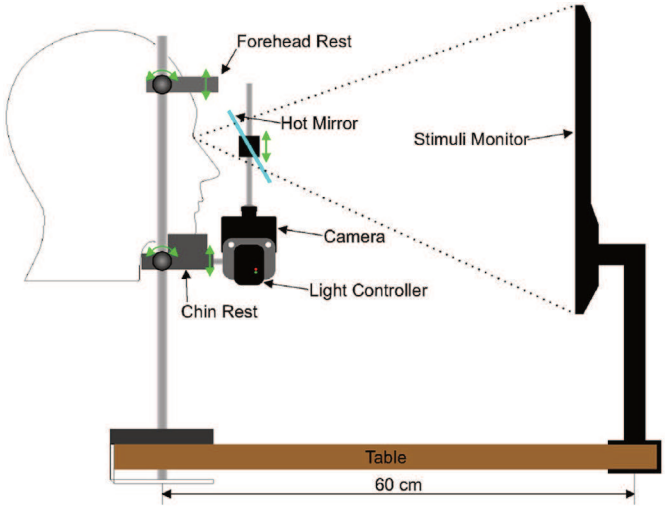


Fig. 5. Example of a gaze tracker with a restraint system (from Hernandez et al. [23]).

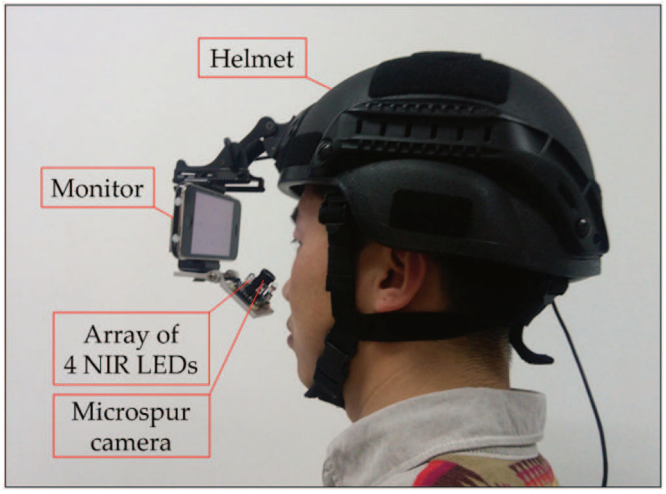


Fig. 6. Example of a head-mounted gaze tracker system (from Wang et al. [49]).

falls, (x_e, y_e) is the coordinate of the pupil-glint vector and a_i; b_i are the polynomial coefficients. These coefficients are calculated in the calibration procedure. During this procedure, the patient is asked to stare at a set of known targets, while a set of corresponding points are obtained. For example, for a 6-point calibration procedure, 6 corresponding points are obtained (x_{ci}, y_{ci}); (x_{ei}, y_{ei}) with $i = 1,2, ...6$ and a system of 12 equations is generated to calculate the polynomial coefficients a_i; b_i, by applying the least squares estimation procedure, that is, minimizing the quadratic error E^2 between the estimations and the calibration points coordinates. The higher is the order of the polynomial mapping function, the greater will be the number of calibration points needed to calculate all coefficients. In Equation (11) the quadratic error function for N calibration points is displayed.

$$E_x^2 = \sum_{i=1}^{N} [x_{ci} - (a_1 + a_2 x_{ei} + a_3 y_{ei} + a_4 x_{ei} y_{ei} + ..)]^2$$
$$E_y^2 = \sum_{i=1}^{N} [y_{ci} - (b_1 + b_2 x_{ei} + b_3 y_{ei} + b_4 x_{ei} y_{ei} + ..)]^2 \quad (11)$$

Mimica et al. [63] use a second order polynomial to minimize the number of calibration points required comparing to those required by a higher order polynomial. Cerrolaza et al. [64,65] carried out a study on the potential effect of the order and systematic inclusion of all polynomial terms, on the accuracy and robustness of the gaze tracker. For this, a real VOG system with different configurations was used. The authors point out that the gaze estimation accuracy of a gaze tracking system is not noticeably increased with the enhancement of polynomial order or with more complete mathematical expressions due to the factors of head motion, and calculation method of the pupil-glint vector.

The choice of the mapping function determines not only the accuracy of the system but also the head movement tolerance and the calibration time. Therefore, when linear regression solution methods are applied to solve the mapping function, a second-order linear polynomial is the most used due to its advantages of less calibration markers and better approximation effect.

Alternatively, Baluja et al. [32] first proposed a method using a simple artificial neural network (ANN) to calculate a non-linear mapping function. First, they mapped images of only the pupil and cornea as the inputs to ANN to the coordinates of the gaze point as the outputs. Then, they included the total eye socket as an input to improve the system accuracy (about 1.5°). In addition, Zhu and Ji [66] utilize generalized regression neural networks to estimate the gaze direction. For this purpose, 6 pupil and glint parameters were used as inputs to the calibration procedure. The parameters were chosen in such a way that they represent eye and head movements and remain relatively unchanged for different people. Therefore, even though the accuracy acquired is not good enough (about 5°), it is a free calibration process and head movements are allowed.

In a similar way, Gneo et al. [67] utilize multilayer neural feed-forward networks to calculate gaze point coordinates based on pupil-glint vectors. In order to minimize the number of output neurons, they use one separate network with the same input for each gaze coordinate (x, y). The reported results were competitive with high accuracy (about 0.6°). More recently, Wang et al. propose in Ref. [49] an improved ANN based on direct least squares regression to calculate the mapping function between pupil-glint vectors and actual gaze points. They combine the advantages of both methods: the high speed of direct least squares regression and the high accuracy of ANN. They achieved a good accuracy (about 0.4°) in a head-mounted device which can be seen in Fig. 6.

Thus, as it was pointed out before, the choice of the model depends on multiple factors: required accuracy, hardware cost, image quality/eye region resolution, available information in the image (e.g., glints), and configuration flexibility. For instance, feature-based methods accuracy may decrease when model assumptions are violated. In some applications such as clinical research or disease diagnosis, where it is possible to control the illumination conditions, the camera's quality, and the system settings, these methods achieve a really high accuracy which is critical to investigate imperfections in the oculomotor system [68].

Moreover, despite the fact that mapping methods provide little information about the intrinsic behavior of the system, they are much simpler to construct than the model-based methods and do not require additional hardware calibration, which makes setup much faster for the system user. That is why, most commercial gaze tracking systems use 2D mapping features-based methods with IR camera and active IR illumination, as it is shown in Fig. 7, to achieve the highly accurate performance of gaze estimation.

2.3. Head movements

So far we have talked about gaze tracking as a process that belongs exclusively to the eyes, but it is known that the gaze is a product of two contributing factors, the head pose (position and orientation) and the eyeball orientation. A person can change gaze direction by rotating the eyeball while keeping the head stationary; similarly, a person can change gaze direction by moving the head while keeping the eye stationary relative to the head.

Usually, a person moves the head to a comfortable position before

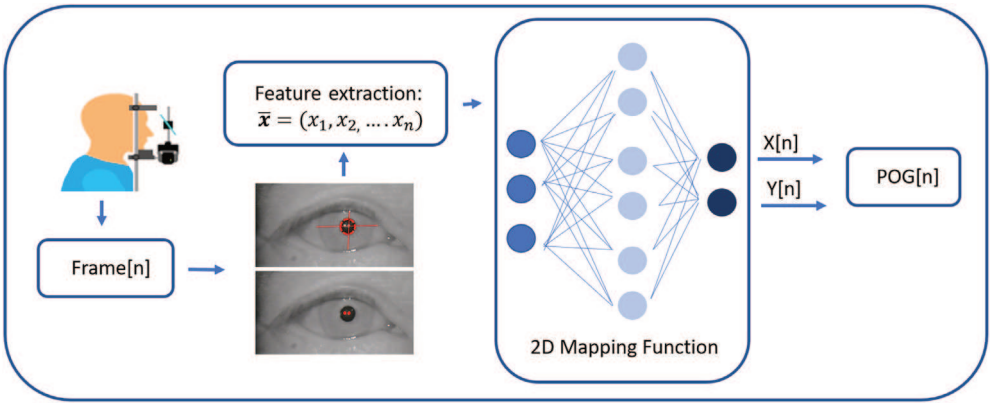


Fig. 7. Diagram of a standard eye tracker with 2D mapping method.

orienting the eye. Head pose, therefore, determines the coarse-scale gaze direction while the eyeball orientation determines the local and detailed gaze direction. For example, when the target is located over 20° of the field of view, it is much more comfortable to rotate the head than to rotate the eyeball.

While in the previous section we reviewed the state-of-the-art in gaze estimation systems, focusing on tracking eye movement, the problem of ensuring the invariance to head movements is also an important and a challenging research topic. In almost all applications, people move their heads while they are using a gaze tracker. For this reason, for an accurate gaze estimation, it is necessary to (either directly or implicitly) model both head pose and eye rotation.

As we pointed out above, appearance-based methods have been developed in order to be applied in environmental conditions without the possibility of monitoring user conditions. This user freedom movement situation requires the method to be robust to changes in head position. For solving this problem two possibilities can be held: learning generic gaze estimators from large amounts of head pose-independent training data or adding head pose information to the eye images.

