UNIVERSIDAD
NACIONAL
DE COLOMBIA

A Deep Learning Approach for Image-based Semantic Segmentation with Preserved Interpretability

Juan Carlos Aguirre Arango

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Department of Electric, Electronic and Computing Engineering
Manizales, Colombia
2023

A Deep Learning Approach for Image-based Semantic Segmentation with Preserved Interpretability

Juan Carlos Aguirre Arango

Dissertation submitted as a partial requirement to receive the grade of:

Master in Engineering - Industrial Automation

Advisor:

Prof. Andrés Marino Álvarez-Meza, Ph.D.

Co-advisor:

Prof. Germán Castellanos-Domínguez, Ph.D.

Academic research group:

Signal Processing and Recognition Group - SPRG

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Department of Electric, Electronic and Computing Engineering
Manizales, Colombia
2023

Un enfoque de aprendizaje profundo para la segmentación semántica de imágenes conservando interpretabilidad

Juan Carlos Aguirre Arango

Disertación presentada como requisito parcial para recibir el título de:

Magíster en Ingeniería - Automatización Industrial

Director:

Prof. Andrés Marino Álvarez-Meza, Ph.D.

Codirector:

Prof. Germán Castellanos-Domínguez, Ph.D.

Grupo de investigación:

Grupo de Control y Procesamiento Digital de Señales - GCPDS

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitecura
Departamento de Ingeniería Eléctrica, Electrónica y Computación
Manizales, Colombia
2023

ACKNOWLEDGEMENTS

I would like to express my heartfelt appreciation to my parents for their unwavering support and encouragement.

I would like to express my gratitude to two esteemed professors, Andrés Marino Álvarez Meza, and Germán Castellanos Domínguez, for providing me with invaluable direction during my research. I would also like to thank the members of the Signal Processing and Recognition Group (SPRG) at the Manizales branch of Universidad Nacional de Colombia for their thoughtful suggestions and educational discussion.

Finally, I recognize that this research would not have been possible without the support given by the projects: *"Herramienta de apoyo a la predicción de los efectos de anestésicos locales vía neuroaxial epidural a partir de termografía por infrarrojo"* (Code 111984468021) funded by MINCIENCIAS, and *"Sistema prototipo de visión por computador utilizando aprendizaje profundo como soporte al monitoreo de zonas urbanas desde unidades aéreas no tripuladas"* (Hermes Code 55261), funded by Universidad Nacional de Colombia.

Juan Carlos Aguirre Arango
2023

ABSTRACT

Semantic segmentation is pivotal in various industries, showcasing its significant impact across numerous applications. Semantic segmentation offers invaluable insights that drive advancements in fields such as autonomous driving, surveillance, robotics, and augmented reality by enabling precise identification and labeling of objects within an image. Accurate segmentation of objects allows autonomous vehicles to navigate complex environments, surveillance systems to detect and track specific objects, robots to manipulate objects efficiently, and augmented reality applications to seamlessly blend virtual objects with the real world. However, in the medical industry, the importance of semantic segmentation has become truly profound. Medical imaging techniques, such as computerized tomography scans and magnetic resonance imaging, generate vast amounts of data that require meticulous annotation for analysis. Manual annotation is a time-consuming and resource-intensive process, leading to diagnosis and treatment planning delays. Semantic segmentation techniques have the potential to automate this process, facilitating faster and more accurate analysis of medical images, thereby enhancing patient care and reducing the burden on healthcare professionals. Moreover, in medical applications, the need for interpretability is critical. Understanding and interpreting the segmentation results is vital for clinicians to make informed decisions. Interpretable semantic segmentation techniques provide transparency and insights into the segmentation process, ensuring that medical professionals can trust and validate the results for accurate diagnosis and treatment.

Medical image analysis faces several challenges, with one of the primary obstacles being the limited availability of datasets specifically tailored for training segmentation models. These models require large and diverse datasets to learn the intricate patterns and features of medical images accurately. However, due to the sensitive nature of medical data and the need for expert annotations, obtaining such datasets can be challenging. Another significant challenge arises from the high variability in the region of interest (ROI) within medical imaging. The ROI can differ significantly from one patient to another due to variations in anatomy, pathology, and imaging parameters. This variability leads to differences in shape, size, and texture, making it difficult for segmentation models to delineate and analyze the regions of interest accurately. Consequently, ensuring consistent and reliable segmentation results across diverse medical images remains a critical challenge. Furthermore, there is a pressing need for systematic and quantitative evaluations of interpretability in deep learning-based segmentation models. Without such evaluations, trusting and relying on these models for clinical decision-making becomes challenging. Medical practitioners must comprehensively understand how and why these models arrive at their conclusions to incorporate them into their practice confidently. The absence of standardized evaluation methods impedes progress in building interpretable and trustworthy medical image analysis systems.

This work addresses challenges in medical image segmentation. Our contributions include optimizing Random Fourier Features for spatial data through gradient descent named CRFFg, enhancing shallow encoder-decoder models for semantic segmentation, and proposing quantitative measures for interpretability. CRFFg takes advantage of the generalization properties of kernel methods and enhances data efficiency for spatial data derived from convolutions, mitigating low sample size and overfitting. To address shape, size, and texture variability in semantic segmentation across patients and imaging protocols, we incorporate a CRFFg layer

into the skip connection of the encoder-decoder models. This improves the representation of low-level features from the encoder and their fusion in the decoder, specifically targeting the challenges of ROI variability. Interpretability is crucial in medical semantic segmentation, but deep learning models present challenges. To enhance interpretability, we propose quantitative measures: CAM-based Cumulative Relevance assesses the location of relevance in specific regions of interest, Mask-based Cumulative Relevance evaluates sensibility across multiple regions of interest, and CAM-Dice measures the homogeneity of relevance in interest regions. These measures provide objective and comprehensive evaluations, surpassing visual inspection and qualitative analysis.

The proposed work has been tested in a medical image application where the mentioned problems occur, specifically in the segmentation of feet for monitoring the effectiveness of analgesia in the obstetric environment. This is achieved by monitoring changes in temperature at the soles of the feet. The proposed methodology demonstrates comparable performance with standard methods while also enhancing interpretability. It is important to note that this project is being developed in conjunction with SES Hospital Universitario de Caldas, under the name ”Sistema prototipo de visión por computador utilizando aprendizaje profundo como soporte al monitoreo de zonas urbanas desde unidades aéreas no tripuladas” (Hermes Code 55261). The project is funded by Universidad Nacional de Colombia.

In our future research, we have identified several promising avenues that can advance our work. By analyzing the spectral representation of the CRFFg layer, we aim to uncover hidden patterns and gain a deeper understanding of the subject. Additionally, incorporating Bayesian approximation techniques will enable us to enhance our decision-making and optimization strategies. We also plan to employ regularization techniques based on the proposed measures, which will effectively address overfitting issues and improve the model’s performance by focusing on the desired behavior of the discriminative regions. By pursuing these paths, we aim to enhance our approach’s overall effectiveness and reliability significantly, thereby pushing the boundaries of knowledge in this field.

Keywords: Infrared Thermal Segmentation, Regional Neuraxial Analgesia, Deep Learning, Random Fourier Features, Class Activation Maps

RESUMEN

La segmentación semántica es fundamental en varias industrias y muestra su impacto significativo en numerosas aplicaciones. La segmentación semántica ofrece información valiosa que impulsa los avances en campos como la conducción autónoma, la vigilancia, la robótica y la realidad aumentada al permitir la identificación y el etiquetado precisos de los objetos dentro de una imagen. La segmentación precisa de objetos permite que los vehículos autónomos naveguen en entornos complejos, los sistemas de vigilancia detecten y rastreen objetos específicos, los robots manipulen objetos de manera eficiente y las aplicaciones de realidad aumentada combinen a la perfección objetos virtuales con el mundo real. Sin embargo, en la industria médica, la importancia de la segmentación semántica se ha vuelto verdaderamente profunda. Las técnicas de imágenes médicas, como las tomografías computarizadas y las resonancias magnéticas, generan grandes cantidades de datos que requieren una anotación meticulosa para su análisis. La anotación manual es un proceso que requiere mucho tiempo y recursos, lo que genera retrasos en el diagnóstico y la planificación del tratamiento. Las técnicas de segmentación semántica tienen el potencial de automatizar este proceso, facilitando un análisis más rápido y preciso de imágenes médicas, mejorando así la atención al paciente y reduciendo la carga de los profesionales de la salud. Además, en aplicaciones médicas, la necesidad de interpretabilidad es crítica. Comprender e interpretar los resultados de la segmentación es vital para que los médicos tomen decisiones informadas. Las técnicas de segmentación semántica interpretables brindan transparencia e información sobre el proceso de segmentación, lo que garantiza que los profesionales médicos puedan confiar y validar los resultados para un diagnóstico y tratamiento precisos.

El análisis de imágenes médicas enfrenta varios desafíos, y uno de los principales obstáculos es la disponibilidad limitada de conjuntos de datos específicamente diseñados para entrenar modelos de segmentación. Estos modelos requieren conjuntos de datos grandes y diversos para aprender con precisión los patrones y características intrincados de las imágenes médicas. Sin embargo, debido a la naturaleza confidencial de los datos médicos y la necesidad de anotaciones de expertos, la obtención de dichos conjuntos de datos puede ser un desafío. Otro desafío significativo surge de la alta variabilidad en la región de interés (ROI) dentro de las imágenes médicas. El ROI puede diferir significativamente de un paciente a otro debido a variaciones en la anatomía, la patología y los parámetros de imagen. Esta variabilidad conduce a diferencias en forma, tamaño y textura, lo que dificulta que los modelos de segmentación delineen y analicen las regiones de interés con precisión. En consecuencia, garantizar resultados de segmentación consistentes y confiables en diversas imágenes médicas sigue siendo un desafío crítico. Además, existe una necesidad apremiante de evaluaciones sistemáticas y cuantitativas de la interpretabilidad en modelos de segmentación basados en aprendizaje profundo. Sin tales evaluaciones, confiar en estos modelos para la toma de decisiones clínicas se convierte en un desafío. Los médicos deben comprender de manera integral cómo y por qué estos modelos llegan a sus conclusiones para incorporarlos a su práctica con confianza. La ausencia de métodos de evaluación estandarizados impide el progreso en la construcción de sistemas de análisis de imágenes médicas interpretables y confiables.

Este trabajo aborda los desafíos en la segmentación de imágenes médicas. Nuestras contribuciones incluyen la optimización de las características aleatorias de Fourier para datos espaciales a través del descenso de gradiente denominado CRFFg, la mejora de los modelos de codificador-decodificador poco profundos para la segmentación semántica y la propuesta de medidas cuantitativas para la

interpretabilidad. CRFFg toma ventajas de las propiedades de generalización de los métodos kernel y mejora la eficiencia de datos para datos espaciales derivados de convoluciones, mitigando el tamaño de muestra bajo y el sobreajuste. Para abordar la variabilidad de forma, tamaño y textura en la segmentación semántica entre pacientes y protocolos de imágenes, incorporamos una capa CRFFg en la conexión de salto de los modelos codificador-decodificador. Esto mejora la representación de características de bajo nivel del codificador y su fusión en el decodificador, apuntando específicamente a los desafíos de la variabilidad del ROI. La interpretabilidad es crucial en la segmentación semántica médica, pero los modelos de aprendizaje profundo presentan desafíos. Para mejorar la interpretabilidad, proponemos medidas cuantitativas: la relevancia acumulada basada en CAM evalúa la ubicación de relevancia en regiones específicas de interés, la relevancia acumulada basada en máscara evalúa la sensibilidad en múltiples regiones de interés y CAM-Dice mide la homogeneidad de relevancia en regiones de interés. Estas medidas proporcionan evaluaciones objetivas y completas, superando la inspección visual y el análisis cualitativo.

El trabajo propuesto ha sido probado en una aplicación de imagen médica donde se presentan los problemas mencionados, específicamente en la segmentación de pies para monitorear la efectividad de la analgesia en el medio obstétrico. Esto se logra monitoreando los cambios de temperatura en las plantas de los pies. La metodología propuesta demuestra un rendimiento comparable con los métodos estándar al tiempo que mejora la interpretabilidad. Es importante señalar que este proyecto se está desarrollando en conjunto con SES Hospital Universitario de Caldas, bajo el nombre de ”Sistema prototipo de visión por computador utilizando aprendizaje profundo como soporte al monitoreo de zonas urbanas desde unidades aéreas no tripuladas” (Código Hermes 55261). El proyecto es financiado por la Universidad Nacional de Colombia

En nuestra investigación futura, hemos identificado varias vías prometedoras que pueden avanzar en nuestro trabajo. Al analizar la representación espectral de la capa CRFFg, nuestro objetivo es descubrir patrones ocultos y obtener una

comprensión más profunda del tema. Además, la incorporación de técnicas de aproximación bayesiana nos permitirá mejorar nuestras estrategias de toma de decisiones y optimización. También planeamos emplear técnicas de regularización basadas en las medidas propuestas, que abordarán de manera efectiva los problemas de sobreajuste y mejorarán el rendimiento del modelo al enfocarse en el comportamiento deseado de las regiones discriminatorias. Al seguir estos caminos, nuestro objetivo es mejorar significativamente la eficacia y confiabilidad general de nuestro enfoque, ampliando así los límites del conocimiento en este campo.

Palabras clave: Segmentación Térmica Infrarroja, Analgesia Neuroaxial Regional, Aprendizaje Profundo, Características Aleatorias de Fourier, Mapas de Activación de Clase

Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

CAD Computer-Aided Diagnosis 5, 7

CAM Class Activation Map 22, 24

CAMs Class Activation Maps xix, xxii, 79, 80, 82–84, 86

CNN Convolutional Neural Networks xxi, 8, 12, 13, 16, 17, 44, 53

CRFFg Convolutional RFF gradient xviii, xix, xxi, 28, 29, 45–48, 51–55, 57, 59, 60, 66, 73, 75, 95, 99

DL Deep Learning xxi, 1, 2, 6–8, 10, 16, 29, 69, 71, 72

MIA Medical Image Analysis xxi, 2, 3, 6, 7, 12

RFF Random Fourier Features xviii, xxi, 14, 15, 24, 25, 27, 37, 39–41, 43–47, 51–54

ROI Region of Interest xvii, xxii, 5–7, 9, 12, 16, 17, 19, 27, 82–84, 96

SPRG Signal Processing and Recognition Group vii, 5, 6

SS Semantic segmentation xvii–xix, xxi, xxii, 1, 2, 6, 9–11, 16, 19, 22–25, 29–31, 33, 35, 59, 60

WHO World Health Organization 2

CHAPTER

ONE

INTRODUCTION

1.1 Motivation

Semantic segmentation (SS) is a computer vision technique that involves dividing an image into distinct and meaningful parts or regions, each corresponding to a specific object or class [Mo et al., 2022]. Thanks to advancements in machine learning, specifically Deep Learning (DL), it has become feasible to apply SS to a broad range of applications, including healthcare, automotive, agriculture, satellite imagery, retail, security industries, among others. For example, in healthcare, SS has been used to identify and segment various organs, tumors, or lesions in medical images, which can help with diagnosis and treatment planning [Qureshi et al., 2023, Khan et al., 2021, Soomro et al., 2023]. In autonomous vehicles, SS is essential as it enables the detection and identification of objects on the road, such as pedestrians and traffic signs, which is crucial for safe and efficient driving [Cakir et al., 2022, Tsai et al., 2023, Rizzoli et al., 2022, Burel et al., 2022]. In agriculture, it is helpful for crop monitoring, disease detection, and yield estimation [Anand et al., 2021, Zou et al., 2021, Singh et al., 2021]. In satellite imagery, it is used for land cover mapping, urban planning, and environmental monitoring [Lilay and Taye, 2023, Wieland et al., 2023]. Therefore, SS is a critical component in the development of artificial intelligence and DL systems, which have wide-ranging applications in various industries [Minaee et al., 2022, Quazi and Musa, 2021].

Medical Image Analysis (MIA) heavily relies on SS, which is a critical task for accurately segmenting and labeling different structures and regions of interest within a medical image [Maier-Hein et al., 2018, Li et al., 2021a, Antonelli et al., 2022]. Figure **1-1** illustrates the distribution of tasks in medical image analysis, with segmentation being the predominant task in MIA. For instance, early detection of Alzheimer's, schizophrenia, and Parkinson's disease requires precise identification of physical changes in the hippocampus, which is often segmented on Magnetic Resonance Images [Ataloglou et al., 2019, Liu et al., 2020, Wu and Tang, 2021, Carmo et al., 2021]. A breast cancer diagnosis is another well-known application. Identifying physical changes in breast tissue, like tumors, in X-rays, magnetic imaging, or ultrasound images is crucial to diagnosis on time [Altameem et al., 2022, Fazilov et al., 2022, Jahwar and Abdulazeez, 2022]. Moreover, to diagnose chest-related diseases, such as pneumonia, tuberculosis, and COVID-19, Computed Radiography is the most commonly used medical imaging together with SS techniques [Yıldırım et al., 2023, Alebiosu et al., 2023]. Besides, in maternal health, particularly pain control before labor, monitoring anesthesia through thermographic images is an objective measure of such effectiveness [Bouvet et al., 2020]. Accordingly, SS has become an indispensable tool in modern medical research and practice by enabling more precise and efficient analysis of medical images [Aljabri and AlGhamdi, 2022].

In this sense, maternal health can be highly benefited from MIA. Maternal health, the health of women during pregnancy, childbirth, and the postnatal period, is an important issue, which is one of the key priorities of World Health Organization (WHO), grounded in a human rights approach and linked to efforts on universal health coverage [World Health Organization, 2023]. For example, in Colombia, a

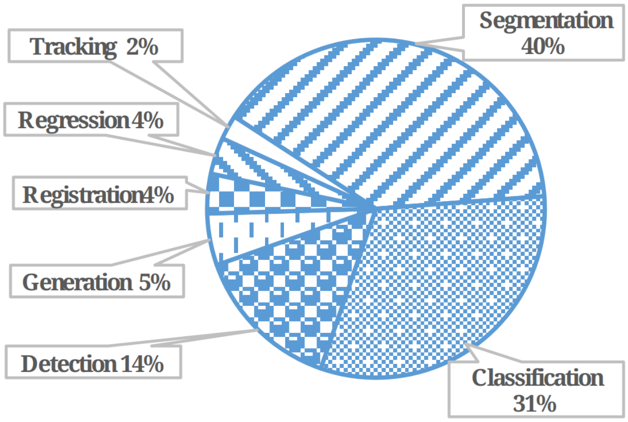


Figure **1-1**. Main tasks in MIA [Li et al., 2021a]

significant number of pregnant women die each year, on average during one day, 40 pregnant women are close to death, most of the time due to preventable causes that are related to hypertension associated with pregnancy, hemorrhage, or infection [Ministerio de Salud y Protección Social de Colombia, 2023]. Moreover, pain management during childbirth is key to ensuring a comfortable and safe delivery for both the mother and baby [Joensuu et al., 2022, Smith et al., 2018]. Childbirth is one of the most painful events [Melesse et al., 2022], and can cause complications that can be dangerous for the mother's and child's safety and wellness [Chughtai et al., 2018, Koyucu and Karaca, 2022]. If the pain is not effectively managed, it can cause stress and anxiety, leading to longer and more difficult labor and increasing the risk of complications such as preterm delivery, cesarean delivery, and postpartum depression [Parise et al., 2021, Cavazos-Rehg et al., 2015]. Therefore, women should be offered appropriate pain management techniques, such as breathing exercises, relaxation techniques, and medication, during labor to help them cope with the pain and have a positive childbirth experience.

Meanwhile, using regional neuraxial analgesia for pain relief during labor is widely recognized as a safe method [Brown et al., 1980, Omer Ibrahim Abdalla et al., 2022]. Compared to other forms of pain relief, regional neuraxial analgesia is considered safe and effective for most women and is associated with lower rates of complications [McCombe and Bogod, 2021]. This technique involves administering medication into the epidural or subarachnoid space in the lower back [Conhaim and Girnius, 2023], which blocks pain signals from the uterus and cervix to the brain. Nonetheless, accurately and quickly assessing its effectiveness is crucial for optimizing healthcare resources and promoting patient well-being [Lee et al., 2022, Gottlieb et al., 2022]. The effectiveness of anesthesia or analgesic block is evaluated using three main modalities: psychophysical, electrophysiological, and imaging techniques. Psychophysical testing evaluates thermosensitive sensory effects through superficial touch, cutaneous pinprick tests, and cold tests [Hoyle and Yentis, 2015]. Although it is the most commonly used modality, its effectiveness depends on subjective patient reports, leading to increased false positive and negative rates [Bruins et al., 2018b]. Electrophysiological testing measures nerve fiber reactions to painful stimuli using electromyography, excitatory or inhibitory reflexes, evoked potentials, electroencephalography, and magnetoencephalography [Chae et al., 2022]. Imaging techniques objectively measure relevant bodily function patterns (such as blood flow, oxygen use, and sugar metabolism) using positron emission tomography, single-photon emission computed tomography, and functional magnetic resonance imaging [Curatolo et al., 2005]. Nevertheless, imaging techniques can be costly, limiting their use, and are generally prohibited in obstetric patients [Whittingham, 2013].

A cost-effective alternative approach involves utilizing thermographic skin images to measure body temperature and predict the distribution and efficacy of epidural anesthesia [Bruins et al., 2018a]. This is achieved by identifying areas of cold sensation [Bruins et al., 2018b]. These areas are related to the dermatomes, an area of skin that is innervated by a single spinal nerve, which acts as a biomarker for the level of pain sensation [Koszewicz et al., 2021]. The use of thermal imaging provides an objective and non-invasive solution to assess the warm modifications resulting from blood flow redistribution after catheter placement [Haren et al., 2013]. However, to accurately assess the effectiveness of epidural analgesia, temperature measurements must be taken from the patient's foot soles. Taking these measurements at different times after catheter placement is crucial to accurately characterize the early thermal changes [Stevens et al., 2006, Werdehausen et al., 2007]. Another significant challenge is the limited data availability together with the high variability of Region of Interest (ROI), particularly in this obstetric environment. This is compounded by the fact that acquiring image data can be even more difficult due to the restricted situation of the mother and the discomfort associated with the process [Willemink et al., 2020, Melesse et al., 2022]. Therefore, developing tools for objective evaluation of the effectiveness of epidural anesthesia under data scarcity plays a critical role in maternal health [Whitburn and Jones, 2020].

In a local context, The Signal Processing and Recognition Group (SPRG) at Universidad Nacional de Colombia has been focused on analyzing biomedical data, aiming to develop machine-learning methodologies to improve Computer-Aided Diagnosis (CAD) systems. For example, their research includes results in diagnosing conditions such as dementia and brain tumors, with the segmentation of the latter achieved through MRI analysis [Bron et al., 2015, Jimenez et al., 2018]. Additionally, SPRG has been utilizing ultrasound images to accurately segment nerve structures, thereby supporting regional anesthesia [Jimenez-Castaño et al., 2021]. Moreover, SPRG has shown an interest in working with computer vision systems for CADs in a variety of investigation and innovations projects (supported by Minciencias, Dirección Nacional de Investigaciones de Manizales, and Vicerrectoría de Investigaciones de la Universidad Nacional de Colombia):

- Desarrollo de un sistema automático de análisis de volumetría cerebral como apoyo en la evaluación clínica de recién nacidos con asfixia perinatal (2019-actual). Financiado por Minciencias.

- Desarrollo de una herramienta de seguimiento de aguja y localización de nervios en ecografía para la práctica de anestesia regional: aplicación al tratamiento de dolor agudo traumático y prevención del dolor neuropático crónico (2019-actual). Financiado por Minciencias

- Caracterización morfológica de estructuras cerebrales por técnicas de imagen para el tratamiento mediante implantación quirúrgica de neuroestimuladores en la enfermedad de Parkinson (2019-actual). Financiado por Minciencias
- Desarrollo de un sistema de apoyo al diagnóstico no invasivo de pacientes con Epilepsia fármaco-resistente asociada a displasias corticales cerebrales: método costo-efectivo basado en procesamiento de imágenes de resonancia magnética (2017-2019). Financiado por Minciencias.

In particular, SPRG has recently expanded its research efforts by collaborating with SES Hospital Universitario de Caldas to develop a novel tool under the project ”Herramienta de apoyo a la predicción de los efectos de anestésicos locales vía neuroaxial epidural a partir de termografía por infrarrojo” (Code 111984468021), which is funded by MINCIENCIAS. This tool uses infrared thermal images to predict the effects of regional neuraxial analgesia on obstetrics patients. The potential impact of this research is substantial, as it could lead to improved patient outcomes and enhance the delivery of medical care in this particular field.

Consequently, developing new SS techniques based on DL with preserved interpretability under scenarios of data scarcity and high variability in the ROI is necessary. Furthermore, effectively developing and implementing these techniques has the potential to yield subsequent tools that can enhance maternal health at SES Hospital de Caldas. For instance, in the context of monitoring anesthesia effectiveness, a shift from subjective assessment to objective evaluation can be achieved by characterizing changes in temperature within a region of interest in the human body. This change not only enhances the quality of service but also ensures more accurate assessments.

1.2 Problem Statement

MIA plays a critical role in diagnosing and treating diseases, as it enables healthcare professionals to obtain valuable insights into the condition of patients [Rashed and

Popescu, 2022, Ghosh et al., 2022]. Yet, the manual analysis can be challenging and time-consuming, often requiring the expertise of highly specialized personnel [Jonaitytė and Petkevičius, 2021, Yang and Yu, 2021]. Furthermore, the complex structures and features present in medical images make it difficult to achieve high levels of accuracy and consistency in the analysis [Scalco and Rizzo, 2017]. This can lead to errors and inconsistencies in diagnosis and treatment, which can have serious consequences for patients [Dong et al., 2023]. In addition, the time required to perform manual image analysis can be a significant bottleneck in the healthcare system, causing delays in diagnosis and treatment and potentially increasing the overall cost of care [Roy et al., 2022]. This is particularly true in cases where multiple images or the time is limited to analyze, as this can quickly become overwhelming for medical professionals [Kumar, 2021].

CAD systems have become increasingly important in medical diagnosis, given the need for reproducibility and scalability [Tsuneki, 2022]. These systems have emerged as vital tools in diagnostic scenarios where medical images are the primary source of information [Loizidou et al., 2022]. Two main approaches for analyzing medical images are classic and DL machine learning techniques. Classic approaches rely on manually designed features to capture certain aspects of images, but they may not adequately capture the intricate and diverse features of medical images or perform well on new data [Rashed and Popescu, 2022]. In contrast, DL approaches use neural networks to automatically learn features, enabling them to capture the full complexity of images and often leading to superior performance [Zhang and Dong, 2019]. Nonetheless, DL approaches require large amounts of data for effective training and may suffer limitations such as overfitting, especially in scenarios of data scarcity, and low trust due to the black-box nature [Sarker, 2021]. Additionally, the black-box nature of DL models restricts their use in MIA [Markus et al., 2021, Amann et al., 2020]. Therefore, developing image-based DL approaches for MIA requires addressing the challenges of limited samples and complex and high variability in structures of ROI while being suitable for interpretability.

1.2.1 Small Sample Size and Overfitting

Medical image segmentation is a complex task that faces several challenges. One of the primary challenges is the limited availability of datasets, especially in the case of obstetric environments, where image data acquisition is even more difficult due to the restricted situation of the mother and the discomfort of the process [Melesse et al., 2022, Willemink et al., 2020]. For instance, acquiring images of the foot soles requires the mother to maintain a relatively stable position, and the foot soles must be visible to the camera [Stevens et al., 2006, Werdehausen et al., 2007]. Additionally, the specialized equipment required for obtaining such images and the reluctance of mothers to participate in research studies make it difficult to acquire annotated data, which is crucial for developing effective segmentation techniques. These factors contribute to the challenge of training accurate and reliable segmentation models, which are essential in clinical diagnosis and treatments [Jonaitytė and Petkevičius, 2021, Yang and Yu, 2021].

Moreover, applying DL models such as Convolutional Neural Networks (CNN)s and transformer-based models for medical image segmentation in scenarios with limited data poses a significant challenge. While these models are powerful and widely utilized in computer vision tasks, their effectiveness heavily relies on extensively annotated datasets for training [Sarker, 2021, Li et al., 2021a]. Overfitting becomes a concern in scarcity scenarios where acquiring annotated data is difficult. The need for more diverse training samples and the tendency to overparameterize the models can result in excessively specialized models that struggle to generalize effectively to new data [Jain et al., 2020, Yang and Yu, 2021]. Therefore, overfitting undermines the reliability and accuracy of segmentation results, potentially compromising clinical diagnoses and treatments [Galati et al., 2022]. Therefore, it is crucial to address the issue of overfitting when dealing with limited data availability [Santos and Papa, 2022]. Doing so is essential for developing robust and effective segmentation techniques in obstetric environments and other scenarios characterized by data scarcity.

1.2.2 High Variability in the ROI in Medical Imaging

In medical image segmentation, the ROI is a critical component in identifying and analyzing the image's specific area of clinical interest [Qureshi et al., 2023]. Yet, one of the challenges of medical image analysis is the high variability of the ROI across different patients, imaging modalities, and acquisition protocols. This variability arises due to differences in anatomy, pathology, and imaging parameters, which can lead to significant variations in the ROI's shape, size, and texture [Li et al., 2021a]. This poses significant challenges for medical image segmentation, as even small inaccuracies in segmentation can significantly impact the accuracy of subsequent analysis and diagnosis [Williams et al., 2022]. Addressing these challenges requires the use of advanced segmentation algorithms and techniques, as well as careful consideration of the specific characteristics of the imaging data being analyzed.

In the context of obstetric environments, SS of feet in infrared thermal images poses significant challenges due to several factors. Firstly, thermal images inherently possess characteristics such as low contrast, blurred edges, and uneven intensity distribution, making it challenging to identify and differentiate between different objects within the image accurately [Zhang et al., 2022, Kütük and Algan, 2022]. Moreover, these characteristics can be further complicated by external factors such as varying ambient temperature, which can cause changes in the thermal patterns of the feet [Maldonado et al., 2020]. Secondly, the high variability of foot positions can lead to images with different orientations, sizes, and shapes. This variability is often present even within the same subject, resulting in a wide range of foot positions, including cases where the feet may overlap or be partially obscured [Arteaga-Marrero et al., 2021]. As a result, accurately distinguishing between the different feet can be quite challenging. Furthermore, the presence of other objects in the image, such as medical equipment, can further complicate the task of foot segmentation, making it difficult to differentiate between the feet and other objects [Bougrine et al., 2022].

1.2.3 Lack of Systematic and Quantitative Evaluations of Interpretability in SS Models

Medical analysis is an important field where SS models are widely used for decision-making. Interpretability of these models is crucial for medical practitioners as it enables them to understand and trust the reasoning behind a model's decision, thereby aiding in the diagnosis and treatment of patients [Teng et al., 2022, Xu et al., 2023, Kolyshkina and Simoff, 2021, Bennetot et al., 2022]. Nevertheless, due to the complex and black-box nature of SS models based on DL, their explainability is challenging [Linardatos et al., 2020]. Interpreting DL models is challenging for several reasons. First, DL models can have millions of parameters, making it difficult to understand how they make predictions [Singh et al., 2020]. Second, these models are often non-linear, which means that small changes in the input can result in significant changes in the output, making it hard to understand how the model makes decisions [Kulathunga et al., 2020]. Third, DL models are often trained on large amounts of data, which means the model may capture complex patterns that humans do not interpret easily [Linardatos et al., 2020].

Several interpretability methods have been developed for DL models. One set of methods, called Back-propagation interpretability methods, aims to uncover the contribution of each input feature to the final output of the model by calculating gradients concerning the input [Teng et al., 2022]. CAM-based relevance analysis methods use class activation maps to visualize regions of an image that contribute the most to a particular class prediction [Singh et al., 2020]. Perturbation-based methods involve perturbing the input data and observing the resulting changes in the output to identify the most influential features [Linardatos et al., 2020]. Lastly, Surrogate interpretability methods involve training an interpretable model, such as a decision tree, to mimic the behavior of the black-box model, providing a more understandable explanation for its decisions [Teng et al., 2022]. Each method has its strengths and weaknesses and can provide different insights into the inner workings of machine learning models. Nonetheless, most current methods for explainability on SS models rely mainly on visual inspection or qualitative analysis to evaluate the model's performance [Wang et al., 2022a, Salahuddin et al., 2022]. While this may be sufficient for evaluating one image at a time, it is not feasible to understand the model's performance or detect any biases that may be present [Zhang et al., 2021b]. Moreover, techniques based on Class Activations Maps rely on choosing a specific layer to compute the relevance, and these methods are often limited to classification models [Zhou et al., 2015a, Selvaraju et al., 2016, Chattopadhyay et al., 2017, Wang et al., 2019, Jiang et al., 2021]. As a result, objective measuress are lacking to systematically evaluate the class-relevance interpretability of the layers of SS models.

Therefore, due to the limited ability of models to generalize to new data, particularly in the context of medical image segmentation, where datasets are small and limited in availability, developing methods to improve the representation for generalization is crucial. The solution must be incorporated in shallow networks to reduce or preserve the same computational cost, and interpretability must be preserved at least in the same way as standard convolution. For these reasons, the following research question arises: **How can we develop a method for generating local and equivariant representations that can improve the generalization of deep-learning models while maintaining interpretability in Semantic Segmentation tasks for medical images?**

1.3 State of the Art

1.3.1 Enhancing Generalization Capabilities under Scenarios of Image-related data scarcity.

Nowadays, automatic segmentation techniques can be broadly classified into two categories: those based on convolution and those based on transformers.

Convolution-based segmentation models use convolutional neural networks to learn local patterns in the input data, making them ideal for processing spatially structured data such as images. Contrary to transformers with minimal inductive biases, CNN have an inductive bias of local and equivariant characterization, allowing them to learn more efficiently from scenarios of low data [Bronstein et al., 2021]. For this reason, Convolution-based models have been widely adopted in MIA applications and have shown strong performance in image segmentation tasks. For example, Fully Convolutional Networks (FCN) [Long et al., 2014] is a popular approach that uses Convolutional layers for pixel-wise classification, but produces coarse ROI and poor boundary definitions for medical images [Bi et al., 2017]. Likewise, U-Net [Ronneberger et al., 2015] consists of encoders and decoders that handle objects of varying scales but have difficulty dealing with opaque or unclear goal masks [Kumar et al., 2018]. ResUNet combines the residual connections and U-Net architecture. Their advantages include efficient memory usage and improved segmentation accuracy, but their disadvantages include longer training time and higher computational costs [Anas et al., 2017a]. U-Net++ [Zhou et al., 2018a] extends U-Net with nested skip connections for highly accurate segmentation but with increased complexity and overfitting risk. Besides, SegNet [Badrinarayanan et al., 2016] is an encoder-decoder architecture that handles objects of different scales but cannot handle fine details. Mask R-CNN [He et al., 2018] extends Faster R-CNN [Ren et al., 2016] for instance segmentation with high accuracy but requires a large amount of training data and has high computational complexity. On the other hand, PSPNet uses a pyramid pooling module for multi-scale contextual information and increased accuracy but with high computational complexity and a tendency to produce fragmented segmentation maps for small objects [Zhao et al., 2017]. Nevertheless, they still struggle to generalize when it comes to handling high variability simultaneously with a scarcity of data.

On the other hand, Transformer-based segmentation models leverage the self-attention mechanism to capture global dependencies among input features, making them well-suited for tasks requiring long-range modeling [Azad et al., 2023]. They can also handle variable-sized inputs and have shown state-of-the-art performance on a range of natural language processing and computer vision tasks [Khan et al., 2022]. For instance, since the proposal of Visual Transformers (VIT) [Dosovitskiy et al., 2020], several recent works have leveraged VIT capabilities to enhance global image representation. For example, in [Chen et al., 2021], a U-Net architecture fused with a VIT-based transformer significantly improves model performance. Nevertheless, this approach requires a pre-trained model and many iterations. Similarly, in [Cao et al., 2021], a pure U-Net-like transformer is proposed to capture long-range dependencies. Another recent work [Zhang et al., 2021a] suggests parallel branches, one based on transformers to capture long-range dependencies and the other on CNN to conserve high resolution. The authors of [Li et al., 2021b] propose a squeeze-and-expansion transformer that combines local and global information to handle diverse representations effectively. This method has unlimited practical receptive fields, even at high feature resolutions. Yet, it relies on a large dataset and has higher computational costs than conventional methods. To address the data-hungry nature of transformer-based models, the work in [Luo et al., 2022] proposes a semi-supervised cross-teaching approach between CNN and Transformers. The most recent work in this field, Meta Segment Anything [Kirillov et al., 2023], relies on an extensive natural database (around 1B images) for general segmentation. Yet, medical and natural images have noticeable differences, including color and blurriness. It is also pertinent to note that accepting ambiguity can incorporate regions that may not be part of the regions of interest. Nonetheless, as has been mentioned, models based on transformer architecture rely on a large amount of data, making these models infeasible in scenarios of limited data availability.

Further than choosing a specific architecture to improve the generalization capability and reduce the overfitting risk, transfer learning, and regularization techniques have been proposed. Transfer learning in medical image segmentation leverages pre-trained models to reduce data and computational resources needed for training and improve generalization performance. Nonetheless, using an unsuitable or non-representative pre-trained model can lead to sub-optimal performance and introduce bias, requiring careful evaluation before use [Alzubaidi et al., 2020, Guan and Liu, 2022]. On the other hand, L1 or L2 regularization adds a penalty term to the loss function to prevent overfitting. However, it may limit the model's ability to capture complex dependencies, leading to poor performance [Goodfellow et al., 2016]. Dropout regularization randomly drops nodes during training to prevent over-reliance on features but can introduce noise and inefficiency [Chen et al., 2020b]. Data augmentation can improve model performance and generalization by increasing the diversity and quantity of training data. Nevertheless, generating unrealistic or irrelevant data can lead to incorrect segmentation results. Domain-specific augmentation techniques are needed to preserve medical images' anatomical and pathological characteristics [Shorten and Khoshgoftaar, 2019, Lv et al., 2020]. Therefore, even though these techniques can help, enhancing generalization capabilities in low-data scenarios is still necessary. Furthermore, reducing the complexity reduces the capability of the models to learn complex-common dependencies.

In order to improve the ability of neural networks to generalize, researchers have investigated the incorporation of kernel methods into these networks. This is because kernel methods can generalize complex non-linear dependencies [Liu et al., 2021]. Rahimi and Recht introduced one noteworthy approach in their seminal work [Rahimi and Recht, 2009], which approximates the mapping of kernel methods using Bochner's theorem. This technique reduces the computational costs of gram matrix calculation and data storage. Several studies have since leveraged this approach to improve the efficiency and effectiveness of neural networks in various scenarios. For instance, the authors in [Morrow et al., 2017] employed a 1D-convolutional form of Random Fourier Features (RFF) to predict transcription factor binding sites from DNA sequence, achieving better inference time and performance. In [Xie et al., 2019] proposed a layer-wise composition of RFF to mimic kernel composition in situations of data scarcity. They also updated the RFF parameters through gradient descent. In [Tancik et al., 2020] employed RFF to mitigate the spectral bias of Multi-Layer Perceptrons (MLPs) and enable the capture of low frequencies. In another study

[Jimenez-Castaño et al., 2021] utilized RFF at the bottleneck of U-Net, FCN, and ResUNet architectures to improve their generalization capabilities for segmenting nerves in ultrasound images. Additionally, in [Peng et al., 2021] proposed Random Fourier Attention (RFA) to approximate attention mechanics using RFF and reduce the computational cost of transformers. Nevertheless, it should be noted that those RFF approaches do not support computation on high-dimensional data such as images without increasing model complexity.

To extend the capabilities of kernel methods to two-dimensional data, most works incorporate the properties of convolution to RFFs. The authors in [Mairal et al., 2014] proposed a Convolutional Kernel Network. Even though this approach does not use RFFs, they proposed an unsupervised approach to approximate the Gaussian kernel mapping through the linear expansion of the kernel, then the selection of random patches in each layer of the network and minimizing the sum of square errors to approximate a Gaussian kernel, which can produce instabilities during the training. In [Mohammadnia-Qaraei et al., 2018] propose a similar approach to [Mairal et al., 2014], Cosine-CKN, but instead of the linear expansions of the Gaussian kernel to approximate the mapping, they use the RFF with and without an unsupervised regime using the same approach of minimizing the sum of squared errors. [Wang et al., 2021], the authors proposed the use of RFF incorporating the optimization of the parameters through Bayes optimization. In [Wang et al., 2022b] follows the same path as the previous work but incorporates modification of the architecture o improve the performance transferring features and the network's quantization to reduce complexity. Yet, using RFF in semantic segmentation models needs to be tested. Finally, Figure 1-2 displays the most relevant approach presented to deal with a small sample size and overfitting.

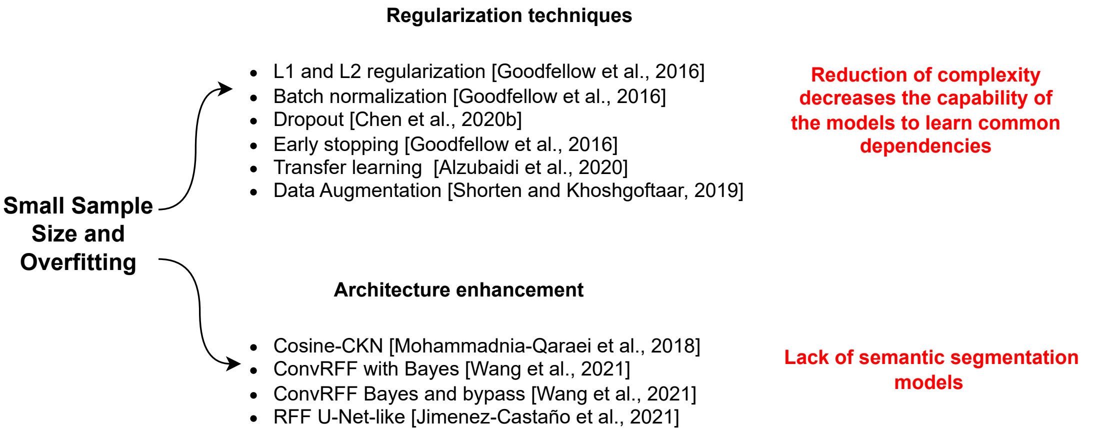

Regularization techniques
L1 and L2 regularization [Goodfellow et al., 2016]
Batch normalization [Goodfellow et al., 2016]
Dropout [Chen et al., 2020b]
Early stopping [Goodfellow et al., 2016]
Transfer learning [Alzubaidi et al., 2020]
Data Augmentation [Shorten and Khoshgoftaar, 2019]
Reduction of complexity decreases the capability of the models to learn common dependencies
Small Sample Size and Overfitting
Architecture enhancement
Cosine-CKN [Mohammadnia-Qaraei et al., 2018]
ConvRFF with Bayes [Wang et al., 2021]
ConvRFF Bayes and bypass [Wang et al., 2021]
RFF U-Net-like [Jimenez-Castaño et al., 2021]
Lack of semantic segmentation models

Figure **1-2**. Summary of state-of-the-art techniques to deal with the small sample size and overfitting issues in Semantic segmentation (SS) based on Deep Learning (DL) approaches.