In particular, Lu et al. in Ref. [69] address the head motion problem by synthesizing new training images for different head poses from those already seen in estimation, while in Ref. [70] they perform the gaze estimation by assuming a fixed head pose and then compensating for the estimation biases caused by the head pose using a head pose tracker. Conversely, Zhang et al. in Ref. [27] used a multimodal convolutional neural network to learn a mapping function from both the head poses and eye images to the unique gaze directions. Lai et al. in Ref. [71] also combine the eye image information with head pose tracking selecting features by means of the neighborhood-based regression algorithm. Although these results outperform the appearance-based methods state of the art, the accuracy achieved is still quite poor what that required for clinical applications.

Regardless of the eye tracker method applied, other researchers have attempted to measure the head movements or head position directly, and to use that information to correct the gaze measurements. These systems usually include two or more cameras and use a complex facial model to track the movement of the face [72,73]. In order to track the face from these models, the FOV of the tracking camera has to be large enough to cover the entire user's head. This is not a problem for daily applications which do not need such a high accuracy like in Ref. [74], but, when feature-based methods are applied, these restrictions make it difficult to locate the small eye features and result in less accurate gaze tracking.

To overcome this drawback, some works have proposed the use of two cameras in combination with pan and tilt mechanisms that allow freedom of person motion while maintaining the feature method accuracy. In Ref. [75], Hennessey et al. proposed a system that rotates an eye tracker with a narrow-angle camera using pan and tilt servo motors, while in Ref. [76] Cho et al. propose a binocular eye gaze tracking system that, using pan, tilt and zoom movements, continue to track the eyes with a narrow camera while the user moves his head freely in depth. Then both estimate the POG by a 2D mapping function modified with the depth of the eyes. All these methods work relatively well but are very complex, expensive and, most notably, slow. These limitations restrict its use so that in practice, head pose information is rarely used directly in the gaze models. It is more common to incorporate this information implicitly either through the mapping function (regression-based method) or through the use of reflections on the cornea (3D model-based approaches).

Apart from that, the 3D model-based methods are the most robust to head pose changes and they can obtain the head pose invariance through various hardware configurations and prior knowledge of the geometry and cameras. Guestrin et al. [44] presented a general study for PCCR covering all the possible system configurations in terms of number and positioning of IR light sources and cameras. In that work the authors claimed that using only one camera with two light sources is the simplest configuration that allows for both the estimation of POG and free head user movements. To estimate the POG with this system configuration, it is necessary to make a subject-specific calibration procedure that requires the subject to fixate on multiple points. To avoid the need of calibration, an additional camera is also necessary.

Using one camera and one light source the POG can be estimated only if the head is completely stationary. This restriction is shared with the 2D regression methods, which assume static head conditions. In general, gaze estimation systems that use one camera and one light source assume that the head movements are negligible. Therefore, it should be noted that the only video oculography method that allows large head movements while maintaining good accuracy are some 3D model-based methods. But as a drawback, they require a really complex and calibrated system setting, difficulting their use in common applications. In addition, the movement of the head is taken into account implicitly, and it is not possible to differentiate between oculocephalic and purely ocular movements.

This is another reason why the eye location algorithms found in commercially available eye trackers use the 2D regression techniques where the 3D eye location is usually unknown and only the relative orientation of the user's eye with respect to the user's head is measured. Particularly, gaze estimation is based on the relative position between pupil and glint. Assuming a static head, methods based on this idea use the glint as a reference point, thus the vector from the glint to the center of the pupil will describe the gaze direction. While contact-free and non-intrusive, these methods work well only for a static head, but even minor head can fail these techniques.

In addition, since the pupil and glints are very small, the FOV of the eye tracking camera has to be confined to obtain a high definition eye

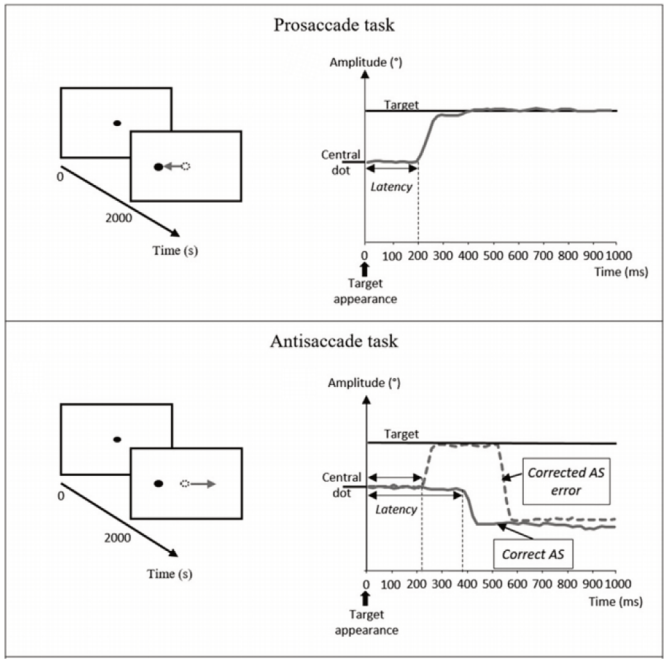


Fig. 8. Dynamics of the saccade as a function of time in the Saccade tests. (from Noiret et al. [83]).

image. This aspect also limits the head so that the eye does not disappear from the FOV and emphasizes the problem of sensitivity to head pose variations, requiring the user to be either equipped with a head-mounted device or to use a high-resolution camera combined with a chin rest to limit the allowed head movements.

In many clinic applications where tests are conducted for only a few minutes and it is not easy to perform complex system calibrations, those restraint systems are suitable and work very well. Even so, despite the fact that the head movements are restricted, in people with certain neurological diseases, it is possible to observe some involuntary movements and it is very important be able to measure them not only to correct the gaze estimation errors but also because these measurements are indicators of the presence or progress of certain diseases.

As it was pointed out before, there are a lot of research works that deal with the problem of obtaining enhanced gaze estimation in presence of large movements and head pose variations for daily applications [74]. But, despite their importance for clinic diagnosis, there are not many studies performing the feasibility of a gaze estimator that considers both the head and eye movements in a really zoomed and high-resolution images with the objective to detect short involuntary head movements.

Some works study the way of achieving head pose invariance for 2D regression methods, for example [66] use Generalized Regression Neural Networks (GRNN) instead of polynomial functions to account head implicitly by the gaze mapping function. They also include more parameters as mapping function inputs like glint coordinates and pupil radio to account for the different head motions. In Ref. [77] authors also employed GRNN but with the aim of compensating the errors generated by the non-linear polynomial for different head poses obtaining a gaze estimation robust to head but with much lower spatial gaze resolution.

Despite these advances, there are no studies done to quantify these slight movements based only on the zoomed eye image which is necessary for some neurological disease diagnosis.

3. Clinical applications

One of the best-known clinical applications of high-precision eye tracking is refractive surgery assistance [78,79]. With this type of surgery, eye refractive disorders (myopia, hypermetropia and astigmatism) are corrected by modelling the cornea by laser ablation [80]. As expected, this procedure requires great precision, and although the patient is asked to look at a fixed point during the few minutes that the surgery lasts, the patient's eye is not fixed and small movements may occur. Therefore, modern laser devices are equipped with sophisticated eye-tracking systems, which by detecting and tracking the pupil and limbus, follow the eye movements and rectify in real time the position of the mirrors that direct the laser. To achieve this adjustment, cameras with acquisition frequencies of the order of 1000 Hz and latencies of less than 3 ms are used [81]. Most of these devices perform the compensation for linear eye movements (in the x/y-axis), rolling eye movements and eye torsions around the visual axis known as cyclotorsions [82].

As discussed in the introduction, eye tracking is used not only as an assistant for clinical applications but also for the diagnosis of neurological diseases, applications on which this review focuses. In the last few years a series of tests have been developed to analyze different alterations in eye movements. In particular, it has been found that the alterations of saccadic eye movements (SEM) provide relevant information.

A widely used test to measure these movements consists of asking the patient to look at a point in the center of the screen until a new point appears on the periphery. At this moment, the patient must look towards it (Pro-Saccade task) or the opposite side (Anti-saccade task) as rapidly and as accurately as possible. In addition, the second target can be turned on before the first target is turned off (Overlap) or a few milliseconds after the first target is turned off (Gap). By combining these paradigms, other metrics can be obtained for analysis. This type of test are very simple to perform for the patient and allow to measure different parameters to be used as markers such as those that can be observed in Fig. 8.

Numerous studies based on this type of test have been carried out in patients with Alzheimer's disease (AD) [84], where it has been possible to distinguish some of the ocular movement alterations, finding significant contributions for the diagnosis of the disease, even in its early stages. In Refs. [83,85] the authors presented a study of SEM in patients with Alzheimer and healthy control patients finding significant differences. For instance, it was found that AD patients had higher latency and latency variability regardless of the tasks. Moreover, AD patients made more uncorrected Anti-Saccade (AS) and took more time-to-correct incorrect AS. In addition, close relationships were found between the majority of SEM variables and dementia screening tests, especially the MiniMental State Examination (MMEE) and episodic memory measure. Also in Ref. [86] the AS test was performed, finding significant differences in both the number of AS errors and the number of errors that remain uncorrected.

Parkinson's disease and other parkinsonian syndromes [87] are another of the conditions in which this type of measurements provides very important information [88,89]. One of the most prominent features of eye movement abnormality in PD are saccade hypometria, abnormally fragmented saccades, called multistep or staircase saccades, an increment of the latency of the saccades and a marked difficulty in inhibiting the saccade movement reflex during the test of the AS [87].