1.3.2 Enhancing the Characterization of Highly Variable Object Patterns in Convolutional Neural Networks for Image Analysis

In order to address the challenges posed by the variability of ROI, several approaches have been proposed in the literature. One approach is to incorporate invariant properties into the neural networks used for image segmentation. In [Ghosh and Gupta, 2019], researchers constructed a locally scale-invariant CNN. They achieved this by incorporating invariant-to-scale characterization using scale steerable filters based on log-radial harmonics, which combine Gaussian functions and complex-valued functions with unit norms. In each convolution layer, the filters are a linear combination of basis filters, and the network learns only the complex coefficients of these filters. Nonetheless, selecting the correct hyperparameters and handling variations in scale can be challenging. Other approaches for achieving invariance include rotation-invariant techniques such as increasing data rotation and feature extraction based on rotation-invariant convolution [Hong et al., 2022]. Another method involves adopting cylindrical sliding windows in a convolutional layer to map the image into a polar coordinate system for achieving rotational invariance [Kim et al., 2020]. Additionally, Capsule Networks have shown promise as they offer advantages over traditional CNNs in handling object spatial relationships having viewpoint invariance [De Sousa Ribeiro et al., 2020]. Nonetheless, they also have some limitations, such as higher computational cost and limited empirical evidence to support their superiority over CNNs in all image recognition tasks [LaLonde et al., 2021, Pan et al., 2022].

One approach for achieving invariance is to train models to be invariant to the texture components of the image, as proposed in [Kim and Byun, 2020]. Yet, this method has limitations since it is based on synthetic data and requires a large amount of data, which may be infeasible in medical images. [Wang and Li, 2023], a different approach to improving feature invariance is presented, which involves learning a disentangled representation using generative adversarial networks (GANs) and modifying the loss function. This technique is promising since it does not require large amounts of data and has shown improved feature invariance in preliminary experiments. Data augmentation through rotation has also improved pattern classification for wafer maps [Kang, 2020]. This method is simple to implement and has shown to be effective in enhancing the performance of image recognition models. Additionally, since machine learning techniques used in computer-aided medical image analysis often suffer from the domain shift problem, which arises due to different distributions between source/reference data and target data, approaches like those proposed in [Dushatskiy et al., 2022] use multiple networks to tackle variability in scans and cross-subjects. Therefore, improving or developing techniques for the high variability of ROI in semantic segmentation is still important.

U-shape networks are popular architectures used in image segmentation tasks. They consist of an encoder and a decoder connected by skip connections. The skip connections, which combine deep, semantic, coarse-grained feature maps from the decoder sub-network with shallow, low-level, fine-grained feature maps from the encoder sub-network, have proven effective in recovering fine-grained details of the target objects even on a complex background [Long et al., 2014, Ronneberger et al., 2015]. Nevertheless, a semantic segmentation gap occurs in those networks due to the difference between the features extracted from the encoder that pass through the skip connections to the decoder [Zhou et al., 2020, Wang et al., 2021]. This difference in feature representation can affect the network's ability to capture variable regions of interest, resulting in poor segmentation performance.

In [Zhou et al., 2020] proposed UNet++, a complex network that connects encoder and decoder networks using nested, dense skip connections to reduce the semantic gap. A dense convolution block is used to enhance the semantic information in the encoder and optimize the learning task. Nevertheless, UNet++ requires additional learnable parameters, and some components may be redundant for specific tasks [Chen et al., 2020a]. In [Mubashar et al., 2022] follow the same idea of U-Net++, but adding recurrent residual blocks over vanilla convolutional blocks to provide a large field of view for layers to extract features enriched with lower-level information. Even though it carries the same limitations of U-Net++ of complexity. In [Huang et al., 2020] propose UNet 3+, which takes advantage of full-scale skip connections and deep supervision. The full-scale skip connections incorporate low-level details with high-level semantics from feature maps in different scales.

At the same time, deep supervision learns hierarchical representations from the full-scale aggregated feature maps. Even though the fusion of features from different levels arises in a mismatch of semantics between features. In the paper [Wang et al., 2021], the authors propose a new method called the Parallel Inception Network (PaI-Net) to enhance segmentation ability. This method integrates an attention parallel module and an output fusion module with the U-Net architecture, which improves the segmentation accuracy. PaI-Net enriches the multi-scale semantic information of the encoder-decoder structure by reducing the information gap between the encoder and decoder. Yet, there is an increase in complexity can lead to a data-hungry model. In [Zunair and Hamza, 2021], a Depthwise Convolutional Network called Sharp U-Net is proposed for biomedical image segmentation. It uses a simple end-to-end encoder-decoder fully convolutional network architecture for binary and multi-class segmentation. Instead of a plain skip connection, a sharpening filter layer is applied to the encoder feature map to merge it with the decoder features, producing a sharpened intermediate feature map of the same size. This improves feature fusion and reduces artifacts during early training stages. Nonetheless, the use of a sharpening filter is sensitive to noise. [Ibtehaz and Rahman, 2020a] proposed the MultiResUNet for Multimodal Biomedical Image Segmentation. This novel approach enhances the UNet architecture by incorporating a chain of convolutional layers with residual connections instead of merely concatenating feature maps between the encoder and decoder paths. By using residual connections, the model is able better to bridge the semantic gap between encoder and decoder features, resulting in more robust segmentation of images from different modalities and scales. Yet, the density connection on the skip connection paths does increase the model's complexity. In the paper [Jafari et al., 2020], the authors introduce DRU-net, an approach for medical image segmentation. DRU-net enhances the network's performance by including extra skip connections in the encoder section, similar to those found in densely connected networks. Additionally, residual connections are incorporated into the decoder section. Nevertheless, it's worth noting that this method focuses on improving the encoder-decoder architecture and does not directly address the skip connections between the encoder and decoder. Figure 1-3 depicts the summary of approaches to enhance the characterization of highly variable ROI.

1.3.3 Quantitative Interpretability of SS Models

Interpretability methods can be divided into two groups: ante-hoc and post-hoc. Ante-hoc interpretability is typically associated with directly interpretable white-box models specifically designed for interpretability [Holzinger et al., 2019]. Attention mechanisms are a popular method that mimics the human visual system to focus on regions of interest and suppress irrelevant regions. This ability allows

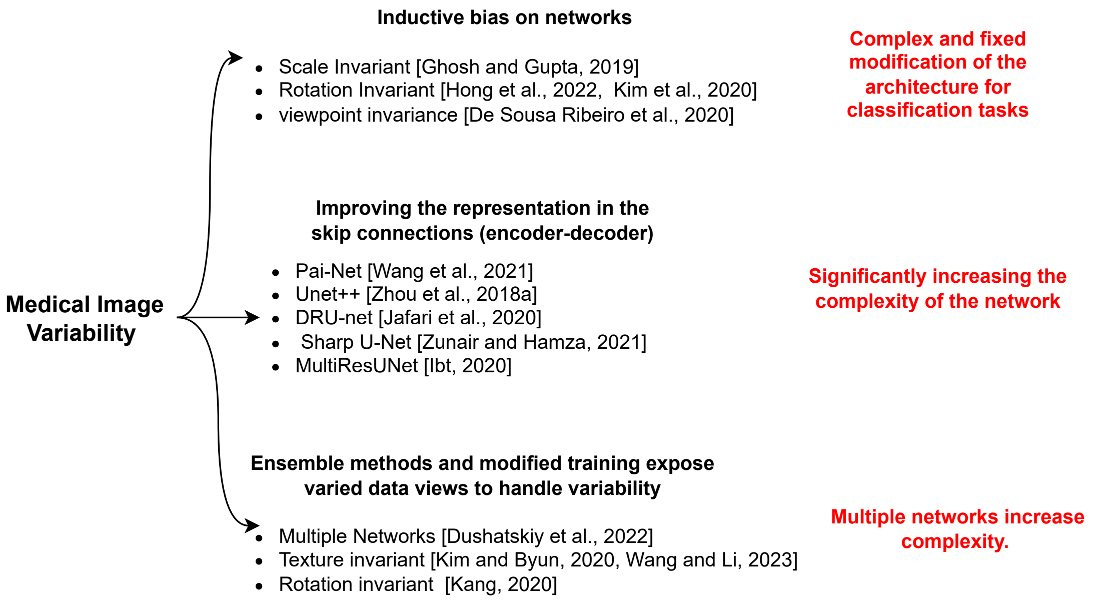

Inductive bias on networks
Scale Invariant [Ghosh and Gupta, 2019]
Rotation Invariant [Hong et al., 2022, Kim et al., 2020]
viewpoint invariance [De Sousa Ribeiro et al., 2020]
Complex and fixed modification of the architecture for classification tasks
Improving the representation in the skip connections (encoder-decoder)
Pai-Net [Wang et al., 2021]
Unet++ [Zhou et al., 2018a]
DRU-net [Jafari et al., 2020]
Sharp U-Net [Zunair and Hamza, 2021]
MultiResUNet [Ibt, 2020]
Significantly increasing the complexity of the network
Medical Image Variability
Ensemble methods and modified training expose varied data views to handle variability
Multiple Networks [Dushatskiy et al., 2022]
Texture invariant [Kim and Byun, 2020, Wang and Li, 2023]
Rotation invariant [Kang, 2020]
Multiple networks increase complexity.

Figure **1**-**3**. Summary of state-of-the-art techniques for enhancing the Characterization of highly variable object patterns in convolutional neural networks for image analysis.

these methods to be interpretable to some extent [Mohankumar et al., 2020]. Due to their simplicity, a decision tree, a linear regression, and logistic regression are also easily interpretable models [Holzinger et al., 2019]. Nevertheless, ante-hoc approach often prioritizes simple models, which may not be sufficient to handle the complexity of certain data types, especially in the case of images [Speith, 2022]. As a result, it may be necessary to explore other methods of interpretability that can accommodate the intricacies of more complex data.

On the other hand, post-hoc focused on techniques to explain a previously trained model or its prediction [Speith, 2022]. In this categories we have, perturbation-based, back-propagation and class activation. Perturbation-based interpretability methods are a type of machine learning interpretability method aimed at explaining how a model makes decisions by observing the effect on the output when the input data is perturbed [Teng et al., 2022, Linardatos et al., 2020]. By identifying the changes in output resulting from perturbations to the input, we can identify which input elements are most important for making accurate inferences. For instance, if the removal of an element causes a significant change in the output, we can conclude that it is an important element for accurate inference [Ivanovs et al., 2021]. One popular method is LIME, or Local Interpretable Model-Agnostic Explanations [Ribeiro et al., 2016], which approximates a model's behavior near a data point to generate an interpretable explanation that is model-agnostic. Yet, its approximations may not always be accurate, and it can be computationally expensive [Doumard et al., 2022]. Another example is SHAP, SHapley Additive exPlanations [Lundberg and Lee, 2017], which attributes a value to each feature in the input data to represent its contribution to the model's prediction based on the Shapley value. Nonetheless, it can also be computationally expensive, particularly for large datasets and complex models [Doumard et al., 2022]. Occlusion is another perturbation-based interpretability method that measures important features by occluding parts of the input, typically by covering different parts of the input with a gray square to measures the change in outcome. If the output of the model changes dramatically, it means that this part has a significant impact on the prediction [Samek et al., 2021]. Nonetheless, one of the main issues is that perturbation itself may introduce artifacts since perturbed images may be out-of-distribution [Brocki and Chung, 2022].

Back-propagation interpretability methods are popular techniques used to understand the workings of neural networks. These methods involve back-propagating signals from the output or a specific layer of interest to the input layer of the model [Teng et al., 2022]. Activation maximization, for example, involves optimizing an input image to maximize the activation of a particular neuron or layer in the network. This can help identify which features the network is focusing on and can be useful for debugging or fine-tuning a model. Nevertheless, visualizations generated using this method can be complex and difficult to understand [Stergiou, 2021]. Another backpropagation method, Layer-Wise Relevance Propagation (LRP), predicts classification results by calculating pixel contributions. Starting from the output layer and moving in the opposite direction, LRP redistributes the relevant score until it reaches the input layer while following the global conservation property. Nonetheless, the heatmaps generated by LRP can be noisy and non-discriminative, making them similar in different predictions [Jung et al., 2021].

The Class Activation Map (CAM) is a widely used visual technique that provides interpretability for image classification tasks. This technique generates class activation maps, which enable researchers to identify the most discriminative regions in an image [Zhou et al., 2015a]. The idea behind CAMs is to create highlighted regions for a specific class using a linear combination of feature maps from a particular layer of a convolutional neural network. Then, the difference between all the approaches of CAMs methods like Grad-CAM [Selvaraju et al., 2016], Grad-CAM++ [Chattopadhyay et al., 2017], Score-CAM [Wang et al., 2019], LayerCAM [Jiang et al., 2021], Shap-CAM[Zheng et al., 2022] etc. Additionally, CAMs approaches are widely applicable to multiple kinds of architectures, simple to implement and interpret, efficient, and produce intuitive visualizations [Teng et al., 2022]. These benefits make CAMs a powerful and versatile tool for researchers and practitioners looking to better understand the features of CNN-based models.

As mentioned, all these different approaches to achieving interpretability have mainly been developed for classification models. Furthermore, visual inspection is often the primary method used to evaluate and capture the interpretability of these models[Linardatos et al., 2020]. [Vinogradova et al., 2020] proposed a technique to use those CAMs methods under SS models through an average of the pixels of the interest class. As mentioned, to analyze these CAMs, a visual inspection is performed, which does not allow one to observe the general behavior of the model and scale to multiple images to get a conclusion. From the quantitative performance of the CAMs, multiple measuress were developed to compare CAM methods like average drop, average increase, and Win [Selvaraju et al., 2016, Chattopadhyay et al., 2017, Wang et al., 2019, Jiang et al., 2021] . Yet, these methods are difficult to interpret in SS models, where we are more interested in spatial information in the input and output space.

[Ventura et al., 2023] introduced EBAnO, a perturbation-based explanation framework for analyzing the decision-making process of deep convolutional neural networks (DCNNs) in image classification tasks. The proposed framework includes indexes for quantifying the influence and precision of input features with respect to a given prediction, enabling both quantitative and qualitative analysis. Nonetheless, the authors noted that this technique cannot be applied to segmentation models due to their requirements, which are unable to capture the relative importance of classes within the region of interest due the coarse of the relevance map. The paper [Dardouillet et al., 2022] introduces a method for visually interpreting SS models using SHAP values. While this approach offers promising insights, it comes with a high computational cost. Moreover, the explanations generated using this method can be somewhat coarse due to the use of superpixels. In the article [Schorr et al., 2021], the authors introduced a comprehensive toolbox for semantic segmentation and classification models, which includes multiple state-of-the-art interpretability methods such as CAM-Based methods. They have also modified these methods to work with semantic segmentation models. Yet, it is important to note that the interpretability of the models still relies on visual inspection and has the restriction of local explainability, which is related to the specific sample and the layers analyzed. In their recent work [Sacha et al., 2023], the authors propose a novel method for enhancing the interpretability of SS, which relies on prototypes. To increase the variability of prototypes, they introduce a loss function. Still, the method has some limitations. It may not be suitable for low database sizes as it relies heavily on prototypes extracted from the datasets. Furthermore, the authors assume that the structures of interest are modular, which is not always the case in practice. The SAU-Net proposed in [Sun et al., 2020] aims to improve the interpretability and robustness of models. To achieve this, the authors modified the U-Net architecture by adding a parallel shape stream based on attention blocks to improve the interpretability of the model's inner workings. Additionally, both spatial and channel attention mechanisms were used in the decoder to provide insight into the model's learning capabilities at each resolution of the U-Net. By extracting the learned shape and spatial attention maps, the highly activated regions of each decoder block can be interpreted. Nevertheless, it should be noted that this approach is specific to the SAU-Net model, and it is not applicable to other models.

In summary, the interpretability methods used in semantic segmentation have challenges and limitations. Perturbation-based techniques can introduce artifacts from out-of-distribution images. Backpropagation-based methods rely on the assumption that increased activation indicates a specific feature, making interpretation difficult. Class activation methods offer local-layer explanations but lack quantitative explanations. These shortcomings highlight the necessity for additional research to improve interpretability in semantic segmentation and provide more accurate and reliable explanations for model decisions. Figure 1-4 summarizes the main approach of qualitative and quantitative methods for image-based interpretability.

In this research context, working with limited data, shallow encode-decoder semantic segmentation models serve as a good alternative to address the challenges of this scenario [Taghanaki et al., 2021]. Given its association with kernel methods, the RFF technique emerges as a valuable option to enhance generalization and mitigate overfitting [Rahimi and Recht, 2009]. However, there is a pressing need to investigate the application of RFF approximation in image-based deep learning for semantic segmentation models. Regarding interpretability, CAMs have demonstrated promising results in identifying significant image regions for specific classes. Nevertheless, extending CAM to tackle the issue of relevance across layers and classes in SS remains an unexplored challenge. These considerations form the basis of our objectives, which aim to exploit shallow decoder-encoder models, explore RFF techniques for spatial data in segmentation, and develop advanced interpretability methods for SS that encompass relevance across layers and classes.

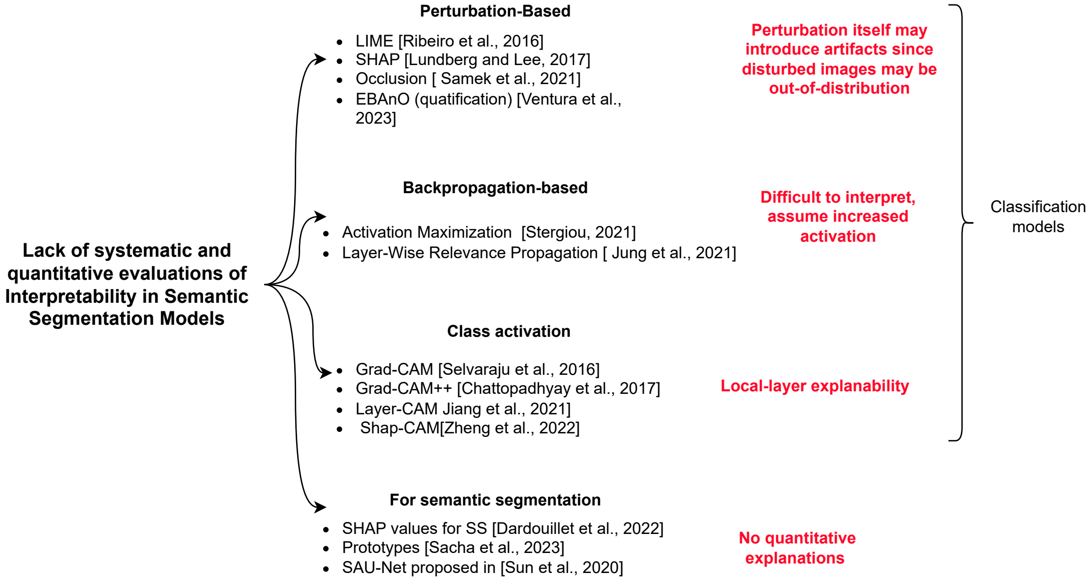

Lack of systematic and quantitative evaluations of Interpretability in Semantic Segmentation Models
Perturbation-Based
LIME [Ribeiro et al., 2016]
SHAP [Lundberg and Lee, 2017]
Occlusion [Samek et al., 2021]
EBAnO (quatification) [Ventura et al., 2023]
Perturbation itself may introduce artifacts since disturbed images may be out-of-distribution
Backpropagation-based
Activation Maximization [Stergiou, 2021]
Layer-Wise Relevance Propagation [Jung et al., 2021]
Difficult to interpret, assume increased activation
Classification models
Class activation
Grad-CAM [Selvaraju et al., 2016]
Grad-CAM++ [Chattopadhyay et al., 2017]
Layer-CAM Jiang et al., 2021]
Shap-CAM[Zheng et al., 2022]
Local-layer explanability
For semantic segmentation
SHAP values for SS [Dardouillet et al., 2022]
Prototypes [Sacha et al., 2023]
SAU-Net proposed in [Sun et al., 2020]
No quantitative explanations

Figure **1-4**. Summary of state-of-the-art Quantitative Interpretability of SS models.