Measurements of eye movement abnormalities have also drawn attention as a biomarker in the diagnosis of multiple sclerosis (MS) [90]. For this task, a particular paradigm known as the endogenously generated sacadde paradigm is used. The same is similar to the pro-sacadde test but an intermediate step is added to it. After the initial target disappears from the screen, an arrow is shown on it whose direction may be equal or opposite to that of the final target. With such studies, saccade latencies were found to increase in patients with MS even as a function of the disease duration [91]. In addition, these patients frequently exhibit fatigue symptoms at any stage of the disease, having a major impact on their quality of life. Recent reports have

Table 1
Comparison of the different methods.

Method	Ref.	Accuracy (degrees)	Head	Calibration	System Setup
Appearance	[39]	6.5°	Free head pose	Free	1 Camera
	[40]	7.8°	Not considered	Free	1 Camera
	[27]	6.9°	Free head pose	Free	1 Camera
	[71]	2° 5°	Allowed (tracked)	Needed	1 Camera
	[69]	2.5°	Free head pose	Needed	1 Camera
	[44]	0.9°	Moderate (2–3 cm)	Multiple points	1 Camera - 2 light sources
Model	[45]	< 1°	Moderate (10 dm^3)	One point	2 Cameras - 2 light sources
	[46]	< 1°	Yawing and pitching allowed	One point	2 Cameras - multiple light
	[47]	> 1°	Natural head movements	4 points	2 Cameras - 2 light sources
	[47]	> 1°	Natural head movements	4 points	2 Cameras - 2 light sources
Shape	[52]	2.42°	Fix	9 points	1 Camera
	[57]	1°	Allowed (tracked)	Needed	2 Cameras
	[60]	4°	Fix	4 points	1 Camera
	[74]	2–5°	Allowed (tracked)	Needed	1 Camera
2D	[63]	0.8°	Fix	9 points	1 Camera - 2 light sources
Mapping	[65]	0.38°	Fix	16 points	1 Camera - 2 light sources
	[23]	0.4°	Fix	9 points	1 Camera - 2 light sources
	[32]	1.5°	Allowed	Moving mouse	1 Camera
	[66]	5°	Natural head movements	Free	1 Camera - 2 IR arrays
	[67]	0.6°	Fix	4 × 5 grid	1 Camera - 3 IR arrays
	[49]	0.4°	Head mounted	16 points	1 Camera - 4 light sources
	[75]	< 2°	Free head pose (Tracked)	9 points	2 Cameras - 1 IR plate 1 pan-tilt mechanism
	[76]	0.69	Back and forth movements	6 points 3 depth	4 Cameras - 16 NIR leds pan-tilt mechanism
	[77]	–	Natural head movements	9 points	1 Camera - 1 light source

shown an increase in endogenous saccade latencies and a reduction in the peak velocities associated with this symptom [92].

4. Conclusions

In this article, several gaze tracking algorithms and their respective advantages and disadvantages for use in disease diagnosis were analyzed. Table 1 summarizes the state of the art of VOG methods by grouping them according to their classification. The method accuracy is calculated as the mean angular error of the gaze estimation. In addition, the system setup (number of cameras and infrared lamps), the number of calibration points used, and the allowed head movements are specified. It should be noted that it would be a significant contribution to be able to provide the resolution of the hardware used, unfortunately this information is not available in many of the original works. In addition, due to the variety of methods presented, providing only the camera resolution would not be enough to account the hardware resolution. Since in some cases the images are taken from the whole face while in others are taken with zoom on the eye, these differences make it very difficult to uniquely characterize acquisition systems.

Appearance-based and passive-shape based methods are not suitable for clinical applications because of their low accuracy. These methods have been researched for daily applications which are carried out in natural illumination conditions and/or with low-resolution cameras but not for clinical applications, where the experimental environment can be controlled with the possibility of even using infrared illumination.

3D model-based methods have an excellent head tolerance. However, the hardware requirements for their implementation is really complex as they need several light sources, multiple cameras and a perfect system calibration. Besides, it is not possible to differentiate between oculocephalic and purely ocular movements.

Thus, we conclude that active 2D regression-based methods are the best option for clinical diagnosis or research applications, since they use features coming from the human eye, such as pupil center and corneal reflections, and they can be implemented using a single camera and a few NIR LEDs. However, these techniques are very vulnerable to head movements and require users to hold their head very still using a headrest, chin rest or bite bar. This restraint system are not enough in presence of some neurological disease, where involuntary head movements occur.

In light of the aforementioned advantages and disadvantages of each methodology considered, it is clear that it is still an open issue to find an optimal way of measuring the head movements, to be able to use both to correct the gaze point estimation and as another biomarker in the disease analysis. We consider that the path to follow to reduce these shortcomings is to focalize on the application for which the device is being designed. In the case of eye trackers for neurological disease diagnosis, it would be helpful to exploit the hardware functionalities where the patient's movements are restricted to a minimum of space during the test duration. In this way, research could be focused on methods that also incorporate measurements of small displacements or rotations from the initial position.

Conflicts of interest

None declared

Acknowledgements

The authors wish to thank the support of (UNL (CAID-PIC-50420150100098LI), ANPCyT (PICT 2016-0651) and RoboCity2030-DIH-CM Madrid Robotics Digital Innovation Hub (S2018/NMT-4331), funded by “Programas de Actividades I+D en la Comunidad de Madrid” and cofunded by Structural Funds of the EU

References

[1] Martin Gorges, H Pinkhardt Elmar, Kassubek Jan, Alterations of eye movement control in neurodegenerative movement disorders, Journal of Ophthalmology 2014 (2014) 11, https://doi.org/10.1155/2013/453402 Article ID 658243.
[2] J Anderson Tim, Michael R. MacAskill, Eye movements in patients with neurodegenerative disorders, Nat. Rev. Neurol. 9 (2) (2013) 74.
[3] H Pinkhardt Elmar, Kassubek Jan, Ocular motor abnormalities in parkinsonian syndromes, Park. Relat. Disord. 17 (4) (2011) 223–230.
[4] Pretegiani Elena, M Optican Lance, Eye movements in Parkinson's disease and inherited parkinsonian syndromes, Front. Neurol. 8 (2017) 592.

[5] Colette Donaghy, Matthew J. Thurtell, Erik P. Pioro, J Mark Gibson, R John Leigh, Eye movements in amyotrophic lateral sclerosis and its mimics: a review with illustrative cases, J. Neurol. Neurosurg. Psychiatry 82 (1) (2011) 110–116.

[6] Stephen L. Hicks, Matthieu PA. Robert, Charlotte VP. Golding, J Tabrizi Sarah, Christopher Kennard, Oculomotor deficits indicate the progression of Huntington's disease, Progress in Brain Research, vol. 171, Elsevier, 2008, pp. 555–558.

[7] Gerardo Fernández, Pablo Mandolesi, Nora P. Rotstein, Oscar Colombo, Osvaldo Agamennoni, Luis E. Politi, Eye movement alterations during reading in patients with early alzheimer disease, Investig. Ophthalmol. Vis. Sci. 54 (13) (2013) 8345–8352.

[8] Sara Montagnese, M Gordon Harriet, Clive Jackson, Justine Smith, Patrizia Tognella, Nutan Jethwa, R Michael Sherratt, Marsha Y. Morgan, Disruption of smooth pursuit eye movements in cirrhosis: relationship to hepatic encephalopathy and its treatment, Hepatology 42 (4) (2005) 772–781.

[9] Jorge Otero-Millan, Jose L. Alba Castro, Stephen L. Macknik, Susana Martinez-Conde, Unsupervised clustering method to detect microsaccades, J. Vis. 14 (2) (2014) 18–18.

[10] Leandro L. Di Stasi, Michael B. McCamy, Andrés Catena, Stephen L. Macknik, J Canas José, Susana Martinez-Conde, Microsaccade and drift dynamics reflect mental fatigue, Eur. J. Neurosci. 38 (3) (2013) 2389–2398.

[11] Dan Witzner Hansen, Qiang Ji, In the eye of the beholder: a survey of models for eyes and gaze, IEEE Trans. Pattern Anal. Mach. Intell. 32 (3) (2010) 478–500.

[12] M Dalton Kim, M Nacewicz Brendon, Tom Johnstone, Hillary S. Schaefer, Ann Gernsbacher Morton, Hill H. Goldsmith, Andrew L. Alexander, Richard J. Davidson, Gaze fixation and the neural circuitry of face processing in autism, Nat. Neurosci. 8 (4) (2005) 519.

[13] Jacob L. Orquin, Simone Mueller Loose, Attention and choice: a review on eye movements in decision making, Acta Psychol. 144 (1) (2013) 190–206.

[14] Martin Stetter, Raimund A. Sendtner, T Timberlake George, A novel method for measuring saccade profiles using the scanning laser ophthalmoscope, Vis. Res. 36 (13) (1996) 1987–1994.

[15] Andreas Bulling, Jamie A. Ward, Hans Gellersen, Gerhard Troster, Eye movement analysis for activity recognition using electrooculography, IEEE Trans. Pattern Anal. Mach. Intell. 33 (4) (2011) 741–753.

[16] Ronald S. Remmel, An inexpensive eye movement monitor using the scleral search coil technique, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. (4) (1984) 388–390.

[17] Moritz Kassner, William Patera, Andreas Bulling, Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction, Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2014, pp. 1151–1160. Adjunct publication.

[18] Enrique Cáceres, Miguel Carrasco, Sebastián Ríos, Evaluation of an eye-pointer interaction device for human-computer interaction, Heliyon 4 (3) (2018) e00574.

[19] Amna Rahman, Mehreen Sirshar, Aliya Khan, Real time drowsiness detection using eye blink monitoring, Software Engineering Conference (NSEC), 2015 National, IEEE, 2015, pp. 1–7.

[20] Emilien Dubois, Colin Blättler, Cyril Camachon, Christophe Hurter, Eye movements data processing for ab initio military pilot training, Intelligent Decision Technologies, Springer, 2015, pp. 125–135.

[21] RG Vishnu Menon, Valdimar Sigurdsson, Nils Magne Larsen, Asle Fagerstrøm, Gordon R. Foxall, Consumer attention to price in social commerce: eye tracking patterns in retail clothing, J. Bus. Res. 69 (11) (2016) 5008–5013.

[22] Luz Rello, Miguel Ballesteros, Detecting readers with dyslexia using machine learning with eye tracking measures, Proceedings of the 12th Web for All Conference, ACM, 2015, p. 16.

[23] Erik Hernández, Hernández Santiago, David Molina, Rafael Acebrón, Cecilia E García Cena, Oscann: technical characterization of a novel gaze tracking analyzer, Sensors 18 (2) (2018) 522.

[24] Juan Biondi, Gerardo Fernandez, Silvia Castro, Osvaldo Agamennoni, Eye-movement Behavior Identification for Ad Diagnosis, (2017) arXiv preprint arXiv:1702.00837.

[25] Anuradha Kar, Peter Corcoran, A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms, IEEE Access 5 (2017) 16495–16519.

[26] Stefania Cristina, Kenneth P. Camilleri, Unobtrusive and Pervasive Video-Based Eye-Gaze Tracking, Image and Vision Computing, 2018.

[27] Xucong Zhang, Yusuke Sugano, Mario Fritz, Andreas Bulling, Appearance-based gaze estimation in the wild, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4511–4520.

[28] Kar-Han Tan, David J. Kriegman, Narendra Ahuja, Appearance-based eye gaze estimation, Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on, IEEE, 2002, pp. 191–195.

[29] Yasuhiro Ono, Takahiro Okabe, Yoichi Sato, Gaze estimation from low resolution images, Pacific-rim Symposium on Image and Video Technology, Springer, 2006, pp. 178–188.

[30] Onur Ferhat, Fernando Vilariño, Low cost eye tracking, Comput. Intell. Neurosci. 17 (2016) 2016.

[31] Zhizhi Guo, Qianxiang Zhou, Zhongqi Liu, Appearance-based gaze estimation under slight head motion, Multimed. Tool. Appl. 76 (2) (2017) 2203–2222.

[32] Shumeet Baluja, Pomerleau Dean, Non-intrusive gaze tracking using artificial neural networks, Advances in Neural Information Processing Systems, 1994, pp. 753–760.

[33] Li-Qun Xu, Dave Machin, Phil Sheppard, A novel approach to real-time non-intrusive gaze finding, BMVC (1998) 1–10.

[34] Weston Sewell, Oleg Komogortsev, Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network, CHI'10 Extended Abstracts on Human Factors in Computing Systems, ACM, 2010, pp. 3739–3744.

[35] Feng Lu, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Inferring human gaze from appearance via adaptive linear regression, Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 153–160.

[36] Kar-Han Tan, David J. Kriegman, Narendra Ahuja, Appearance-based eye gaze estimation, Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on, IEEE, 2002, pp. 191–195.

[37] Yusuke Sugano, Yasuyuki Matsushita, Yoichi Sato, Appearance-based gaze estimation using visual saliency, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 329–341.

[38] Ke Liang, Youssef Chahir, Michèle Molina, Charles Tijus, François Jouen, Appearance-based gaze tracking with spectral clustering and semi-supervised Gaussian process regression, Proceedings of the 2013 Conference on Eye Tracking South Africa, ACM, 2013, pp. 17–23.

[39] Yusuke Sugano, Yasuyuki Matsushita, Yoichi Sato, Learning-by-synthesis for appearance-based 3d gaze estimation, Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE, 2014, pp. 1821–1828.

[40] Feng Lu, Xiaowu Chen, Person-independent eye gaze prediction from eye images using patch-based features, Neurocomputing 182 (2016) 10–17.

[41] Timo Schneider, Boris Schauerte, Rainer Stiefelhagen, Manifold alignment for person independent appearance-based gaze estimation, Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE, 2014, pp. 1167–1172.

[42] Jixu Chen, Qiang Ji, A probabilistic approach to online eye gaze tracking without explicit personal calibration, IEEE Trans. Image Process. 24 (3) (2015) 1076–1086.

[43] Ryoung Park Kang, A real-time gaze position estimation method based on a 3-d eye model, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 37 (1) (2007) 199–212.

[44] Elias Daniel Guestrin, Moshe Eizenman, General theory of remote gaze estimation using the pupil center and corneal reflections, IEEE Trans. Biomed. Eng. 53 (6) (2006) 1124–1133.

[45] Elias Daniel Guestrin, Moshe Eizenman, Remote point-of-gaze estimation requiring a single-point calibration for applications with infants, Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, ACM, 2008, pp. 267–274.

[46] Sheng-Wen Shih, Jin Liu, A novel approach to 3-D gaze tracking using stereo cameras, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34 (1) (2004) 234–245.

[47] Jixu Chen, Yan Tong, Wayne Gray, Qiang Ji, A robust 3D eye gaze tracking system using noise reduction, Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, ACM, 2008, pp. 189–196.

[48] Carlos H. Morimoto, Marcio RM. Mimica, Eye gaze tracking techniques for interactive applications, Comput. Vis. Image Understand. 98 (1) (2005) 4–24.

[49] Jianzhong Wang, Guangyue Zhang, Jiadong Shi, 2d gaze estimation based on pupil-glint vector using an artificial neural network, Appl. Sci. 6 (6) (2016) 174.

[50] Krafka Kyle, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, Antonio Torralba, Eye tracking for everyone, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2176–2184.

[51] Onur Ferhat, Fernando Vilariño, Low cost eye tracking, Comput. Intell. Neurosci. 17 (2016) 2016.

[52] Seung-Jin Baek, Kang-A. Choi, Chunfei Ma, Young-Hyun Kim, Sung-Jea Ko, Eyeball model-based iris center localization for visible image-based eye-gaze tracking systems, IEEE Trans. Consum. Electron. 59 (2) (2013) 415–421.

[53] Ronda Venkateswarlu, et al., Eye gaze estimation from a single image of one eye, Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE, 2003, pp. 136–143.

[54] David Young, Hilary Tunley, Richard Samuels, Specialised Hough Transform and Active Contour Methods for Real-Time Eye Tracking, School of Cognitive and Computing Science, University of Sussex, Cognitive & Computing Science, 1995 CSRP no. 386.

[55] Peng Yang, Bo Du, Shiguang Shan, Wen Gao, A novel pupil localization method based on GaborEye model and radial symmetry operator, Image Processing, 2004. ICIP'04. 2004 International Conference on, vol. 1, IEEE, 2004, pp. 67–70.

[56] Anjith George, Aurobinda Routray, Fast and accurate algorithm for eye localization for gaze tracking in low-resolution images, ET Computer Vision 10 (7) (2016) 660–669.

[57] Jian-Gang Wang, Eric Sung, Ronda Venkateswarlu, Estimating the eye gaze from one eye, Comput. Vis. Image Understand. 98 (1) (2005) 83–103.

[58] Kyung-Nam Kim, R.S. Ramakrishna, Vision-based eye-gaze tracking for human computer interface, Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on, vol. 2, IEEE, 1999, pp. 324–329.

[59] Roberto Valenti, Theo Gevers, Accurate eye center location and tracking using isophote curvature, Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[60] Dan Witzner Hansen, EC Pece Arthur, Eye tracking in the wild, Comput. Vis. Image Understand. 98 (1) (2005) 155–181.

[61] Dongheng Li, David Winfield, Derrick J. Parkhurst, Starburst: a hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches, Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on, IEEE, 200579–79.

[62] Thomas E Hutchinson, K Preston White, Worthy N. Martin, Kelly C. Reichert, Lisa A. Frey, Human-computer interaction using eye-gaze input, IEEE Transactions on systems, man, and cybernetics 19 (6) (1989) 1527–1534.

[63] Marcio RM. Mimica, Carlos Hitoshi Morimoto, A computer vision framework for eye gaze tracking, Computer Graphics and Image Processing, 2003. SIBGRAPI 2003. XVI Brazilian Symposium on, IEEE, 2003, pp. 406–412.