1.4 Aims

1.4.1 General Aim

Develop a deep-learning-based semantic image-segmentation methodology incorporating a **convolutional** layer based on **Random Fourier Features** and comprehensive **interpretability measures** to encode high variable and relevant patterns related to the region of interest and improve generalization performance under conditions of **scarce data**.

1.4.2 Specific Aims

- To design an extension of RFF for **spatial data** with optimization through gradient descent for generalization under scarcity data through local and equivariant characterization.

- To develop a semantic segmentation approach based on type encoder-decoder architectures that incorporate Random Fourier Features for enhanced skip representation for improved capture of small and variable objects in semantic segmentation tasks.

- To develop a post-hoc interpretability approach based on measures for quantitative assessment of relevance maps taking into account the spatial information of semantic segmentation tasks for global and layer-wise relevance analysis.

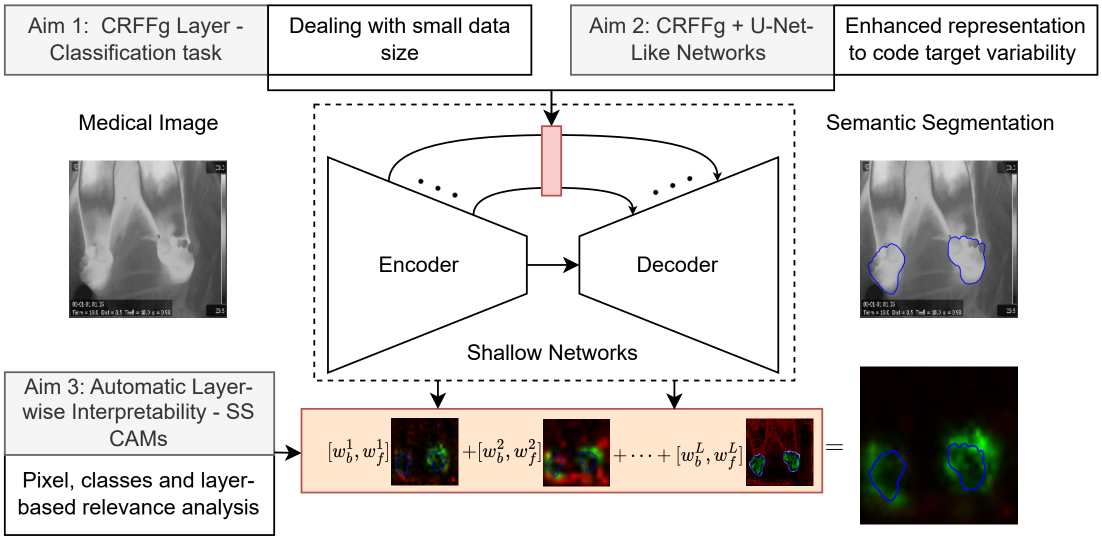

Aim 1: CRFFg Layer - Classification task
Dealing with small data size
Aim 2: CRFFg + U-Net-Like Networks
Enhanced representation to code target variability
Medical Image
Semantic Segmentation
Encoder
Decoder
Shallow Networks
Aim 3: Automatic Layer-wise Interpretability - SS CAMs
Pixel, classes and layer-based relevance analysis
$[w_b^1, w_f^1]$
$+[w_b^2, w_f^2]$
$+\cdots+[w_b^L, w_f^L]$
=

Figure **1-5**. Thesis Contributions. In this study, we present a novel extension of RFF specifically designed for analyzing spatial data in a data-driven manner. Additionally, we leverage this extension to improve the representation of shallow encoder-decoder models at the skip connections for semantic segmentation. Furthermore, we introduce a new quantitative measure aimed at enhancing the interpretability of semantic segmentation models.

1.5 Outline and Contributions

In this section, we provide a brief overview of the key contributions presented in this thesis, summarized in Figure 1-5.

1.5.1 Random Fourier Features in Convolutional Form

Medical image segmentation is challenging due to the limited availability of datasets, especially in obstetric environments. Image acquisition is difficult due to the restricted situation of the mother and the discomfort involved. For example, obtaining images of ROI requires a stable position and visibility to the camera. Specialized equipment and the reluctance of mothers to participate in research studies further hinder data acquisition for developing effective segmentation techniques. This challenges training accurate and reliable segmentation models crucial for clinical diagnosis and treatments.

With that in mind, in Chapter 2, we present a novel approach to improve Random Fourier Features for spatial data by optimizing them through gradient descent named Convolutional RFF gradient (CRFFg). This extension enables us to harness the advantageous generalization properties of kernel methods while also obtaining local and equivariant characterization, thereby enhancing data efficiency for spatial data derived from convolutions.

1.5.2 Enhanced Shallow Encoder-decoder Models for Semantic Segmentation

Medical image segmentation relies on accurately identifying and analyzing the region of interest (ROI). Still, it faces challenges due to variability in shape, size, and texture across patients and imaging protocols. Segmenting feet in thermal images in obstetric environments is particularly challenging due to low contrast, blurred edges, uneven intensity distribution, varying ambient temperature, and diverse foot positions. Additionally, the presence of other objects further complicates the segmentation task. Various approaches have been proposed to address ROI variability, including incorporating invariant properties into neural networks, such as scale and rotation invariance. Capsule Networks and disentangled representations show promise but have limitations. Data augmentation through rotation and techniques addressing domain shift can also improve segmentation. Overall, improving techniques for handling ROI variability in semantic segmentation remains a study of interest.

In Chapter 3, we present a novel enhancement to decoder-encoder architectures for semantic segmentation. Our approach focuses on improving the representation at the skip connection by incorporating a CRFFg layer. This enhancement aims to improve the capture of low-level features from the encoder and the fusion of these features into the decoder.

1.5.3 Quantitative measuress for Relevance Maps in Semantic Segmentation

In the medical field, interpretability plays a crucial role in decision-making for SS models. Yet, the complex and opaque nature of DL models in this domain presents challenges for achieving interpretability. With millions of parameters, non-linear behavior, and the ability to capture intricate patterns from extensive data, these models are inherently difficult to understand. Several interpretability methods have been developed, including Back-propagation, CAM-based, Perturbation-based, and Surrogate methods. Each method offers unique strengths and weaknesses, providing different insights into the models. Yet, current methods often rely on visual inspection or qualitative analysis, which may not be sufficient for evaluating overall model performance and detecting biases. Additionally, Class Activation Maps-based methods have limitations and are primarily applicable to classification models. Consequently, there is a lack of objective measuress to systematically evaluate the interpretability of layers in SS models.

Therefore, in Chapter 4, we propose novel measuress to enhance the interpretability of semantic segmentation models. These measuress assess three crucial aspects of the models. Firstly, we introduce CAM-based Cumulative Relevance, which identifies the location of relevance in specific regions of interest. Secondly, we introduce Mask-based Cumulative Relevance, which quantifies sensibility across multiple regions of interest. Lastly, we propose CAM-Dice to assess the homogeneity of relevance in interest regions. By incorporating these measures, we aim to provide comprehensive and global quantifications for interpretability in semantic segmentation models.

The implementation of CRFFg and the proposed measures are in a GitHub repository [1] [2].

[1] https://github.com/aguirrejuan/ConvRFF

[2] https://github.com/aguirrejuan/Foot-segmentation-CRFFg

1.6 Semantic segmentation (SS) Databases

In order to test the proposed work, we used the following databases.

1.6.1 Fashion Mnist

The dataset comprises 70,000 grayscale images of fashion products, each sized 28x28 pixels. These images are sourced from Zalando's collection of article images, encompassing 10 distinct categories. Each category is represented by 7,000 images, ensuring a balanced distribution. The training set is composed of 60,000 images, while the remaining 10,000 images form the test set [Xiao et al., 2017a]. Figure 1-6 shows examples of images in this dataset.

1.6.2 Infrared Thermal Images - ThermalFeet

ThermalFeet: This thermography image database was collected during epidural anesthesia administration in labor. Due to the complexities involved in labor, the sample size is relatively small, and capturing images of both feet in the same position was not always feasible. The clinicians at SES Hospital Universitario de Caldas devised a timeline for data acquisition: the first thermal picture was taken at catheter placement, followed by another picture taken one minute later, and subsequently every five minutes, resulting in a total of six images per patient.

The initial set of images consisted of 196 samples from 22 women. These images were captured using a FLIR A320 infrared camera with a resolution of 640x480 and a spectral range of 7.5 to 13 μm. The second set of images (128 in total) was captured using a FLIR E95 thermal camera, which offered improved sensitivity and flexibility and improved image quality. Both sets of images were annotated for semantic segmentation by three researchers using the CVAT Computer Vision

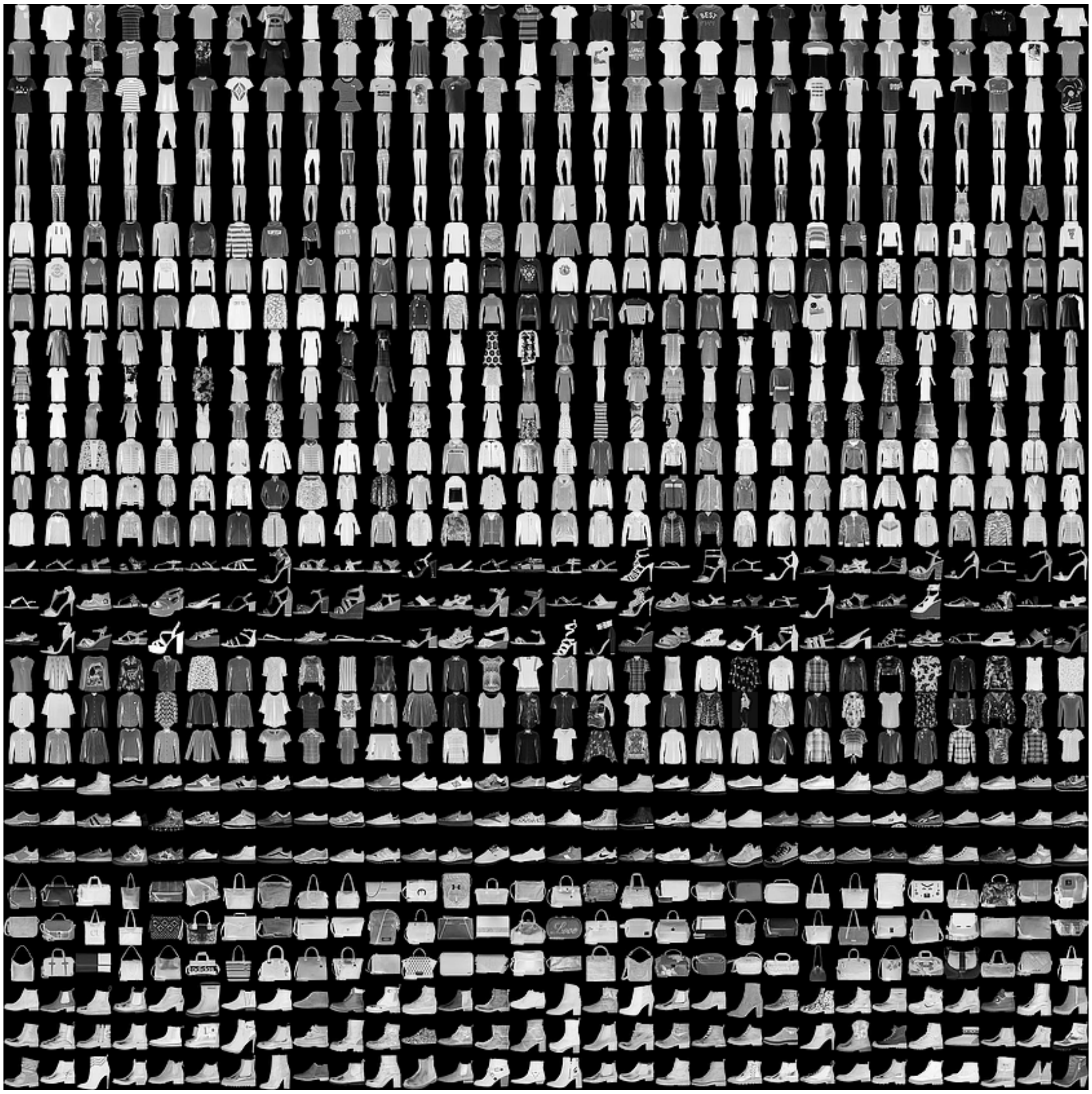


Figure **1-6**. Fashion Mnist dataset [Xiao et al., 2017a].

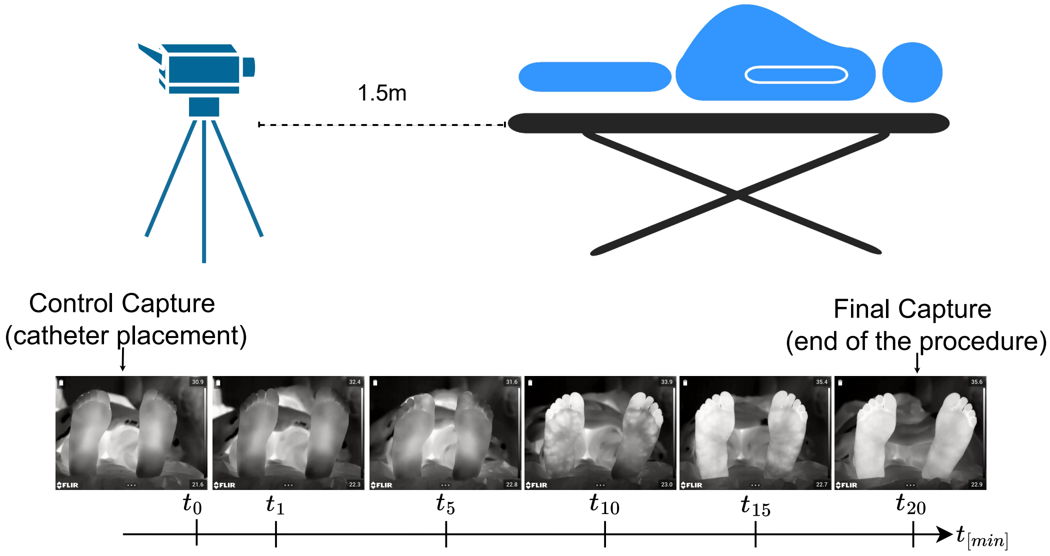


Figure **1-7**. Infrared Thermal database acquisition.

Annotation Tool[3]. From the combined first and second sets, a total of 166 high-quality images were selected [Mejia-Zuluaga et al., 2022].

The dataset is available at [4].

1.6.3 Natural Images - Oxford IIIt Pet

Oxford-IIIT Pet [Parkhi et al., 2012]: This dataset features 37 different pet categories, each with approximately 200 images, resulting in 3,680 for training and 3,669 for testing. The images have significant scale, pose, and lighting variations, and each is accompanied by a labeled breed annotation [Parkhi et al., 2012]. The dataset is available at [5]

[3] https://cvat.org/
[4] https://gcpds-image-segmentation.readthedocs.io/en/latest/notebooks/02-datasets.html
[5] https://www.tensorflow.org/datasets/catalog/oxford_iiit_pet

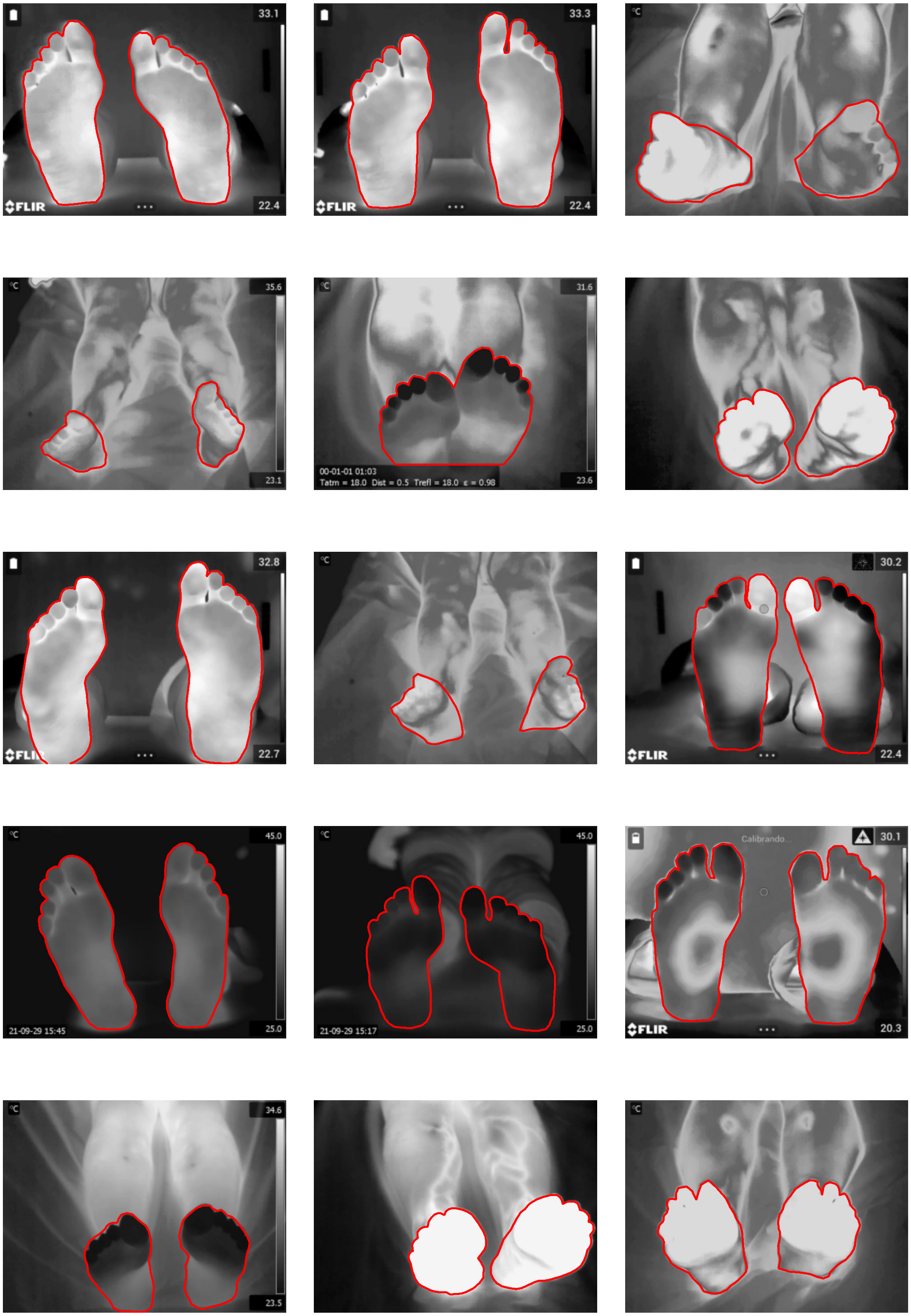


Figure **1-8**. Infrared Thermal Example Images.

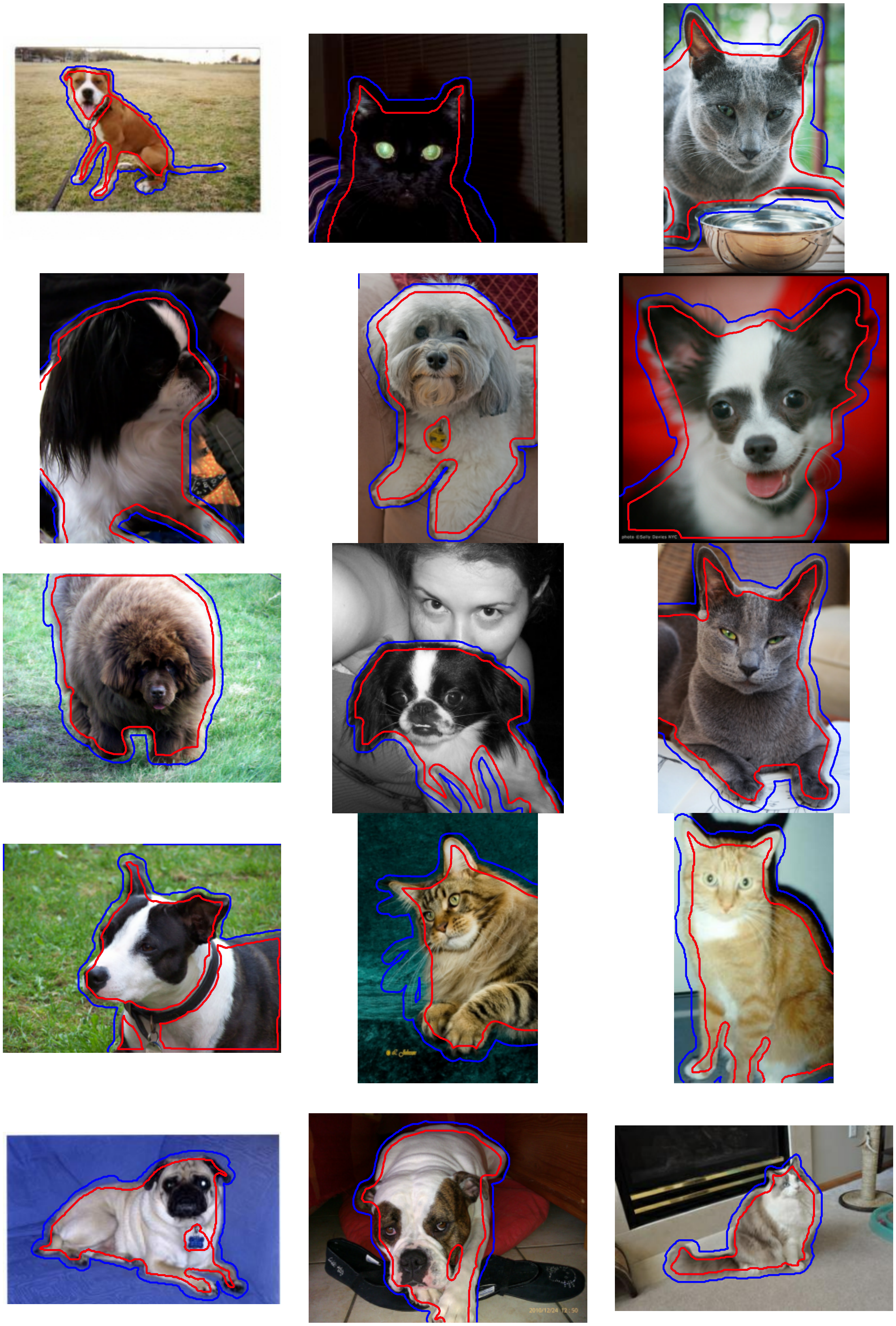


Figure **1-9**. Oxford Pet Example Images.