[64] Juan J. Cerrolaza, Arantxa Villanueva, Rafael Cabeza, Taxonomic study of polynomial regressions applied to the calibration of video-oculographic systems, Proceedings of the 2008 Symposium on Eye Tracking Research & Applications, ACM, 2008, pp. 259–266.
[65] Juan J. Cerrolaza, Arantxa Villanueva, Rafael Cabeza, Study of polynomial mapping functions in video-oculography eye trackers, ACM Trans. Comput. Hum. Interact. 19 (2) (2012) 10.
[66] Zhiwei Zhu, Qiang Ji, Eye and gaze tracking for interactive graphic display, Mach. Vis. Appl. 15 (3) (2004) 139–148.
[67] Massimo Gneo, Maurizio Schmid, Silvia Conforto, Tommaso DAlessio, A free geometry model-independent neural eye-gaze tracking system, J. NeuroEng. Rehabil. 9 (1) (2012) 82.
[68] Marcus Nyström, Richard Andersson, Kenneth Holmqvist, Joost Van De Weijer, The influence of calibration method and eye physiology on eyetracking data quality, Behav. Res. Methods 45 (1) (2013) 272–288.
[69] Feng Lu, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Gaze estimation from eye appearance: a head pose-free method via eye image synthesis, IEEE Trans. Image Process. 24 (11) (2015) 3680–3693.
[70] Feng Lu, Takahiro Okabe, Yusuke Sugano, Yoichi Sato, Learning gaze biases with head motion for head pose-free gaze estimation, Image Vis Comput. 32 (3) (2014) 169–179.
[71] Chih-Chuan Lai, Yu-Ting Chen, Kuan-Wen Chen, Shen-Chi Chen, Sheng-Wen Shih, Yi-Ping Hung, Appearance-based gaze tracking with free head movement, Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE, 2014, pp. 1869–1873.
[72] Kwang Ho An, Myung Jin Chung, 3d head tracking and pose-robust 2d texture map-based face recognition using a simple ellipsoid model, Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on, IEEE, 2008, pp. 307–312.
[73] Jilin Tu, Thomas Huang, Hai Tao, Accurate head pose tracking in low resolution video, Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, IEEE, 2006, pp. 573–578.
[74] Roberto Valenti, Nicu Sebe, Theo Gevers, Combining head pose and eye location information for gaze estimation, IEEE Trans. Image Process. 21 (2) (2012) 802–815.
[75] Hennessey Craig, Fiset Jacob, Long range eye tracking: bringing eye tracking into the living room, Proceedings of the Symposium on Eye Tracking Research and Applications, ACM, 2012, pp. 249–252.
[76] Dong-Chan Cho, Wah-Seng Yap, HeeKyung Lee, Injae Lee, Whoi-Yul Kim, Long range eye gaze tracking system for a large screen, IEEE Trans. Consum. Electron. 58 (4) (2012).
[77] Chi Jian-nan, Chuang Zhang, Yan Yan-tao, Yang Liu, Han Zhang, Eye gaze calculation based on nonlinear polynomial and generalized regression neural network, Natural Computation, 2009. ICNC'09. Fifth International Conference on, vol. 3, IEEE, 2009, pp. 617–623.
[78] Michael Mrochen, Mostafa Salah Eldine, Maik Kaemmerer, Theo Seiler, Hütz Werner, Improvement in photorefractive corneal laser surgery results using an active eye-tracking system, J. Cataract Refract. Surg. 27 (7) (2001) 1000–1006.
[79] Chieh Lee Yuan, Active eye-tracking improves lasik results, J. Refract. Surg. 23 (6) (2007) 581–585.
[80] Kerry D. Solomon, Luis E. Fernández de Castro, Helga P. Sandoval, Joseph M. Biber, Brian Groat, Kristiana D. Neff, S Ying Michelle, John W. French, D Donnenfeld Eric, Richard L. Lindstrom, et al., Lasik world literature review: quality of life and patient satisfaction, Ophthalmology 116 (4) (2009) 691–701.
[81] L Alió Mohamed El Bahrawy and Jorge, Excimer laser 6 th generation: state of the art and refractive surgical outcomes, Eye and Vision 2 (1) (2015) 6.
[82] M Aslanides Ioannis, Georgia Toliou, Sara Padroni, Samuel Arba Mosquera, Sai Kolli, The effect of static cyclotorsion compensation on refractive and visual outcomes using the schwind amaris laser platform for the correction of high astigmatism, Contact Lens Anterior Eye 34 (3) (2011) 114–120.
[83] Nicolas Noiret, Nicolas Carvalho, Éric Laurent, Gilles Chopard, Mickaël Binetruy, Magali Nicolier, Julie Monnin, Eloi Magnin, Pierre Vandel, Saccadic eye movements and attentional control in alzheimer's disease, Arch. Clin. Neuropsychol. 33 (1) (2017) 1–13.
[84] Robert J. Molitor, Philip C. Ko, A Ally Brandon, Eye movements in alzheimer's disease, J. Alzheimer's Dis. 44 (1) (2015) 1–12.
[85] Qing Yang, Tao Wang, Ning Su, Shifu Xiao, Zoi Kapoula, Specific saccade deficits in patients with alzheimers disease at mild to moderate stage and in patients with amnestic mild cognitive impairment, Age 35 (4) (2013) 1287–1298.
[86] Liam D. Kaufman, Jay Pratt, Brian Levine, Sandra E. Black, Executive deficits detected in mild alzheimer's disease using the antisaccade task, Brain and behavior 2 (1) (2012) 15–21.
[87] Pretegiani Elena, M Optican Lance, Eye movements in Parkinsons disease and inherited parkinsonian syndromes, Front. Neurol. 8 (2017) 592.
[88] Anshul Srivastava, Ratna Sharma, Sanjay K. Sood, Garima Shukla, Vinay Goyal, Madhuri Behari, Saccadic eye movements in Parkinson's disease, Indian J. Ophthalmol. 62 (5) (2014) 538.
[89] Martin Gorges, Hans-Peter Müller, Dorothée Lulé, H Pinkhardt Elmar, Albert C. Ludolph, Kassubek JanLANDSCAPE Consortium, The association between alterations of eye movement control and cerebral intrinsic functional connectivity in Parkinsons disease, Brain imaging and behavior 10 (1) (2016) 79–91.
[90] K Sheehy Christy, Alexandra Beaudry-Richard, Ethan Bensinger, Jacqueline Theis, J Green Ari, Methods to assess ocular motor dysfunction in multiple sclerosis, J. Neuro Ophthalmol. 38 (4) (2018) 488–493.
[91] Meaghan Clough, Lynette Millist, Nathaniel Lizak, Beh Shin, Teresa C. Frohman, Elliot M. Frohman, Owen B. White, Joanne Fielding, Ocular motor measures of cognitive dysfunction in multiple sclerosis i: inhibitory control, J. Neurol. 262 (5) (2015) 1130–1137.
[92] Marisa Ferreira, Paulo A. Pereira, Marta Parreira, Inês Sousa, José Figueiredo, João J Cerqueira, Antonio F. Macedo, Using endogenous saccades to characterize fatigue in multiple sclerosis, Multiple sclerosis and related disorders 14 (2017) 16–22.

Anexo F

Eye corners tracking for head movement estimation

Eye corners tracking for head movement estimation

Agostina J. Larrazabal
Research institute for signals, systems and computational intelligence , sinc(i)
FICH-UNL/CONICET
3000 Santa Fe, Argentina
alarrazabal@sinc.unl.edu.ar

Cecilia E. García Cena
Centre for Robotics and Automation
UPM-CSIC
28006 Madrid, Spain
cecilia.garcia@upm.es

César E. Martínez
Research institute for signals, systems and computational intelligence , sinc(i)
FICH-UNL/CONICET
3000 Santa Fe, Argentina
cmartinez@sinc.unl.edu.ar

***Abstract*—Recently, video-oculographic gaze tracking has begun to be used in the diagnosis of a wide variety of neurological diseases, such as Parkinson and Alzheimer. For this application, the so-called feature-based methods are used, more precisely, 2D regression-based methods. They use geometrically derived eye features from high-resolution eye images captured by zooming into the user's eyes. The main weakness of these methods is that the head of the user must remain motionless to avoid estimation errors. In some patients, some involuntary movements cannot be avoided and it is necessary to measure them. In this paper, we tackle the measurement of head position as a way to improve the gaze tracking on these precision demanding medical applications. As a first stage, we propose to obtain the eye corners coordinates as a reference point, since they are the most stable points in front of the eyeball and eyelids movements. The problem was handled as a regression problem using a coarse-to-fine cascaded convolutional neural network in order to accurately regress the coordinates of the eye corner. Particularly, with the aim of achieving high precision we cascade two levels of convolutional networks. Finally, we added temporal information to increase accuracy and decrease computation time. The accuracy of the estimation was calculated from the mean square error between the predictions and the ground truth. Subjective performance was also evaluated through video inspection. In both cases, satisfactory results were obtained.**

***Index Terms*—Landmark Tracking, Convolutional Neural Networks, Head Movements**

I. Introduction

Studies have shown that activity related to eye movements is observed in cortical and subcortical areas, which are directly and indirectly connected with several neural systems. They interact with each other to control the suitable performance of the ocular and ocular-cephalic movements. As a result, a broad spectrum of oculomotor alterations are usually observed in the presence of neurodegenerative motor disorders. An accurate and detailed analysis of eye movements becomes a unique opportunity to detect the presence of different injuries in the nervous center. These injuries involve a wide variety of neurological diseases including parkinsonian syndromes [1], amyotrophic lateral sclerosis [2], Huntingtons disease [3] Alzheimer disease [4] and minimal hepatic encephalopathy [5], among others. In recent years, the video-oculography (VOG) has become the most widely used eye-movements assessment method, as it is the only one considered non-invasive; also it allows for easy coordination of test design and stimuli provision that make it possible to automatically analyse the data [6], [7].

In order to be able to measure alterations in eye movements, the measurement of them must be performed with high precision and accuracy. That is why methods that use extracted eye features, such as pupil center or eye reflections, called feature-based methods, are the most popular approaches to gaze estimation in these kinds of applications. Feature-based methods, more precisely 2D regression-based methods, use geometrically derived eye features from high-resolution eye images captured by zooming in on the user's eyes. Then, they find a mapping function from the 2D feature space to gaze directions or the computer screen coordinates. These techniques are widely used and achieve really good results, but they have one major issue: the head of the person must remain motionless, otherwise there will be large errors between the actual and estimated directions. In order to avoid these errors, head restraint systems are often used. In spite of the fact that the head movements are restricted, in people with certain neurological diseases, it is possible to observe some involuntary movements. Being able to measure these head movements would be very important not only to correct the gaze estimation errors but also because these measurements seem to be another indicators of the presence or progress of certain diseases.

As of today, despite its importance for clinical diagnosis, there are not many studies about gaze estimation techniques that considers both the head and eye movements in the detailed conditions. We have also not found any research work aimed at detecting short involuntary head movements for future analysis. In this paper we tackle the measurement of the head position with the goal of being able to register small head movements only from videos of the patient's eye. With this aim, as a first stage we develop a method for estimating the eye corners coordinates. Since the eye corners are the most stable points in front of the eyeball and eyelids movements, the changes in their position may be used as a reference for the head movements estimation.

Facial landmark detection is a topic of much interest these days. Different algorithms aim to automatically identify the locations of the facial key points on facial images or videos for different applications. Most of them are designed for landmarks detection throughout the entire face [8], [9] and

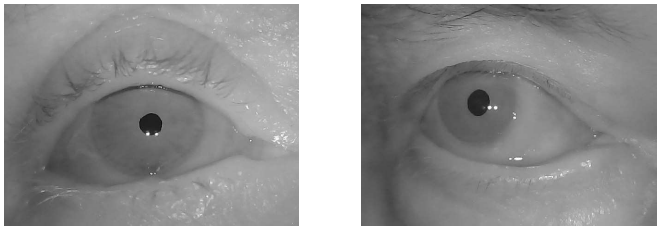
therefore require the whole face to be present in the image, regardless of variations in position, angle or illumination. Others are intended to detect the eye's corners exclusively by extracting the eye region from a given face image, mainly for low cost gaze tracking systems [10], [11]. In such circumstances, the extracted region is usually small in comparison to the image size and therefore the eye has a poor resolution. Contrary to this, in the studied application, the eyes are recorded directly at very high resolution. Therefore, their appearance and the their corners shape change significantly through the subjects or through small eye movements. This is why the goal of regressing the exact position of the corner becomes more challenging.

In recent time, there is a trend to shift from traditional methods to deep learning based methods, more specifically Convolutional Neural Network (CNN) for the task of facial landmark detection and tracking [12]–[14]. In this work, the problem is handled as a regression problem and a deep CNN is proposed to regress the landmark positions from the image appearance. Particularly, with the aim of achieving high precision, we cascade two levels of convolutional neural networks to make a coarse-to-fine prediction of the eye corners position in each frame.

Another point to consider, is the fact that many works address the landmark detection task in the same way for both static images and videos. Thus, when detection is performed in each individual frame, information from preceding frames is not used. Instead, the location of facial landmarks in preceding frames could be used to make it easier to find the facial landmarks in the current frame. One way to add temporal information consists of *tracking* the landmark region instead of detecting it each time. The problem with the direct application of a conventional tracker is that they are sensitive to appearance changes. At high image resolutions, when the person blinks the eye's corner region appearance changes completely and the tracker get lost. In these cases it is really complex to recover the region-of-interest (ROI) tracked and when is recovered, the position is not exactly the same, decreasing the accuracy. Alternatively, in this work we propose a mix approach: during the first frames a tracker is initialized with the two level networks prediction, afterwards the first network is no longer used. Instead, the corner patch is tracked along the video and the fine estimation is obtained with the second network from these patches. This approach showed to provide a fast estimation and a more accurate measurement. The latter is accomplished since when the eye's appearance does not change, the second level network receive the same patch and give a similar landmark estimation.

Finally, we can sum up the main contributions of this paper as follows.

- We created a specific database for this type of videos.
- We designed a coarse-to-fine cascaded convolutional neural network in order to accurately regress the coordinates of the eye corner.
- We added temporal information to increase accuracy and decrease computation time.



(a) (b)

Fig. 1: Eye images from different patients.

II. Data

Unlike the large number of facial landmark datasets available [15]–[17], there are no public datasets for this kind of application. Therefore, it was necessary to generate a customized database by selecting eye images with accurate eye related landmark labels for training and testing, and annotate them with ground truth. Eighteen videos were recorded from eighteen different patients using the Oscann device [18]. An infrared camera and an image resolution of 640x480 pixels were used. As these clinical studies are carry out on either of the two eyes and the choice is made by the doctor, the videos were recorded from both eyes. Therefore it was decided to mirror some of them in order to make them all look the same. Figure 1 shows an example of eye images from different patients.

Then, fifteen frames per video were carefully selected. Due to the large size of the eye in the image, its shape varies greatly depending on the person and in the presence of eye movements or blinks. As a consequence and in order to have a wider range of training data, those frames in which the eye position and the eyeball direction differ as much as possible were chosen. Each frame was identified with the patient number, so that we can split the data in training and validation sets without blending the patient. Since the eye's corner are kept relatively stable again the eye deformations, its accurate detection is critical to estimate the head movements only from the videos. To this end, the ground truth was carefully generated by hand for each eye's corner.

III. Methodology

A. Pre-processing

One of the main challenges faced in implementing this landmark regression cascade is the limited amount of available data. Deep learning methods have outperformed state of the art in various tasks but, due to the large number of parameters to be learned in the model, they require large amount of data. One way to add more data to the training process without the need to collect it is to perform *data augmentation*. It consists of a series of methods applied to each image in a random way during training, in order to train the networks in a robust manner avoiding overfitting at the same time. The individual transformations considered are:

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
A. Larrazabal, C. García Cena & C. E. Martínez; "Eye corners tracking for head movement estimation"
IEEE International Work Conference on Bioinspired Intelligence (IWOBI 2019), 2019.

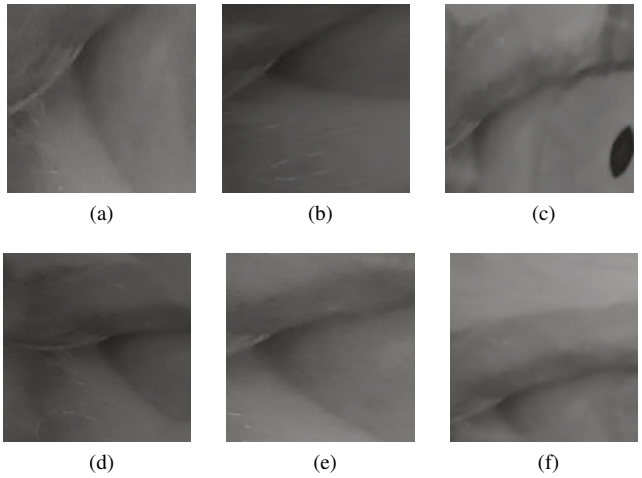


Fig. 2: Different appearance obtained from data augmentation techniques.

- Image rotations by an angle θ between $\pm 40°$ about the centre.
- Gaussian blurring with a 5×5 kernel.
- Change in brightness level.
- Changes in the image shape applying an affine transformation with random matrices.
- Elastic deformations [19].

For training the second level network, different ROIs were obtained from the original training images using the eye's corner ground truth as the reference point. To make the network more robust to the varying position of the eye corner, in each frame the ROI was cropped with a random offset from the centered reference point. In this way, all possible corner positions could be learned from the network. In addition, the ROIs were cropped with random width and height, and then interpolated to the pre-established size. This also varies the eye corner aspect and adds new data to the training process. After that, the data augmentation techniques previously explained were performed to these patches. Figure 2 shows six different patches generated by applying these transformations to a single frame. It should be noted that both, the appearance and the position of the corner, vary considerably. Finally, the intensity of the transformed images was normalized between 0 and 1 to be used as network inputs.

In the augmentation process, the landmarks coordinates were mapped to their new positions. Lastly, these ground truth coordinates lx and ly were normalized between -1 and 1 using the following equations:

$$lxn = \frac{lx - 0.5 \cdot Iw}{0.5 \cdot Iw} \quad lyn = \frac{ly - 0.5 \cdot Ih}{0.5 \cdot Ih} \tag{1}$$

where Iw and Ih are the image (or patch) width and height, respectively.

B. Convolutional Neural Network Structure

For landmark regression, we cascade two levels of convolutional neural networks to make a coarse-to-fine prediction.

TABLE I: Network architecture

	First level network		Second level network	
Block	**Layer (type)**	**Filter size**	**Layer (type)**	**Filter size**
Input	eye image (640x480)		image patch (140x140)	
Conv 1	convolutional	3x3(32)	convolutional	3x3(32)
	convolutional	3x3(32)	convolutional	3x3(32)
	Max pooling		Max pooling	
Conv 2	convolutional	3x3(64)	convolutional	3x3(64)
	convolutional	3x3(64)	convolutional	3x3(64)
	Max pooling		Max pooling	
Conv 3	convolutional	3x3(128)	convolutional	3x3(128)
	convolutional	3x3(128)	convolutional	3x3(128)
	Max pooling		Max pooling	
Conv 4	convolutional	3x3(128)	convolutional	3x3(128)
	convolutional	3x3(128)	convolutional	3x3(128)
	Max pooling		Max pooling	
Conv 5	convolutional	3x3(256)		
	convolutional	3x3(256)		
	Max pooling			
Conv 6	convolutional	3x3(256)		
	convolutional	3x3(256)		
	Max pooling			
Fully connected	flatten		flatten	
	dense	600	dense	600
	dense	300	dense	300
	dense	2	dense	2

The first level network makes an initial prediction of the eye corner landmarks. The second level network receives the image patch cropped from the original image centered in the first prediction, and implement a local refinement of the landmark coordinate. Since not all videos show the complete eye, the network was designed to estimate only the coordinates of one eye corner at a time. After preliminary experiments, the selected network architecture detailed in Table I was used. The ReLu activation function was applied after each convolutional layer and after the first two dense layer. In the last layer, a linear activation function was used.