1.7 Thesis Structure

The next parts of this thesis are organized as follows. In Chapter 2, we introduce an extension of Random Fourier Features to handle spatial data, acquire locality and equivariant to work high dimensional data, such as images, name *CRFFg*. Chapter 3 implements the CRFFg on high-resolution tasks such as SS. Finally, Chapter 4 introduces a new set of interpretability measures for Class Activation Maps on SS models.

CHAPTER

TWO

RANDOM FOURIER FEATURES EXTENDED TO CONVOLUTIONAL FORM

This chapter extends the Random Fourier Features (RFF) to work with spatial data. We begin by describing the RFF method. Then, we extend the one-dimensional operation to a two-dimensional operation through a convolution form. Finally, we present the experiments' results to validate our approach.

2.1 Theoretical Background RFF

Kernel methods are widely used in machine learning for no-linear dependence problems. However, as the dataset's size grows, kernel methods' computational complexity also increases, making applying them to large datasets impractical. RFF were introduced to approximate kernel methods using random projections to address this issue [Rahimi and Recht, 2009]. RFF can effectively scale kernel machines and enable them to handle large datasets by mapping input data to a

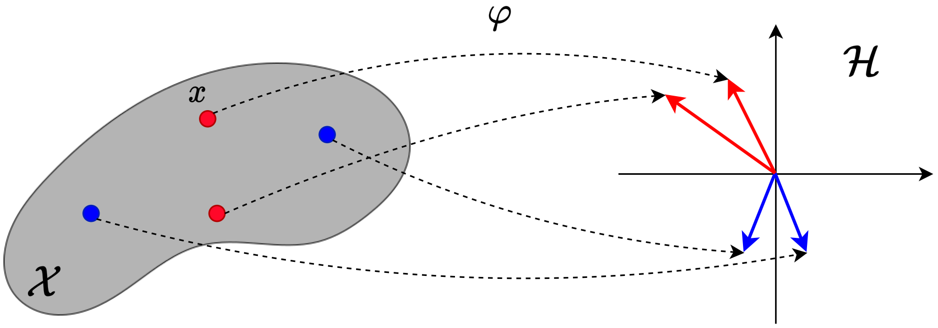


Figure **2-1**. Hilbert space benefits representation. Unstructured space to richer structure space.

high-dimensional space through random Fourier transforms. This technique has gained popularity recently due to its ability to provide fast and accurate approximations of kernel methods, making it a practical solution for many real-world applications [Liu et al., 2021]. This section will discuss the theoretical background of RFF in more detail and explore its benefits and limitations in scaling kernel machines.

2.1.1 Kernel Machines

A learning problem with data and targets $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, where $\mathbf{x}_n \in \mathcal{X}$ and $y_n \in \mathcal{Y}$, can be approached with a simple linear model defined as Equation 2-1

$$y = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} \tag{2-1}$$

However, this model is limited in that it can only accurately classify linearly separable data. To overcome this limitation, we can use a mapping $\varphi : \mathcal{X} \to \mathcal{H}$ to transform the vector $\mathbf{x}$ into a high-dimensional Hilbert space, where the data can now be linearly separable as is shown in the Figure 2-2. Moreover, Hilbert spaces

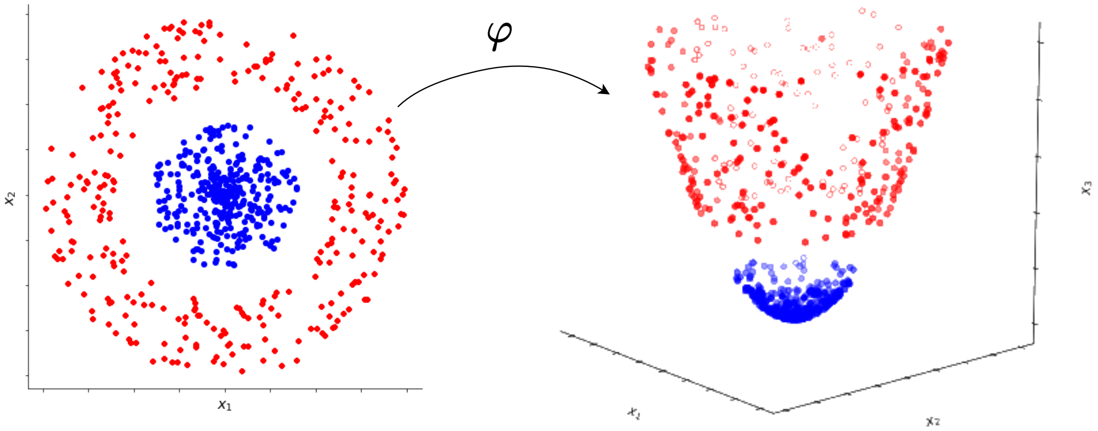


Figure **2-2**. Mapping function to a high dimensional space where the data can be linearly separable.

offer rich structure, efficient approximation of complex data, generalization, and flexibility. These spaces are useful for various mathematical and computational applications beyond the linear separation of data [Mairal and Vert, 2018]. Nonetheless, knowing such a map beforehand is often difficult, and the computational cost in high-dimensional space can also be a limitation. Kernel machines overcome these limitations by using the kernel trick, which can induce a high dimensional mapping, and allows the computation of dot products in the high-dimensional space in the input space, which is affordable [Kutateladze, 2022].

The definition of a kernel is presented in Equation 2-2.

Kernel Method

Given samples **x**,**x**′ from a set $\mathcal{X}$ a kernel function $k(\mathbf{x}, \mathbf{x}')$ maps each pair of input points to a scalar value:

$$k(\mathbf{x}, \mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle_{\mathcal{H}} \quad \text{where } \varphi : \mathcal{X} \to \mathcal{H} \tag{2-2}$$

Where $\mathcal{H}$ denotes the Hilbert space, which can be infinite-dimensional vector

space, and φ is the mapping function related with the kernel .
The kernel function satisfies the Mercer's condition

$$\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \mathbf{x}'_j)c_i c_j \geq 0 \tag{2-3}$$

Kernel functions satisfy the Mercer's condition described in Equation 2-3, which requires it to be symmetric and positive definite. This condition ensures that the kernel matrix, $\mathbf{K} \in \mathbb{R}^{n\times n}$, obtained by applying the kernel function to all pairs of input points, is positive semi-definite.

Then the function from Equation 2-1 can be written then as Equation 2-4 through the property of reproducibility of the kernel [Mairal and Vert, 2018]. However, this is difficult to scale to large datasets due to the need to calculate the Gram matrix and store the data [Bengio and Lecun, 2017].

$$f(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n k(\mathbf{x}, \mathbf{x}_n) = \langle \boldsymbol{\omega}, \varphi(\mathbf{x}) \rangle_{\mathcal{H}} \tag{2-4}$$

2.1.2 RFF to Scale Kernel Machines

To overcome the limitations of kernel machines in large scales data scenarios, RFF defines a finite-dimensional explicit map that approximates shift-invariant kernels [Rahimi and Recht, 2009]. As shown in Equation 2-5, we can have a z mapping from input space to a finite dimension space.

$$k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle_{\mathcal{H}} \approx z(\mathbf{x})^{\top} z(\mathbf{x}') \quad z : \mathcal{X} \to \mathbb{R}^{R} \tag{2-5}$$

Where R is finite.

Then we can rewrite the function f of Equation 2-4 as Equation 2-6:

$$\begin{aligned} f(\mathbf{x}) &= \sum_{n=1}^{N} \alpha_n k(\mathbf{x}_n, \mathbf{x}) \\ &= \sum_{n=1}^{N} \alpha_n \langle \varphi(\mathbf{x}_n), \varphi(\mathbf{x}) \rangle_{\mathcal{H}} \\ &\approx \sum_{n=1}^{N} \alpha_n \mathbf{z}(\mathbf{x}_n)^\top \mathbf{z}(\mathbf{x}) \\ &= \boldsymbol{\beta}^\top \mathbf{z}(\mathbf{x}) \end{aligned} \tag{2-6}$$

Which is a simpler linear model in the resulting space of the mapping $\mathbf{z}$. This mapping $\mathbf{z}$ can be defined from the Bochner's theorem [Rudin, 1976] that states:

Bochner's theorem

A continuous function of the form $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}')$ is a positive definite if only if $k(\delta)$ is the Fourier transform of a non-negative measure.

$$k(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^p} p(\boldsymbol{\omega}) \exp(j\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')) d\boldsymbol{\omega} \tag{2-7}$$

Where $\boldsymbol{\omega} \sim p(\boldsymbol{\omega})$

$p(\boldsymbol{w})$ is a probability density function of $\boldsymbol{w}$ and which defines the type of kernel. For example, the Gaussian kernel in the Equation 2-8, which is preferred because of its universal approximating property and mathematical tractability [Álvarez-Meza et al., 2014], can be obtained setting $p(\boldsymbol{w})$ equal to a Gaussian distribution.

$$k(\mathbf{x}, \mathbf{x}') = \exp(\frac{-||\mathbf{x} - \mathbf{x}'||_2^2}{2\sigma^2}) \tag{2-8}$$

With the resulting Equation 2-7, we can go further, expressing these results in terms of expected value as shown in Equation 2-9, which allows us to end up with the sample mean in Equation 2-10, and finally, expand it as a dot product in Equation 2-11.

$$
\begin{aligned}
k(\mathbf{x}-\mathbf{x}') = \mathbb{E}_{\omega}\{\, \exp(i\boldsymbol{\omega}^{\top}(\mathbf{x}-\mathbf{x}'))\} & \qquad (2\text{-}9) \\
\approx \frac{1}{R}\sum_{r=1}^{R} \exp(i\boldsymbol{\omega}_r^{\top}(\mathbf{x}-\mathbf{x}')) & \qquad (2\text{-}10) \\
= \begin{bmatrix} \frac{1}{\sqrt{R}}\exp(i\boldsymbol{\omega}_1^{\top}\mathbf{x}) \\ \frac{1}{\sqrt{R}}\exp(i\boldsymbol{\omega}_2^{\top}\mathbf{x}) \\ \vdots \\ \frac{1}{\sqrt{R}}\exp(i\boldsymbol{\omega}_R^{\top}\mathbf{x}) \end{bmatrix}^{\top} \begin{bmatrix} \frac{1}{\sqrt{R}}\exp(-i\boldsymbol{\omega}_1^{\top}\mathbf{x}') \\ \frac{1}{\sqrt{R}}\exp(-i\boldsymbol{\omega}_2^{\top}\mathbf{x}') \\ \vdots \\ \frac{1}{\sqrt{R}}\exp(-i\boldsymbol{\omega}_R^{\top}\mathbf{x}') \end{bmatrix} & \qquad (2\text{-}11)
\end{aligned}
$$

We know that the kernel is a real function, then the imaginary part is discarded [Gundersen, 2019], as shown in Equation 2-12

$$
\begin{aligned}
\exp(i\boldsymbol{\omega}^{\top}(\mathbf{x}-\mathbf{x}')) &= \cos(\boldsymbol{\omega}^{\top}(\mathbf{x}-\mathbf{x}')) - \cancel{i\sin(\boldsymbol{\omega}^{\top}(\mathbf{x}-\mathbf{x}'))} \\
&= \cos(\boldsymbol{\omega}^{\top}(\mathbf{x}-\mathbf{x}')) \qquad (2\text{-}12)
\end{aligned}
$$

Then our z mapping can be set as Equation 2-13 or Equation 2-14. For simplicity, in this work, we use the mapping using the Equation 2-13.

$$
z_{\boldsymbol{\omega}}(\mathbf{x}) = \sqrt{2}\cos(\boldsymbol{\omega}^{\top}\mathbf{x} + b) \qquad (2\text{-}13)
$$

$$
z_{\boldsymbol{\omega}}(\mathbf{x}) = \begin{bmatrix} \cos(\boldsymbol{\omega}^{\top}\mathbf{x}) \\ \sin(\boldsymbol{\omega}^{\top}\mathbf{x}) \end{bmatrix} \qquad (2\text{-}14)
$$

where $\boldsymbol{\omega} \sim p(\boldsymbol{\omega})$ and $b \sim$ Uniform$(0, 2\pi)$.

We can briefly prove that this z mapping holds the result of Equation 2-12

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\omega}}[z_{\boldsymbol{\omega}}(\mathbf{x}) z_{\boldsymbol{\omega}}(\mathbf{x}')] &= \mathbb{E}_{\boldsymbol{\omega}}[\sqrt{2}\cos(\boldsymbol{\omega}^\top \mathbf{x} + b)\sqrt{2}\cos(\boldsymbol{\omega}^\top \mathbf{x}' + b)] \\
&= \mathbb{E}_{\boldsymbol{\omega}}[\cos(\boldsymbol{\omega}^\top (\mathbf{x} + \mathbf{x}') + 2b)] + \mathbb{E}_{\boldsymbol{\omega}}[\cos(\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}'))] \\
&= \mathbb{E}_{\boldsymbol{\omega}}[\cos(\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}'))].
\end{aligned}$$

Then, we can rewrite the Equation 2-10 as Equtation 2-15

$$\begin{aligned}
\mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{x}') &= \frac{1}{R}\sum_{r=1}^{R} z_{\boldsymbol{\omega}_r}(\mathbf{x}) z_{\boldsymbol{\omega}_r}(\mathbf{x}') \qquad (2\text{-}15) \\
&= \frac{1}{R}\sum_{r=1}^{R} 2\cos(\boldsymbol{\omega}_r^\top \mathbf{x} + b_r)\cos(\boldsymbol{\omega}_r^\top \mathbf{x}' + b_r) \\
&= \frac{1}{R}\sum_{r=1}^{R} \cos(\boldsymbol{\omega}_r^\top (\mathbf{x} - \mathbf{x}')) \approx \mathbb{E}_{\boldsymbol{\omega}}[\cos(\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}'))] = k(\mathbf{x}, \mathbf{x}')
\end{aligned}$$

This mapping can be implemented as a dense layer, Equation 2-16 with activation function $\cos(\cdot)$ and proper initialization as in Equation 2-13 and R units.

$$\mathbf{z}(\mathbf{x}) = \begin{bmatrix} \frac{1}{\sqrt{R}} z_{\boldsymbol{\omega}_1}(\mathbf{x}) \\ \frac{1}{\sqrt{R}} z_{\boldsymbol{\omega}_2}(\mathbf{x}) \\ \vdots \\ \frac{1}{\sqrt{R}} z_{\boldsymbol{\omega}_R}(\mathbf{x}) \end{bmatrix} \qquad (2\text{-}16)$$

As mentioned before, the type of mapping depends on the distribution chosen for $\boldsymbol{w}$. Gaussian distribution for Gaussian Kernel is usually used for the properties capturing no linear dependencies and its generalization capability. Furthermore, the parameter σ is usually trained through gradient descent, making the trick of writing the mapping $z_{\boldsymbol{\omega}}(\mathbf{x})$ adding the σ divide the $\boldsymbol{\omega}$ as Equation 2-17.

$$z_{\boldsymbol{\omega}}(\mathbf{x}) = \sqrt{2}\cos(\frac{\boldsymbol{\omega}}{\sigma}^{\top}\mathbf{x} + b) \tag{2-17}$$

While RFF is a useful approximation technique, it may not be suitable for high-dimensional data such as images.

2.2 RFF Extended to Convolutional Form

The principal characteristics of convolutional operations are its properties of translation equivariance and notions of locality. Translation equivariance refers to the property that a convolutional layer produces the same output regardless of the position of the input. In other words, if the input is translated, the output of the layer is also translated in the same way. This property is important for image recognition tasks since objects can appear at different positions in an image. By being translation equivariant, convolutional layers can recognize objects regardless of their position in the image [Bronstein et al., 2021]. Notions of locality refer to the property that a convolutional layer only looks at a small part of the input at a time, known as the receptive field. This is achieved through the use of small filters that slide over the input to compute local features. This property is important for two reasons. Firstly, it reduces the number of parameters in the model, making it more efficient to train. Secondly, it allows the model to capture local patterns in the input, which are often important for object recognition [Goodfellow et al., 2016]. Those properties are depicted in Figure 2-3. Then, those inductive bias of local and equivariant allow the CNN models to learn more efficiently from scenarios of low data.

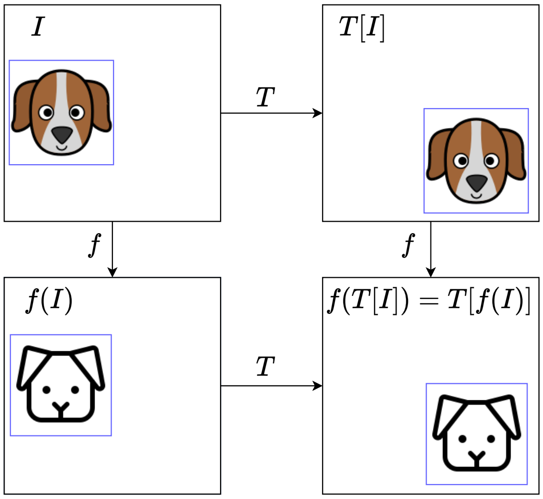


Figure **2-3**. Translation Equivariance and locally Properties of Convolutional Layers.

To add these property to RFF, we perform z mapping in localities of the grid input space. Suppose we have the feature map $\mathbf{F}_l \in \mathbb{R}^{H_l \times \tilde{W}_l \times Q_l}$ we constructed a mapping such that $z : \mathbb{R}^{H_{l-1} \times \tilde{W}_{l-1} \times D_{l-1}} \to \mathbb{R}^{H_l \times \tilde{W}_l \times Q_l}$ as shown in Equation 2-18. As mentioned before, we are interested in the properties of the Gaussian kernel, then follow a similar approach as the Equation 2-17 with $\sigma \in \mathbb{R}^+$ Scale. This maps is exemplified in Figure **2-4**

$$\mathbf{F}_l = z(\mathbf{F}_{l-1}) = cos(\frac{\mathbf{W}_l}{\sigma} \otimes \mathbf{F}_{l-1} + \mathbf{b}_l) \tag{2-18}$$

It is worth noting that the parameters $\{\mathbf{W}, \mathbf{b}, \sigma\}$ of this layer are easily updated during the training step through gradient descent under a back-propagation-based optimization. We call this layer CRFFg and the implementation can be found in [1]. Then, We have extended the use of Random Fourier features to spatial data by optimizing it through gradient descent. This approach allows us to acquire the desirable properties of generalization from kernel methods and enables us to obtain local and equivariant characterization for spatial data using convolutions.

[1] https://github.com/aguirrejuan/ConvRFF/blob/master/convRFF/layers/convRFF.py

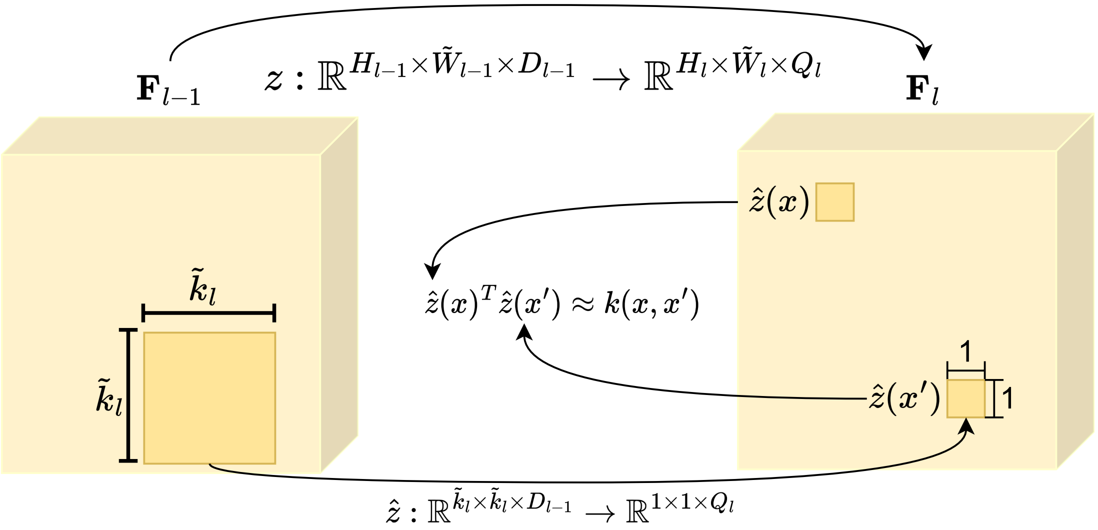

$\mathbf{F}_{l-1}$
$z : \mathbb{R}^{H_{l-1}\times\tilde{W}_{l-1}\times D_{l-1}} \to \mathbb{R}^{H_l\times\tilde{W}_l\times Q_l}$
$\mathbf{F}_l$
$\hat{z}(x)$
$\tilde{k}_l$
$\tilde{k}_l$
$\hat{z}(x)^T\hat{z}(x') \approx k(x,x')$
1
$\hat{z}(x')$
1
$\hat{z} : \mathbb{R}^{\tilde{k}_l\times\tilde{k}_l\times D_{l-1}} \to \mathbb{R}^{1\times 1\times Q_l}$

Figure **2-4**. RFF in Convolutional Form. The $\hat{\mathbf{z}}$ represents the one convolution operation in a certain location. Properties inherited from the convolution are observed in the $\hat{\mathbf{z}}$ and the properties of the kernel method in the representation of each pixel in the projected space though the dot product $\hat{\mathbf{z}}(\mathbf{x})^\top\hat{\mathbf{z}}(\mathbf{x}')$.

2.3 Experimental set-up

This section will detail the experimental configurations used to validate our CRFFg model. We conducted two types of experiments to evaluate the effectiveness of our model. First, we compared the dot product results of the mapped space with the Random Fourier Features (RFF) method. Second, we evaluated our model's performance on a classification task.

2.3.1 Equivariant characterization of CRFFg

To showcase the equivariant characterization property of the CRFFg, we contrast empirically both the input and output of the CRFFg across various scenarios involving the translation of the input interest region. Referencing Figure 2-5, we provide a visual representation of the synthetic images utilized in our study.

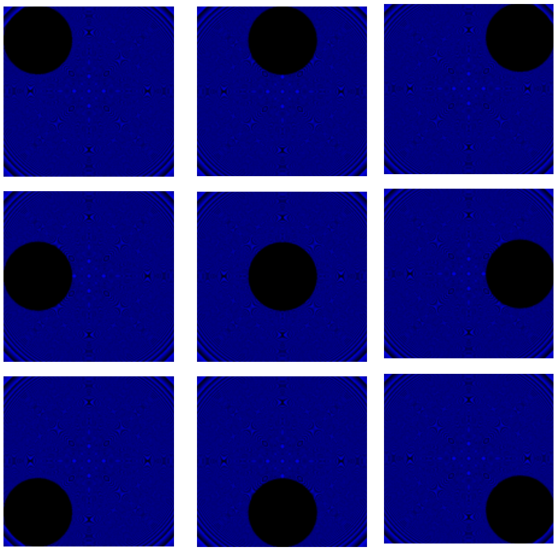


Figure **2-5**. Synthetic set of images to test the equivariant characterization property of CRFFg.

2.3.2 Comparison between CRFFg and RFF in the projected space

This study aims to compare two layers, namely CRFFg and RFF, using the difference of the dot product of each mapping $\mathbf{z}_{\text{RFF}}(\cdot; R)$ and $\mathbf{z}_{\text{CRFFg}}(\cdot; R)$ respectibly, where R denotes the output dimension. To carry out the experiment, we used 100 randomly selected images from the Fashion-MNIST dataset as data points. We varied the output dimension from 1 to 5000 for each layer and kept the receptive field of the CRFFg equal to the input image $\mathbf{x} \in \mathbb{R}^{H\times W}$, while for the RFF, the flattened result of the image was passed to it $\mathbf{x}' \in \mathbb{R}^{HW}$. Equation 2-19 shows the equation for comparing the resulted dot product of the projected space of each mapping.

$$\epsilon_{n,l} = |\langle \mathbf{z}_{\text{RFF}}(\mathbf{x}'_n; R_l), \mathbf{z}_{\text{RFF}}(\mathbf{x}'_n; R_l)\rangle - \langle \mathbf{z}_{\text{CRFFg}}(\mathbf{x}_n; R_l), \mathbf{z}_{\text{CRFFg}}(\mathbf{x}_n; R_l)\rangle| \tag{2-19}$$

2.3.3 CRFFg in Classification Tasks

To demonstrate the effectiveness of CRFFg in scenarios of data scarcity in classification tasks, we performed a classification task on Fashion-MNIST [Xiao et al., 2017b], comparing models with and without it.

Image classification is the task of assigning a label to an input image. Let $\{\mathbf{I}_n \in \mathbb{R}^{H\times W\times D},\ \mathbf{y}_n \in \{0,1\}^C\}_{n=1}^N$ be a set of N labeled images and their corresponding class labels. Each image $\mathbf{I}_n$ has a height of H, a width of W, and D color channels. In this case, the image classification problem is to predict the class label in one hot encoding $\mathbf{y}_n$ for each image.

A deep learning architecture for image classification often has a stack of convolutional layers, followed by one or more fully connected layers, as defined in Equation 2-20.

$$\hat{\mathbf{y}} = (\psi_L \circ \cdots \circ \psi_1)(\mathbf{I}) \quad \in [0,1]^C \tag{2-20}$$

where ψ_l represents the computation of new represented feature map $\mathbf{F}_l$ from previous $\mathbf{F}_{l-1}$, being for convolution $\psi_l : \mathbb{R}^{H_{l-1}\times W_{l-1}\times D_{l-1}} \to \mathbb{R}^{H_l\times W_l\times D_l}$ and for dense layers $\psi_l : \mathbb{R}^{Q_{l-1}} \to \mathbb{R}^{Q_l}$

$$\mathbf{F}_l = \psi_l(\mathbf{F}_{l-1}) = \nu(\Lambda(\mathbf{W}_l, \mathbf{F}_{l-1}) + \mathbf{b}_l) \tag{2-21}$$

where Λ is the operator for convolution or dense layers, $\mathbf{W}_l$ and $\mathbf{b}_l$ are the parameters, ν is the activation function. Then, the performance of the networks relies on the set of parameters $\Theta = \{\mathbf{W}_l, \mathbf{b}_l\}_{l=1}^L$, ending in the optimization problem shown in the Equation 2-23 through sample mean.

Name Layer	Type	Output Shape	Param #
Input	InputLayer	[(None, 28, 28, 1)]	0
CRFFg1 (Conv01*)	ConvRFF (Conv2D)	(None, 26, 26, 16)	161
MaxPool01	MaxPooling2D	(None, 13, 13, 16)	0
CRFFg2 (Conv02*)	ConvRFF (Conv2D)	(None, 11, 11, 32)	4640
MaxPool02	MaxPooling2D	(None, 5, 5, 32)	0
CRFFg3 (Conv03*)	ConvRFF (Conv2D)	(None, 3, 3, 64)	18486
Flatten	Flatten	(None, 576)	0
Dense(RFF$^+$)	Dense(RFF)	(None, 32)	18464
Output	Dense	(None, 10)	330

Table **2-1**. Architecture For Classification Experiment.

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}\big\{\mathcal{L}(\mathbf{y}_n, \hat{\mathbf{y}}_n|\Theta) \ : \ \forall\, n \in \{1, 2, \dots, N\}\big\} \tag{2-22}$$

$$\approx \arg\min_{\Theta} \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(\mathbf{y}_n, \hat{\mathbf{y}}_n|\Theta) \tag{2-23}$$

where $\mathcal{L} : \{0,1\}^C \times [0,1]^C \to \mathbb{R}$ is the loss function.

Deep Learning Architectures

Table 2-1 lists the deep learning architectures utilized in this study The first architecture described in this chapter is based on CRFFg, followed by a second architecture that utilizes a standard convolutional layer instead of CRFFg. Lastly, the second architecture replaces the dense layer with an RFF layer. The models maintain a consistent number of filters and kernel sizes throughout.

To optimize the parameters of the models, a cross-entropy loss function is used, as shown in Equation 2-24

$$\Theta^* \approx \arg\min_{\Theta} \frac{1}{N}\sum_{n=1}^{N} -\mathbf{y}_n^\top \log(\hat{\mathbf{y}_n}) \tag{2-24}$$

All experiments are carried out in Python 3.8, with the Tensorflow 2.4.1 API, on a Google Colaboratory environment (Code Colab).

Training details and method comparison

To measure the performance of the models, the next metrics are used:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \\
\text{Precision} &= \frac{TP}{TP+FP} \\
\text{Recall} &= \frac{TP}{TP+FN} \\
\text{F1-score} &= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

In these equations, TP represents the number of true positive predictions, TN represents the number of true negative predictions, FP represents the number of false positive predictions, and FN represents the number of false negative predictions. These measures are commonly used in machine learning and data analysis to evaluate the performance of classification models. As for the validation strategy, we selected the hold-out cross-validation strategy with the following partitions: 84% of the samples for training, 1% for validation, and 15% for testing. Moreover, we vary the size of the training partition keeping the same test partition. To ensure reliable results, we experimented 22 times per variation on the train size, mitigating the variability and potential differences in the training sets.

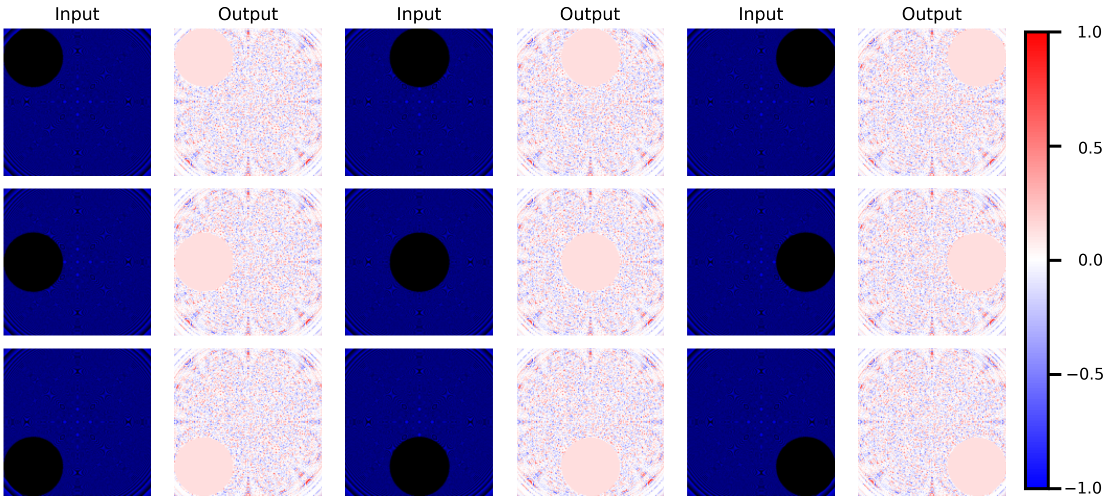


Figure **2-6**. Translation equivariant characterization property of CRFFg. The input and output are interleaved by the translation equivariant property of CRFFg

2.4 Results and discussion

2.4.1 Equivariant characterization of CRFFg

Figure 2-6 vividly illustrates the input-output relationship within the context of the CRFFg. Evidently, the output features faithfully mirror the input translations across all scenarios, highlighting a remarkable property of translation equivariance inherent in the CRFFg. This entwining of input and output underscores the CRFFg's ability to maintain consistent characteristics amidst varying translations.

Furthermore, we can observe the locality property. Even when the circle's translation varies across different locations, the structural integrity of regions distant from the circle remains unaffected.

2.4.2 Comparison between CRFFg and RFF in the projected space

Figure 2-7 displays the comparison results using Equation 2-19 by calculating the mean and standard deviation from the 100 data points. Ideally, both the mean and

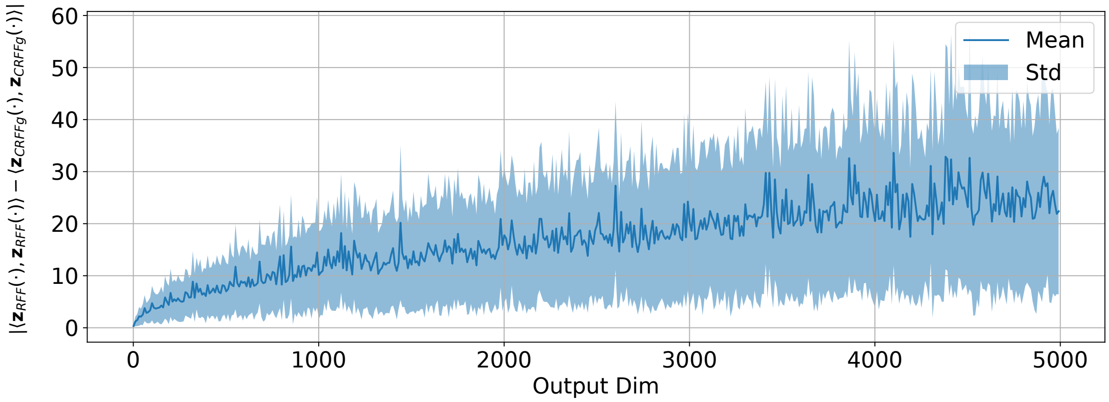


Figure **2-7**. Comparison between the RFF and CRFFg on Fashion-MNIST dataset.

standard deviation should approach zero, indicating that both methods perform similarly under these conditions. However, as the output dimension increases, the inherent randomness also increases, leading to an observed increase in both the mean and standard deviation. Despite this increase, the difference remains relatively low compared to the output dimension. Therefore, both operations RFF and CRFFg are comparable regarding the dot product on the mapped space.

In conclusion, the results from the comparison of RFF and CRFFg using Equation 2-19 suggest that both methods perform similarly under the given conditions. Although the mean and standard deviation increase with the output dimension, the difference between the two methods remains relatively low compared to the output dimension. As a result, both RFF and CRFFg can be considered comparable in terms of their performance.

2.4.3 CRFFg in Classification Tasks

The results of our classification experiment are summarized in Figure 2-8, where we evaluate the performance of three models: CRFFg, CNN, and RFF. Our analysis reveals that the differences in performance are most pronounced in small datasets containing less than 400 samples. In this range, CRFFg outperforms both CNN and

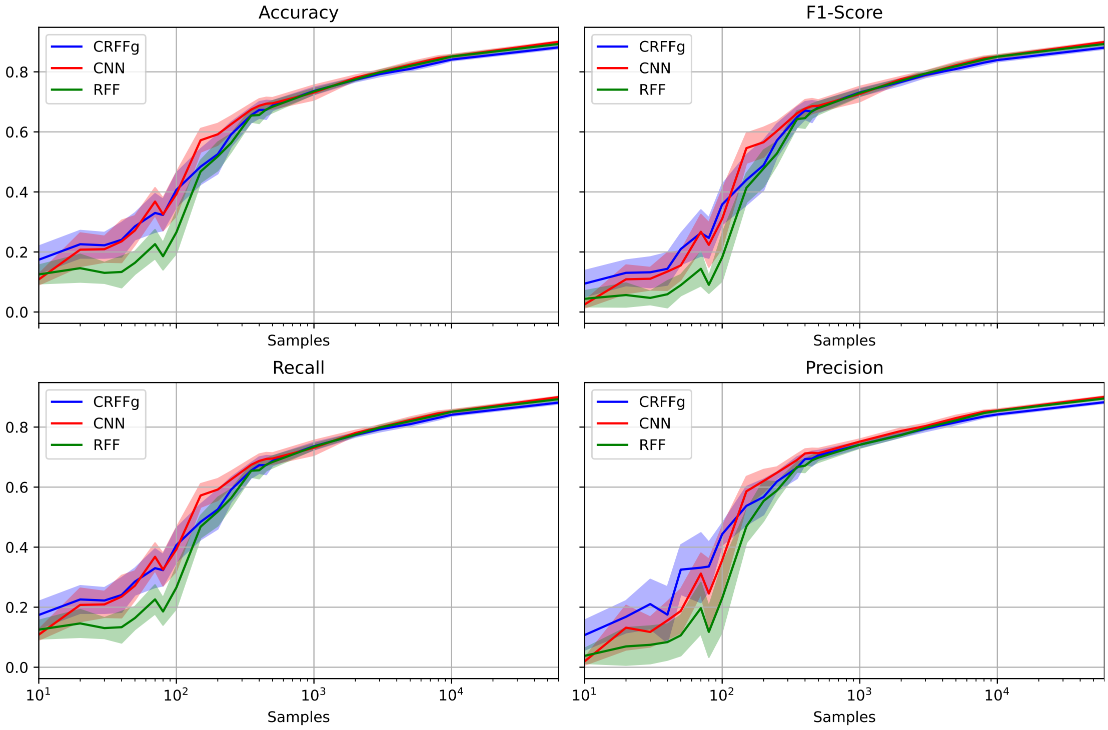


Figure **2-8**. Comparison CRFFg, CNN and RFF in metrics performances vs size of data using Fashion-MNIST dataset.

RFF, while RFF performs the worst. However, as the dataset size increases, the performance of the three models becomes comparable. It's worth noting that the trainable parameters of CRFFg allow the model to perform similarly to standard convolutional networks in larger datasets. Overall, our results demonstrate that CRFFg can capture both kernel methods' generalization capability under low sample sizes and the local and equivariant characterization of CNNs.

Weights Distributions

Figure 2-9 illustrates the weight distributions of models at three convolutional layers, corresponding to two specific sizes of the training set: 30 and 100,000 samples. Upon initial observation, it is evident that the weight behavior is significantly influenced by the value of σ, as indicated by the orange color. Furthermore, the introduction of σ helps to narrow the weight distributions, aligning them with those of the other models. Notably, as we progress from shallower to deeper layers, the weight distributions exhibit a tendency to become increasingly narrow. This can be attributed to the deeper layers needing to capture more specialized features compared to the shallow layers. Lastly, a substantial change in the weight distribution is noticeable when transitioning from models with standard convolutional layers to the model with the CRFFg layer, particularly when increasing the partition set from 30 to 10,000 samples.

2.5 Summary

In this chapter, we presented an extension of the RFF for spatial data, which optimizes the technique through gradient descent to improve generalization when data is scarce. Our proposed method, the CRFFg, provides a local and equivariant characterization, enhancing its effectiveness for spatial data. Our experiments demonstrate that the CRFFg performs comparably to both the RFF in the mapped space and the standard convolution in image classification tasks, particularly improving the performance in low-data scenarios. However, as previously discussed, the utilization of CRFFg holds great potential in pixel-wise classification tasks, primarily due to its ability to approximate the kernel in the projected space. Therefore, in the upcoming chapter, we delve into the practical application of CRFFg in pixel-wise classification, focusing specifically on semantic segmentation.

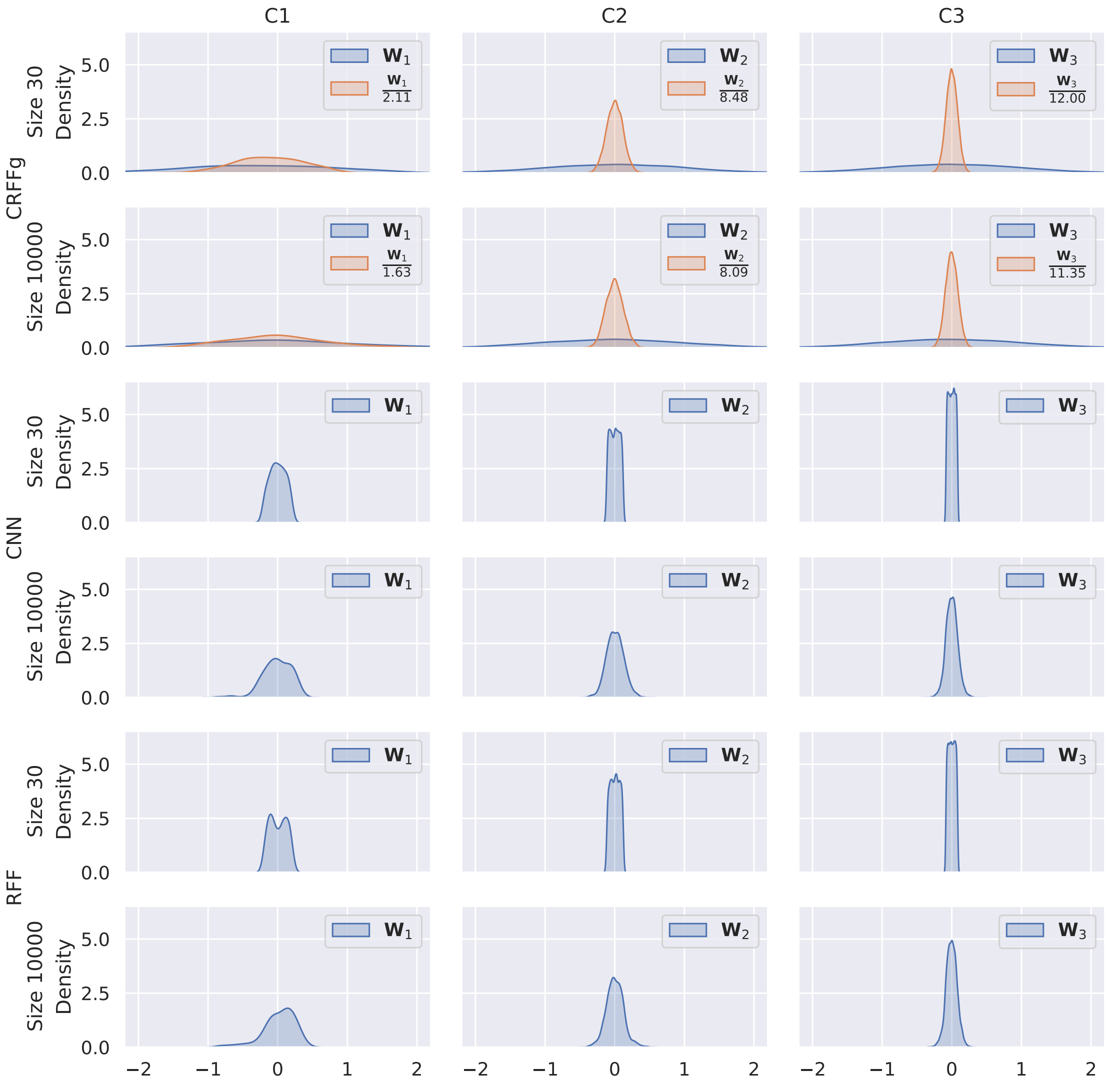


Figure **2-9**. Weights Distributions of models at three convolutional layers, correspond to two specific sizes of the training set: 30 and 100,000 samples. The orange color denoted the weights of CRFFg models divided by the scale σ.