C. Training

The first and second networks were trained separately. During training, the coarse regression network received the 640x480px augmented images as inputs and the refinement network received the 140x140px patches generated as it was detailed above. In some videos the eye does not appear complete within the image. This is why only sixteen of the patients could be used to train the inner corner model and seventeen could be used to train the outer corner model. The following steps were shared by both CNN.

1) Cross-Validation: An eigth-fold cross-validation [20] approach was used in this study. The data were split into eight folders depending on the patient number, i.e. in each iteration two patients were left for testing and the rest was used to train the neural network. For the outer corner models, as we had an odd number of patients, three of them were used for validation

in the first run. The partition was designed in this way to test the generalization ability of the network to correctly regress the position of the eye corner on previously unseen patients. The models were then compared using the average and the standard deviation of the eight folds.

2) Network parameters: The network was trained to minimise the mean square error (mse) between the landmarks labels and predictions applying the Adam Optimizer. To select the training parameters, a grid was designed in which different learning rates, batch sizes, patch sizes and dropout rates were tested. The best performance was achieve with learning rate: 0.0001; batch size: 16; patch size for the second network: 140x140px.

3) Testing: For testing, the original 15 frames of each patient were used following the validation scheme. To this end, the first network made the coarse prediction and, centered in its prediction, the patch was cropped and it was used as input for the second network which regressed the final coordinate. This prediction was transformed to pixel value by means of:

$$\begin{aligned} lxp &= lxn \cdot 0.5 \cdot Pw + 0.5 \cdot Pw + l \\ lyp &= lyn \cdot 0.5 \cdot Ph + 0.5 \cdot Ph + t \end{aligned} \tag{2}$$

being l and t the right and top patch positions and (lxp, lyp) the eye corner coordinate in pixels. For each model, the mean error and the standard deviation were calculated among all patients.

In addition, a subjective analysis with videos was carried out. Due to the laboriousness of annotating hundreds of data, it is really difficult to have an objective performance evaluation over the large videos. Furthermore, some aspects such as the tremor of the estimated point around the corner are complex to evaluate only by means of these metrics, but can be easily evaluated by inspection.

4) Tracking: The tracker used to follow the patch throughout the video was the Kernelized Correlation Filter (KCF) from OpenCV3 [21]. To initialize the tracker, the two-step network prediction was applied to the first frame and the ROI centered in its prediction was selected. After that, the first part of the network was no longer used and the input patch to the second network was update with this tracker along the videos.

IV. Results

In Figure 3 it can be seen a diagram of the entire process applied as an example to a particular frame. It is possible to observe how the patch is cropped on test time and how the second network improve the first network prediction. Figure 4 shows the mean error and standard deviation of the predictions given by the first and second networks on the complete dataset. The results show separately the predictions made on the inner and the outer corner of the eye. It can be seen that the second network greatly improves the results of the first one achieving a significantly better prediction.

Despite the fact that the mean accuracy obtained was quite good, and it improved even more by adding the tracker, the jiterring did not disappear completely. For this reason, we decided to add an off-line filter that smooths the prediction over time [22]. This feature was subjectively evaluated and it was found that the filter significantly improves this undesirable effect. These qualitative results for difficult validation videos can be seen in the videos of the following link: *https://agostinal.github.io/Corner-detector-project*. It is important to point out that for this application it is not necessary to work online since the final purpose is to use the predictions to post-process an eye tracking algorithm.

Finally, the tracking instance was also evaluated subjectively throughout the videos, finding that it is robust to blinking and rapid eye movements. Furthermore, the computation time was measured during each video both applying the cascade to each frame and following the tracking approach where only the fine-convolutional neural network is applied to each frame. The analysis was made in CPU as the algorithm it is thought to be used in a doctor office. It was found that applying patch tracking, the processing time was reduced 4 times. While on average the double cascade lasts 300 ms per frame, applying the tracker reduces the time per frame to 70 ms. This becomes very significant if it is considered that the videos are taken at 100 fps.

V. Conclusion

A model for eye corner coordinate estimation was constructed from scratch. The model was designed for a particular application for which it was necessary to collect and annotate the complete database. A coarse-to-fine cascaded convolutional neural network was implemented for static-frames landmark regression. Temporal information was added both from the patch tracking along the videos and from the implementation of a temporal filter to smooth the estimation. The experimental results confirmed the efficacy of the proposed method in different videos. With these promising results, we expect to provide a valuable information to be able to measure changes in the head position during clinical routine trials in patients.

Acknowledgments

The authors gratefully acknowledge NVIDIA Corporation with the donation of the Titan Xp GPU used for this research, and the support of UNL (CAID-PIC-50420150100098LI) and ANPCyT (PICT 2016-0651).

References

[1] Elena Pretegiani and Lance M Optican. Eye Movements in Parkinson's Disease and Inherited Parkinsonian Syndromes. *Frontiers in Neurology*, 8:592, 2017.

[2] Colette Donaghy, Matthew J Thurtell, Erik P Pioro, J Mark Gibson, and R John Leigh. Eye movements in amyotrophic lateral sclerosis and its mimics: a review with illustrative cases. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(1):110–116, 2011.

[3] Stephen L Hicks, Matthieu PA Robert, Charlotte VP Golding, Sarah J Tabrizi, and Christopher Kennard. Oculomotor deficits indicate the progression of Huntington's disease. In *Progress in brain research*, volume 171, pages 555–558. Elsevier, 2008.

[4] Gerardo Fernández, Pablo Mandolesi, Nora P Rotstein, Oscar Colombo, Osvaldo Agamennoni, and Luis E Politi. Eye movement alterations during reading in patients with early alzheimer disease. *Investigative ophthalmology & visual science*, 54(13):8345–8352, 2013.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
A. Larrazabal, C. García Cena & C. E. Martínez; "Eye corners tracking for head movement estimation"
IEEE International Work Conference on Bioinspired Intelligence (IWOBI 2019), 2019.

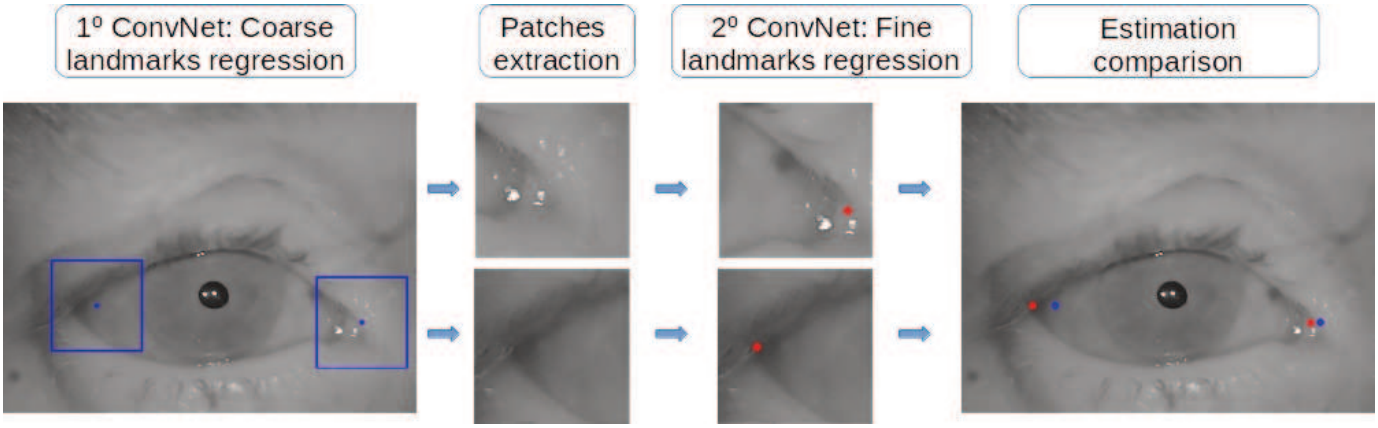


Fig. 3: Eye corner coordinate estimation from the first level network (blue) and second level network (red)

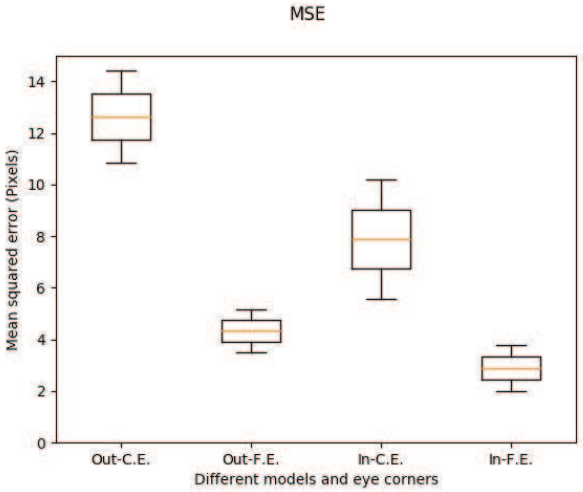


Fig. 4: MSE and standard deviation obtained from the first (C-E) and the second (F-E) networks.