CHAPTER

THREE

CONVOLUTIONAL RANDOM FOURIER FEATURES ON SEMANTIC SEGMENTATION TASKS

In the previous chapter, we introduced the concept of Random Fourier Features in convolution form, also known as CRFFg. In this chapter, we will explore its practical application in the specific task of semantic segmentation. By leveraging the power of CRFFg, we aim to enhance the representation of features captured at the skip connection of encoder-decoder architectures, thereby improving the accuracy and efficiency of semantic segmentation algorithms.

3.1 Semantic Segmentation Tasks

Semantic segmentation (SS) is defined as pixel-level classification, where each label belongs to a specific semantic category [Mo et al., 2022]. Let $\{\mathbf{I}_n \in \mathbb{R}^{H\times W\times D}, \mathbf{M}_n \in \{0,1\}^{H\times W\times C}\}_{n=1}^{N}$ be a set of N labeled images and their corresponding masks. Each image $\mathbf{I}_n$ has a height of H, a width of W, and D color channels. The mask $\mathbf{M}_n$

encodes the membership of each pixel to a particular class $c \in \{0, 1, \ldots, C\}$. In this case, the semantic segmentation problem is limited to binary segmentation, such as distinguishing between the background and the foreground.

A deep learning architecture for semantic segmentation often has a stack of convolutional layers, as defined in Equation 3-1.

$$\hat{\mathbf{M}} = (\psi_L \circ \cdots \circ \psi_1)(\mathbf{I}) \quad \in [0,1]^{H \times W \times C} \tag{3-1}$$

where $\psi_l : \mathbb{R}^{H_{l-1} \times W_{l-1} \times D_{l-1}} \to \mathbb{R}^{H_l \times W_l \times D_l}$ does the computation of new represented feature map $\mathbf{F}_l$ from previous $\mathbf{F}_{l-1}$ as Equation 3-2.

$$\mathbf{F}_l = \psi_l(\mathbf{F}_{l-1}) = \nu(\mathbf{W}_l \otimes \mathbf{F}_{l-1} + \mathbf{b}_l) \tag{3-2}$$

where $\mathbf{W}_l \in \mathbb{R}^{Q_l \times Q_l \times D_{l-1} \times D_l}$ and $\mathbf{b}_l \in \mathbb{R}^{D_l}$ are the parameters, ν is the activation function, and $\otimes$ stands for image-based convolution. Then, the performance of the networks relies on the set of parameters $\Theta = \{\mathbf{W}_l, \mathbf{b}_l\}_{l=1}^{L}$, ending in the optimization problem shown in the Equation 3-4 through sample mean.

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}\big\{\mathcal{L}(\mathbf{M}_n, \hat{\mathbf{M}}_n | \Theta) \ : \ \forall\, n \in \{1, 2, \ldots, N\}\big\} \tag{3-3}$$

$$\approx \arg\min_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\mathbf{M}_n, \hat{\mathbf{M}}_n | \Theta) \tag{3-4}$$

where $\mathcal{L} : \{0,1\}^{H \times W \times C} \times [0,1]^{H \times W \times C} \to \mathbb{R}$ is the loss function.

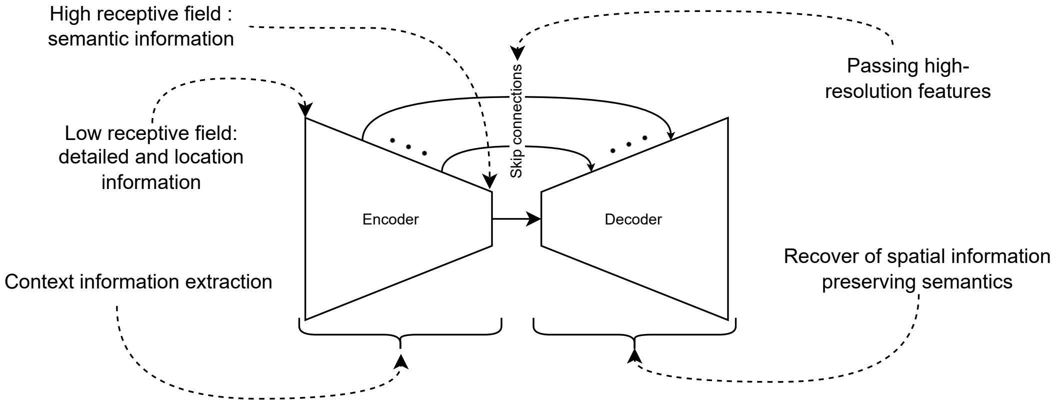


Figure **3-1**. Encoder-decoer architecture for SS.

3.2 Convolutional Neural Networks Architectures for SS

To perform SS, a convolutional neural network must incorporate location and semantic information [Mo et al., 2022]. This is why the architectures for SS are typically divided into two main parts: the encoder and the decoder [Long et al., 2014, Ronneberger et al., 2015, Huang et al., 2020, Zhou et al., 2020, Zunair and Hamza, 2021, Jafari et al., 2020]. The encoder is responsible for extracting high-level features from the input image by capturing global information through layers with large receptive fields. However, this can result in the loss of location information. On the other hand, the decoder aims to recover the lost location information while retaining the semantic information. This is typically achieved through the use of upsampling or transpose convolution, which gradually expands the semantics to a higher resolution. In addition, skip connections between the encoder and decoder parts are often used to recover lost location information. By combining the encoder and decoder components, a convolutional neural network is able to effectively capture both spatial and semantic information, enabling it to segment complex images accurately. However, there is a mismatch between the semantics of the encoder and the decoder [Ibtehaz and Rahman, 2020b, Zhou et al., 2018b, Wang et al., 2021]. Therefore, we decided to locate our CRFFg layer

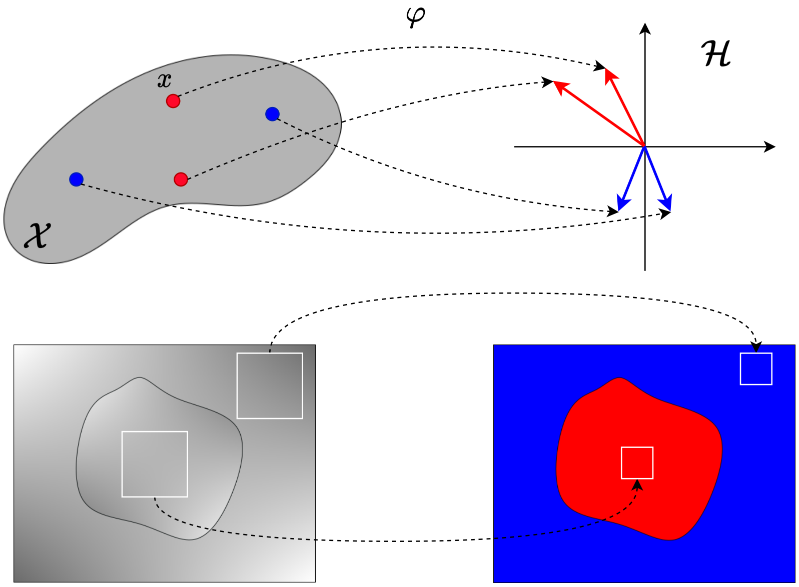


Figure 3-2. Analogy on mapping of CRFFg in case of images in SS in ideal case.

at the skip connections of encoder-decoder architectures to improve the representation of low-level features and the localization information.

3.3 Experimental Set-Up

This section will detail the experimental configurations used to validate our CRFFg model in semantic segmentation tasks. We perform semantic segmentation on the databases described in Section 1.6.

Deep Learning Architectures

In order to test the CRFFg layer, we use the following well-known shallow architectures that hold the definitions of Section 3.1:

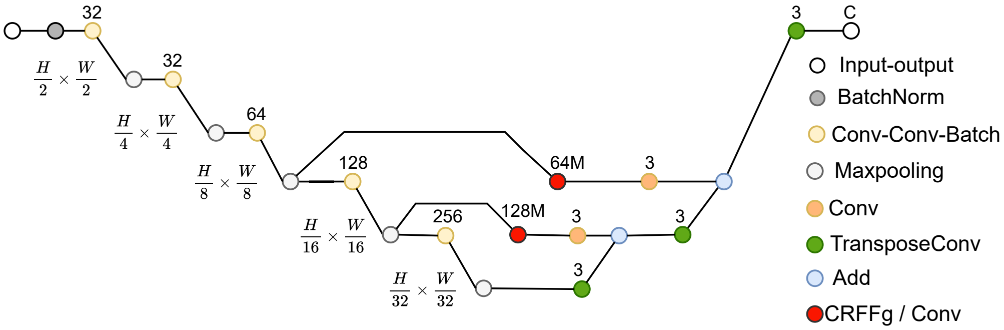

32
$\frac{H}{2} \times \frac{W}{2}$
32
$\frac{H}{4} \times \frac{W}{4}$
64
$\frac{H}{8} \times \frac{W}{8}$
128
$\frac{H}{16} \times \frac{W}{16}$
256
$\frac{H}{32} \times \frac{W}{32}$
64M
3
128M
3
3
3
3
C
Input-output
BatchNorm
Conv-Conv-Batch
Maxpooling
Conv
TransposeConv
Add
CRFFg / Conv

Figure 3-3. The FCN Architecture.

- **Fully convolutional network (FCN)** [Long et al., 2014]: The FCN architecture is based on the VGG (Very Deep Convolutional Network) [Simonyan and Zisserman, 2014] model, which is well-known for its ability to recognize large-scale images. By exclusively using convolutional layers, the FCN can output a segmentation map with pixel-level accuracy while also reducing computational costs.

- **U-Net** [Ronneberger et al., 2015]: The U-Net architecture comprises two main parts, the encoder and the decoder. The encoder consists of a series of convolutional layers that reduce the spatial dimensions of the input image. The decoder is a series of upsampling layers that upsample the encoded features back to the original input image size.

- **ResUNet** [Anas et al., 2017b]: The ResUNet architecture is an extension of the U-Net architecture that uses residual connections to improve performance. Residual connections allow for the gradient to flow directly through the network, improving the training of deeper models.

These architectures, FCN, U-Net and ResUNet, are depicted in Figures 3-3, 3-5, and 3-4 respectively.

To compare the performance of the proposed CRFFg-layer strategy, we implemented a standard convolutional layer with an equal number of filters and a

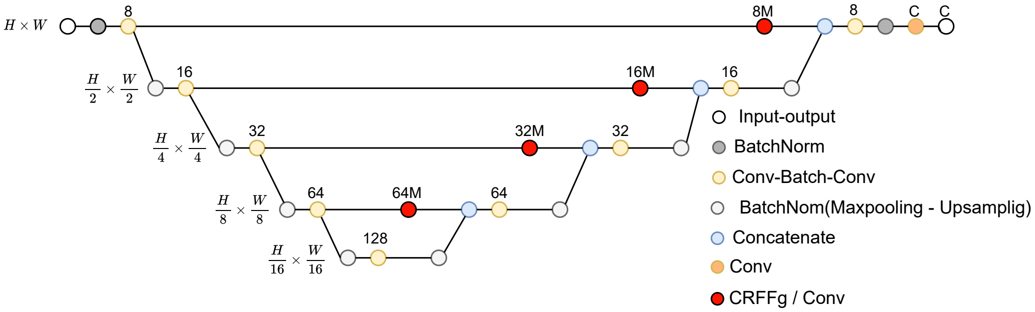

$H \times W$
8
8M
8
C
C
$\frac{H}{2} \times \frac{W}{2}$
16
16M
16
$\frac{H}{4} \times \frac{W}{4}$
32
32M
32
$\frac{H}{8} \times \frac{W}{8}$
64
64M
64
$\frac{H}{16} \times \frac{W}{16}$
128
Input-output
BatchNorm
Conv-Batch-Conv
BatchNom(Maxpooling - Upsamplig)
Concatenate
Conv
CRFFg / Conv

Figure **3-4**. U-Net Architecture.

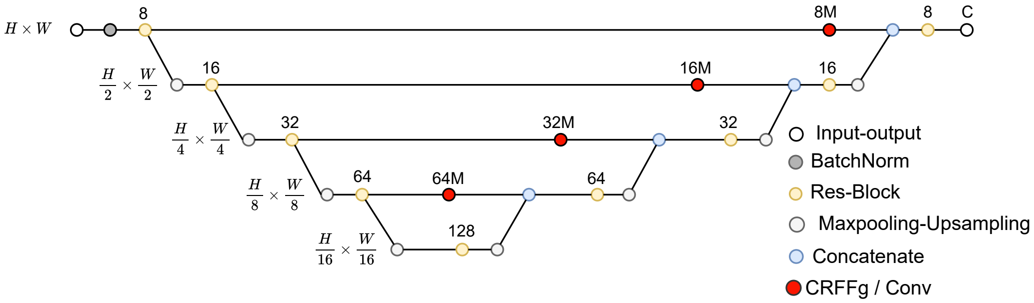

$H \times W$
8
8M
8
C
$\frac{H}{2} \times \frac{W}{2}$
16
16M
16
$\frac{H}{4} \times \frac{W}{4}$
32
32M
32
$\frac{H}{8} \times \frac{W}{8}$
64
64M
64
$\frac{H}{16} \times \frac{W}{16}$
128
Input-output
BatchNorm
Res-Block
Maxpooling-Upsampling
Concatenate
CRFFg / Conv

Figure **3-5**. ResUNet Architecture.

ReLU activation function at the same position in the architecture. We assessed the influence of the CRFFg layer dimension on segmentation performance by testing two multiplication values (one and three). To examine the impact of CRFFg, we set all model hyperparameters to the same values. We used the Adam optimizer with default Keras parameter values and employed the dice-based loss function shown in Eq. 3-5:

$$\mathcal{L}_{Dice}(\mathbf{M}_n, \hat{\mathbf{M}}_n) = 2\frac{\mathbf{1}^\top(\mathbf{M}_n \odot \hat{\mathbf{M}}_n)\mathbf{1} + \epsilon}{\mathbf{1}^\top\mathbf{M}_n\mathbf{1} + \mathbf{1}^\top\hat{\mathbf{M}}_n\mathbf{1} + \epsilon} \tag{3-5}$$

where $\mathbf{M}_n$ is the ground truth segmentation map for the n-th sample, $\hat{\mathbf{M}}_n$ is the predicted segmentation map for the n-th sample, $\odot$ denotes the element-wise product, and ϵ is a small constant added to avoid umerical instability. All experiments are carried out in Python 3.8, with the Tensorflow 2.4.1 API, on a Google Colaboratory environment (code repository: `https://github.com/aguirrejuan/ConvRFF`, accessed on 25 April 2023).

Training details and method comparisons

To evaluate the models' performance, three dataset scenarios are considered. ThermalFeet, ThermalFeet with data augmentation, and Oxford IIIt pet. The Data augmentation procedure is presented in Table **3-1**

Parameter	**Infrared Therma Images**	**Oxford Pet**
Shape	224,224,1	256,256,3
Partions	Train: 0.81 Val: 0.09 Test: 0.1	Train: 0.64 Val: 0.16 Test: 0.2
Repeat	7	1
Flip Left-Right	True	True
Flip Up-Down	False	True
Rotation Range	(-15,15)	(-10,10)
Translation	0.10	None
Zoom	0.15	None

Table **3-1**. Dataset Parameters.

To evaluate and compare the performance of the models, we used the following metrics:

$$\begin{aligned}
\text{Dice} &= \frac{2|\mathbf{M} \cap \hat{\mathbf{M}}|}{|\mathbf{M}| + |\hat{\mathbf{M}}|} = \frac{2TP}{2TP + FP + FN} \\
\text{Jaccard} &= \frac{|\mathbf{M} \cap \hat{\mathbf{M}}|}{|\mathbf{M} \cup \hat{\mathbf{M}}|} = \frac{TP}{FN + FP + TP} \\
\text{Sensitivity} &= \frac{|\mathbf{M} \cap \hat{\mathbf{M}}|}{|\mathbf{M} \cap \hat{\mathbf{M}}| + |\mathbf{M} \cap \neg\hat{\mathbf{M}}|} = \frac{TP}{TP + FN} \\
\text{Specificity} &= \frac{|\neg\mathbf{M} \cap \neg\hat{\mathbf{M}}|}{|\neg\mathbf{M} \cap \neg\hat{\mathbf{M}}| + |\neg\mathbf{M} \cap \hat{\mathbf{M}}|} = \frac{TN}{TN + FP}
\end{aligned}$$

Where TP (True Positives) measures the pixels correctly labeled as target foreground, TN (True Negatives) the ones correctly labeled as background, FP (False Positives) the background pixels labeled as foreground, and FN the foreground pixels, which are labeled as background. The metrics are depicted in Figure 3-6.

Finally, the hyperparameters for training were all the same for all the experiments conducted, and those are: Epochs equal to 200, and Adam Optimizer with a learning rate equal to 1e-3

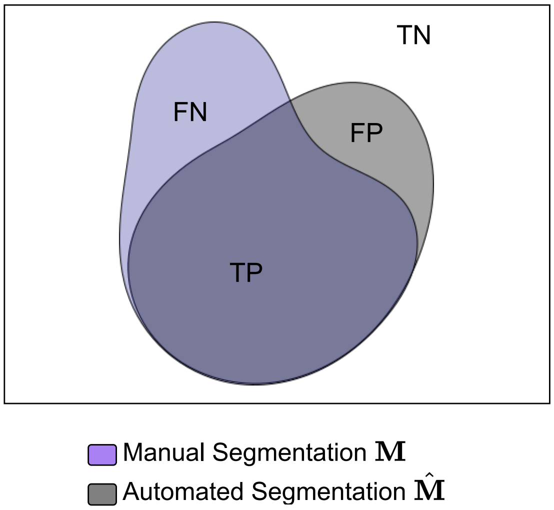

TN
FN
FP
TP
🟪 Manual Segmentation $\mathbf{M}$
⬛ Automated Segmentation $\hat{\mathbf{M}}$

Figure **3-6**. Visual Representation of performance metrics.

3.4 Results and Discussion

This section presents the key findings from our experiments for each dataset analyzed.

3.4.1 Visual Inspection Results

Figure 3-7 shows results obtained from the ThermalFeet database without data augmentation, where each row represents a different architecture: FCN in the first row, U-Net in the second row, and ResUNet in the third row. As expected, the performance of the models under a small-size dataset is poor. The regions of faster change in temperature, which characterize the dataset, are where the models struggle more. At first glance, we observe the FCN architecture is the one that struggles the most, having high false positives regions in regions that exhibit low-high temperatures.

Figure 3-8 shows results obtained incorporating data augmentation. At an initial glance, the positive impact of the data augmentation on the resulting segmentation of all the models is visible. Moreover, FCN architectures produce smoother borders and fewer false positives than other architectures. This can be explained due to the high receptive field that possesses the FCN architecture, allowing it to capture complex and heterogeneous regions (the variability of the temperatures) that compose the feet.

When comparing FCN models with a multiplication factor of 1 (M1), the model with CRFFg (blue) generally outperforms the models without it (orange) in terms of pixel membership prediction (sensitivity). However, This trend does not hold when the multiplication factor is increased to 3 (M3), probably because the large model is a propensity to overfit, making the prediction less confident in new data points.

On the other hand, U-Net models struggle with regions that exhibit fast temperature changes. This can be explained by the same characteristic the FCN model posses, but the U-Net model does not have a high receptive field that allows it to characterize high heterogeneous feet. Among the U-Net models, U-Net CRFFg Skips-M1 performs satisfactorily with low false positives and high false negatives, while its direct competitor, U-Net Skips-M1, shows the opposite trend. Similarly, using CRFFg in the other U-Net models reduces the number of false positives. The reason behind the reduction of false positives by the CRFFg layer could be related to overfitting.