[5] Sara Montagnese, Harriet M Gordon, Clive Jackson, Justine Smith, Patrizia Tognella, Nutan Jethwa, R Michael Sherratt, and Marsha Y Morgan. Disruption of smooth pursuit eye movements in cirrhosis: relationship to hepatic encephalopathy and its treatment. *Hepatology*, 42(4):772–781, 2005.

[6] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2010.

[7] Anuradha Kar and Peter Corcoran. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*, 5:16495–16519, 2017.

[8] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, pages 1–28, 2017.

[9] Chao Gou, Yue Wu, Kang Wang, Kunfeng Wang, Fei-Yue Wang, and Qiang Ji. A joint cascaded framework for simultaneous eye detection and eye state estimation. *Pattern Recognition*, 67:23–31, 2017.

[10] Jose Javier Bengoechea, Juan J Cerrolaza, Arantxa Villanueva, and Rafael Cabeza. Evaluation of accurate eye corner detection methods for gaze estimation. *Journal of Eye Movement Research*, 7(3), 2014.

[11] Yiu-ming Cheung and Qinmu Peng. Eye gaze tracking with a web camera in a desktop environment. *IEEE Transactions on Human-Machine Systems*, 45(4):419–430, 2015.

[12] Haoqiang Fan and Erjin Zhou. Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35, 2016.

[13] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[14] Yue Wu, Tal Hassner, KangGeon Kim, Gerard Medioni, and Prem Natarajan. Facial landmark detection with tweaked convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):3067–3074, 2018.

[15] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011.

[16] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[17] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 896–903, 2013.

[18] Erik Hernández, Santiago Hernández, David Molina, Rafael Acebrón, and Cecilia E García Cena. Oscann: Technical characterization of a novel gaze tracking analyzer. *Sensors*, 18(2):522, 2018.

[19] Patrice Y Simard, Dave Steinkraus, and John C Platt. Best practices for convolutional neural networks applied to visual document analysis. In *null*, page 958. IEEE, 2003.

[20] Richard O Duda, Peter E Hart, David G Stork, et al. Pattern classification. 2nd. *Edition. New York*, 55, 2001.

[21] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *European conference on computer vision*, pages 702–715. Springer, 2012.

[22] Damien Garcia. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational statistics & data analysis*, 54(4):1167–1178, 2010.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
A. Larrazabal, C. García Cena & C. E. Martínez; "Eye corners tracking for head movement estimation"
IEEE International Work Conference on Bioinspired Intelligence (IWOBI 2019), 2019.

Bibliografía

[1] Dinesh D Patil and Sonal G Deore. Medical image segmentation: a review. *International Journal of Computer Science and Mobile Computing*, 2(1):22–27, 2013.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[3] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[4] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[5] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[6] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *arXiv preprint arXiv:1903.02026*, 2019.

[7] Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. In *International conference on medical image computing and computer-assisted intervention*, pages 10–18. Springer, 2016.

[8] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506, 2016.

[9] Heung-Il Suk, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Deep learning in diagnosis of brain disorders. In *Recent Progress in Brain and Cognitive Engineering*, pages 203–213. Springer, 2015.

[10] Sandra Vieira, Walter HL Pinaya, and Andrea Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, 74:58–75, 2017.

[11] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Regressing heatmaps for multiple landmark localization using cnns. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–238. Springer, 2016.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. of MICCAI*, 2015.

[13] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, , et al. Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61 – 78, 2017.

[14] Mahsa Shakeri, Stavros Tsogkas, Enzo Ferrante, Sarah Lippe, Samuel Kadoury, Nikos Paragios, et al. Sub-cortical brain structure segmentation using F-CNN's. In *Proc. of ISBI*, 2016.

[15] Aïcha BenTaieb and Ghassan Hamarneh. Topology aware fully convolutional networks for histology gland segmentation. In *Proc. of MICCAI*, pages 460–468. Springer, 2016.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017.

[17] Davood Karimi and Ali Gholipour. Improving calibration and out-of-distribution detection in medical image segmentation with convolutional neural networks. *arXiv preprint arXiv:2004.06569*, 2020.

[18] Robert Robinson, Vanya V Valindria, Wenjia Bai, Ozan Oktay, Bernhard Kainz, Hideaki Suzuki, Mihir M Sanghvi, Nay Aung, José Miguel Paiva, Filip Zemrak, et al. Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study. *Journal of Cardiovascular Magnetic Resonance*, 21(1):1–14, 2019.

[19] Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. Is segmentation uncertainty useful? In *International Conference on Information Processing in Medical Imaging*, pages 715–726. Springer, 2021.

[20] Masoud S Nosrati and Ghassan Hamarneh. Incorporating prior knowledge in medical image segmentation: a survey. *arXiv preprint arXiv:1607.01092*, 2016.

[21] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, et al. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. *IEEE TMI*, 37(2):384–395, 2018.

[22] Hariharan Ravishankar, Rahul Venkataramani, Sheshadri Thiruvenkadam, Prasad Sudhakar, and Vivek Vaidya. Learning and incorporating shape models for semantic segmentation. In *Proc. of MICCAI*, 2017.

[23] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11:3371–3408, 2010.

[24] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. *Semi-supervised learning*. MIT Press, 2009.

[25] Philipp Krähenbühl and Vladlen Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Proc. of Nips*, 2011.

[26] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. In *Medical Imaging with Deep Learning*, 2018.

[27] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

[28] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[29] Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282, 2020.

[30] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

[31] Asa Cooper Stickland and Iain Murray. Diverse ensembles improve calibration. In *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

[32] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. *arXiv preprint arXiv:2002.09437*, 2020.

[33] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020.

[34] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proc. of Fourth International Conference on 3D Vision (3DV)*, 2016.

[35] Jörg Sander, Bob D de Vos, Jelmer M Wolterink, and Ivana Išgum. Towards increased trustworthiness of deep learning segmentation methods on cardiac mri. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094919. International Society for Optics and Photonics, 2019.

[36] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

[37] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *ICLR*, 2020.

[38] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. In *NeurIPS*, 2020.

[39] Samarth Sinha, Homanga Bharadhwaj, Anirudh Goyal, Hugo Larochelle, Animesh Garg, and Florian Shkurti. Dibs: Diversity inducing information bottleneck in model ensembles. In *AAAI*, 2020.

[40] Wonsik Kim, Bhavya Goyal, Kunal Chawla, Jungmin Lee, and Keunjoo Kwon. Attention-based ensemble for deep metric learning. In *ECCV*, pages 736–751, 2018.

[41] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[42] Mobarakol Islam and Ben Glocker. Spatially varying label smoothing: Capturing uncertainty from expert annotations. In *International Conference on Information Processing in Medical Imaging*, pages 677–688. Springer, 2021.

[43] Babajide O Ayinde, Tamer Inanc, and Jacek M Zurada. Regularizing deep neural networks by enhancing diversity in feature extraction. *IEEE transactions on neural networks and learning systems*, 30(9):2650–2661, 2019.

[44] Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu. Orthogonal convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11505–11515, 2020.

[45] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4:170117, 2017.

[46] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[47] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[48] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019.

[49] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Caizi Li, Qianqian Tong, Weixin Si, et al. A global benchmark of algorithms for segmenting late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, 2020.

[50] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[51] Byron C Wallace and Issa J Dahabreh. Improving class probability estimates for imbalanced data. *Knowledge and information systems*, 41(1):33–52, 2014.

[52] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[53] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, pages 1–28, 2017.

[54] Chao Gou, Yue Wu, Kang Wang, Kunfeng Wang, Fei-Yue Wang, and Qiang Ji. A joint cascaded framework for simultaneous eye detection and eye state estimation. *Pattern Recognition*, 67:23–31, 2017.

[55] Jose Javier Bengoechea, Juan J Cerrolaza, Arantxa Villanueva, and Rafael Cabeza. Evaluation of accurate eye corner detection methods for gaze estimation. *Journal of Eye Movement Research*, 7(3), 2014.

[56] Yiu-ming Cheung and Qinmu Peng. Eye gaze tracking with a web camera in a desktop environment. *IEEE Transactions on Human-Machine Systems*, 45(4):419–430, 2015.

[57] Haoqiang Fan and Erjin Zhou. Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47:27–35, 2016.

[58] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[59] Yue Wu, Tal Hassner, KangGeon Kim, Gerard Medioni, and Prem Natarajan. Facial landmark detection with tweaked convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):3067–3074, 2018.

[60] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *European conference on computer vision*, pages 702–715. Springer, 2012.

[61] Damien Garcia. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational statistics & data analysis*, 54(4):1167–1178, 2010.

[62] Erik Hernández, Santiago Hernández, David Molina, Rafael Acebrón, and Cecilia E García Cena. Oscann: Technical characterization of a novel gaze tracking analyzer. *Sensors*, 18(2):522, 2018.

[63] Patrice Y Simard, Dave Steinkraus, and John C Platt. Best practices for convolutional neural networks applied to visual document analysis. In *null*, page 958. IEEE, 2003.

Doctorado en Ingeniería
Mención en Inteligencia Computacional, Señales y Sistemas

Título de la obra:

Aprendizaje profundo aplicado al análisis de imágenes médicas

Autora: Agostina Juliana Larrazabal

Lugar: Santa Fe, Argentina

Palabras Claves:

Autocodificadores,
Redes neuronales convolucionales,
Segmentación de imágenes,
Aprendizaje de representaciones,
Ensambles de redes neuronales,
Restricciones ortogonales,
Penalización de entropía.