Finally, the ResUNet architecture seems to have the same behavior as the U-Net architecture but with smoother borders, which can be explained due to the multiple stack layers at the ResBlock, which increase multiple steps of representation, allowing to capture of useful representation. The ResUNet Skips-M1 model works better on average than the other models; adding layers at the skip connections appears to reduce performance, creating false positives and false negatives. This can be explained by the small size of the dataset and the large model. Specifically, using CRFFg with ResUNet does not result in significant improvements.

Figure 3-9 shows results on the Oxford-IIIT Pet database. Once again, FCN architectures produce smoother borders than the other models. Additionally,

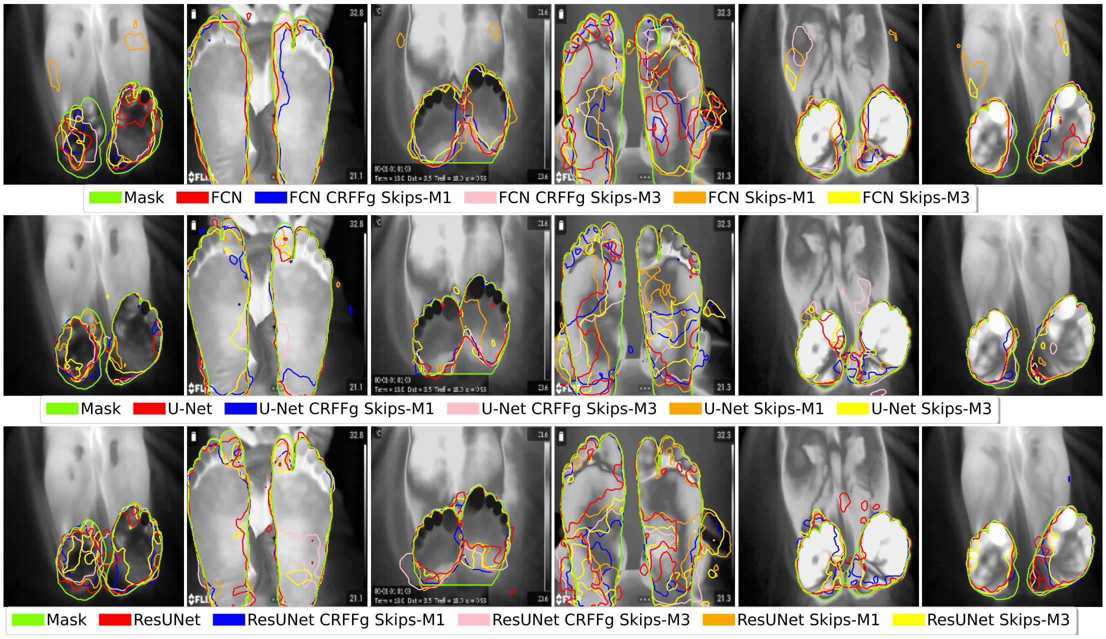


Figure **3-7**. Visual inspection of the results on the ThermalFeet database without data augmentation. The first row displays results for FCN, the second for U-Net, and the third for ResUNet. A unique color differentiates each model within an architecture.

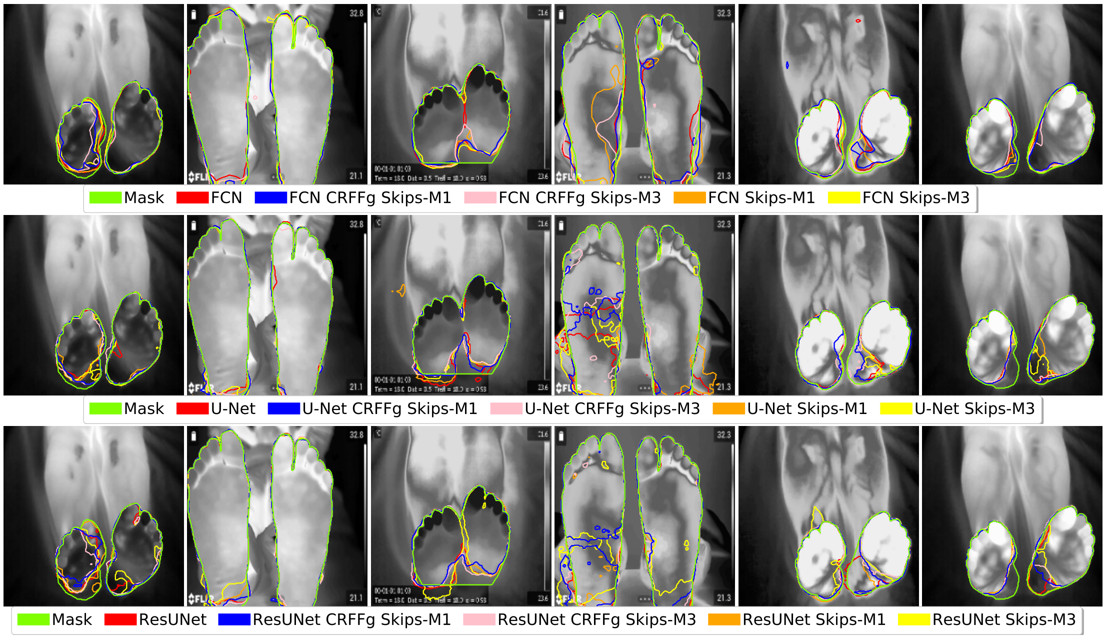


Figure **3-8**. Visual inspection of the results on the ThermalFeet database with data augmentation for all the models.

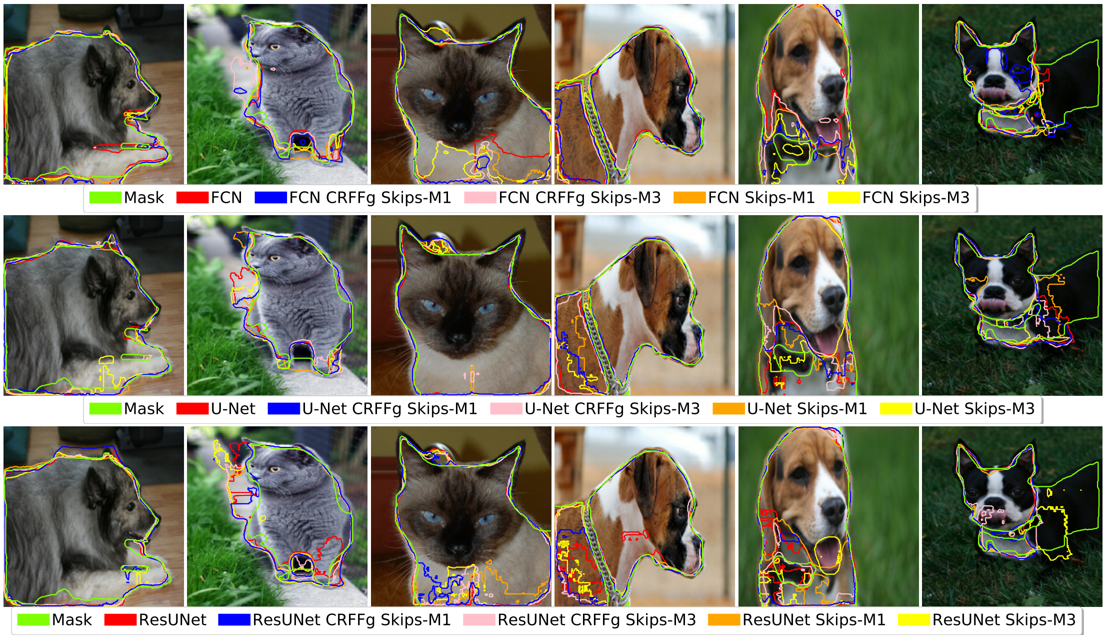


Figure **3-9**. Visual inspection of the results on Oxford-IIIT Pet database, for all the models. The first row shows the results for the FCN architecture, the second row for U-Net, and the third row for ResUNet. A unique color differentiates each model within an architecture.

adding layers to the skip connections improves overall performance by reducing false negatives. Most models can avoid over-segmentation, although the region with a similar color to the cat is also segmented in the case of the gray cat. Even though adding layers to the skip connections tends to help, which could be explained by the large dataset size.

3.4.2 Performance Metrics

Figures 3-10, 3-11, and 3-12 illustrate the learning curves of the compared models, depicting the training loss versus epochs. A visual inspection reveals notable differences between the curves with and without data augmentation. When data augmentation is not applied, the algorithms initially exhibit higher validation loss for the first 40 epochs, but subsequently demonstrate a downward trend. The learning curves exhibit increased noise due to the limited dataset size, which poses challenges in capturing generalized features early in training. In the validation partition, when data augmentation is not applied, certain models display a pattern of initial loss increase followed by a subsequent decrease. This phenomenon can be attributed to the impact of the momentum within the Adam optimizer. Such behavior can lead the optimization process towards local minima. In contrast, the data augmentation scenario consistently shows decreasing training and validation losses with minor noise in the validation partition. The FCN CRFFg S-M3 and, to a lesser extent, FCN CRFFg S-M1 demonstrate faster decreases in validation loss, attributed to the generalization capabilities of the RFF from kernel methods. In the ResUNet architectures, ResUNet S-M3 experiences an early decline but reaches a non-minimum early, while no apparent differences are observed within the U-Net architectures. It is worth noting that the models in the data augmentation scenario exhibit similar behavior. While we do not display the learning curves of certain models, specifically the one trained in Oxford IIIt Pet, in this section, they can be found in Appendix A for those interested in visually inspecting the training process.

Figure 3-13 displays the values of semantic segmentation performance for the ThermalFeet dataset achieved by each compared DL architecture: FCN (colored in purple), ResNet (orange), U-Net (green). For interpretation purposes, the results are presented for the evaluation measures separately, but keeping with and without data augmentation scenarios at the same plot. At first glance, the models with data augmentation, more saturated color, perform better than the contrary scenario, as expected. As seen, the specificity estimates are very close to the maximal value and show the lowest variability. This result can be explained by the relatively small feet sizes compared with the background, making their correct detection and segmentation more difficult. On the contrary, sensitivity assessments are of less value and have much more variability, accounting for the diversity in the regions of interest (i.e., size, shape, and location). Due to the changing behavior of thermal patterns and the limited datasets available, learners have difficulty obtaining an accurate model.

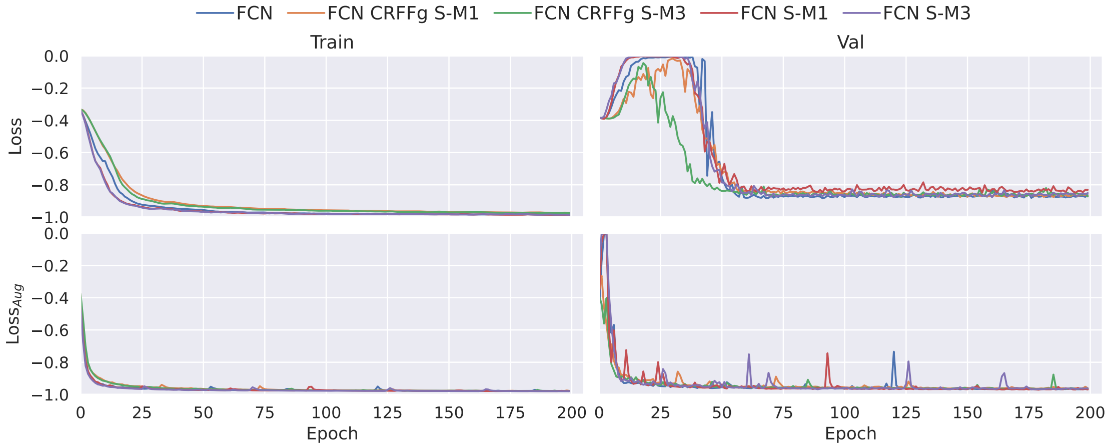


Figure **3-10**. Learning curve of FCN models.

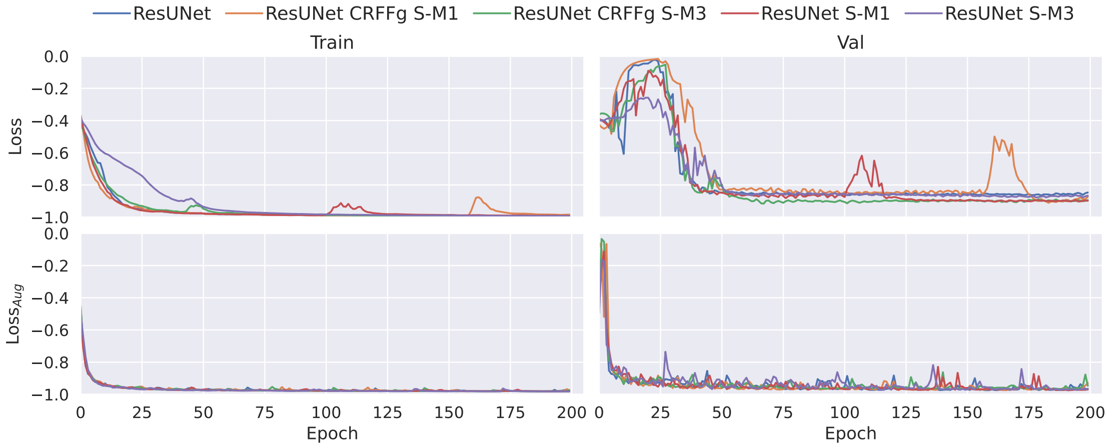


Figure **3-11**. Learning curve of ResUNet models.

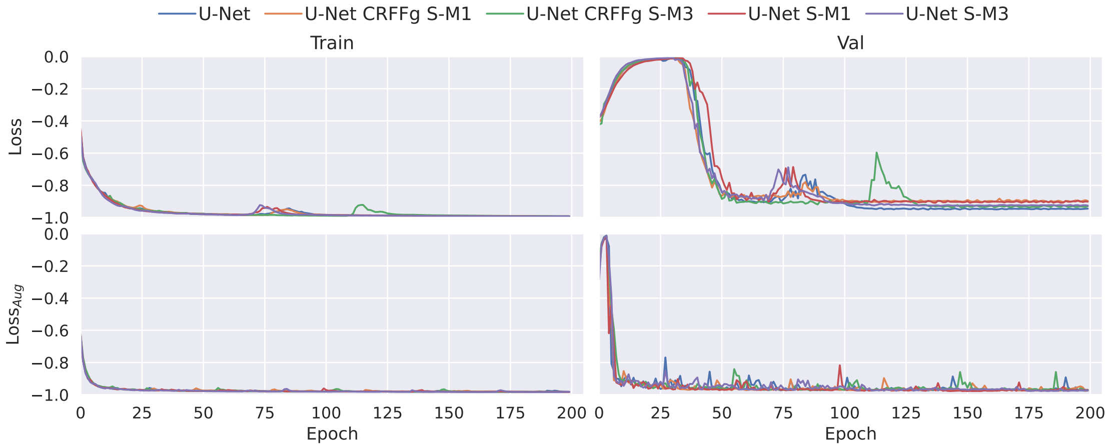


Figure **3-12**. Learning curve of U-Net models.

Regarding overlapping between estimated thermal masks, the Dice value is acceptable but with higher variance values for FCN, implying that other tested models segment complex shapes more accurately. As expected, the Jaccaed index mean values resemble the Dice assessments, although with increased variance, highlighting the mismatch between the ground truth and the predicted mask even more.

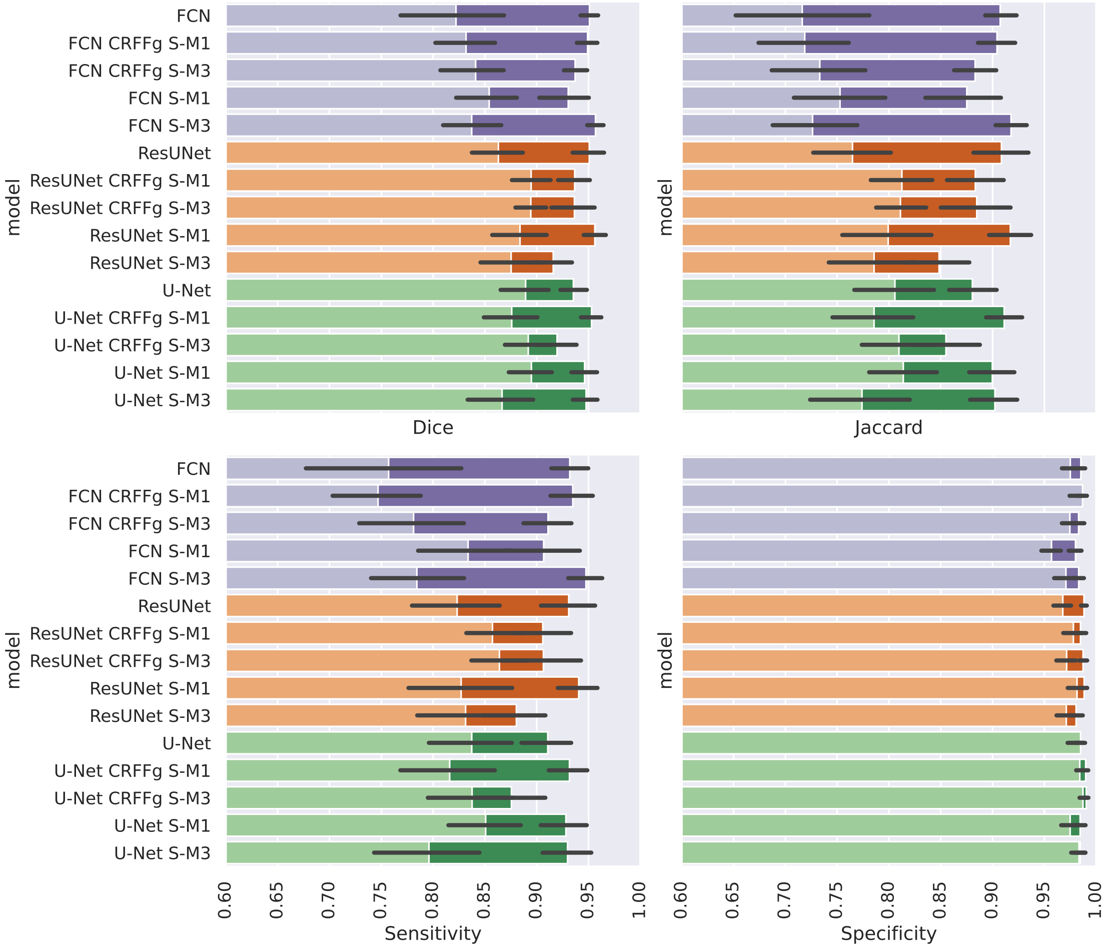


Figure **3-13**. Performance metrics results on ThermalFeet. More saturated color denotes data augmentation.

A comparison between the segmentation metric value achieved by the baseline DL architecture (without any modifications) and the value estimated for every

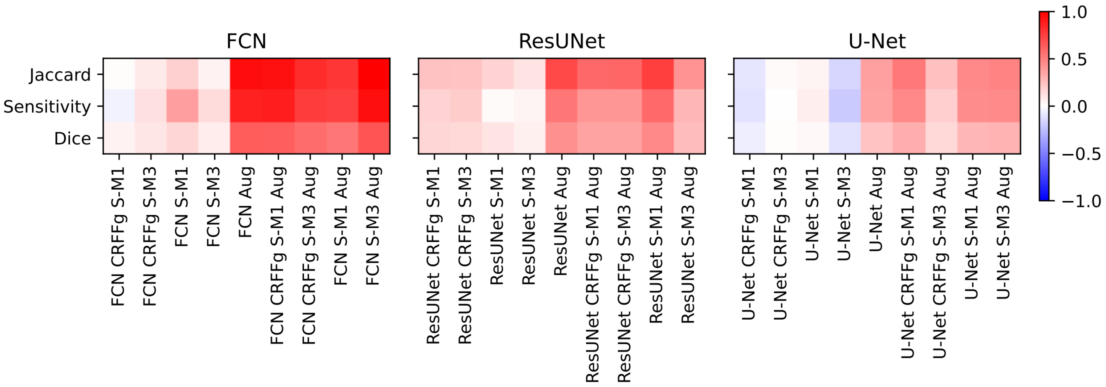


Figure **3-14**. Comparison with baseline models on ThermalFeet.

evaluated DL semantic segmentation strategy is presented in Figure 3-14. Note that specificity is removed because its estimates are obtained with minimal variations.

As seen, the performance improvement depends on the learner model size (also called algorithm complexity). Namely, the baseline architecture of FCN holds 1,197,375 parameters, baseline ResUnet - 643,549, and baseline Unet - 494,093. Thus, the FCN model contains the largest tuning parameter set and achieves the poorest performance, but it benefits the most from the evaluated architectures. As data augmentation is also applied, this finding becomes more evident. It may be pointed out that adding new data decreases model overfitting inherent to massive model sizes. Likewise, the following ResUnet model takes advantage of the enhanced architecture strategy and improves performance. It increases more by generating new data points, however, to a lesser extent. Lastly, the learner with the lowest parameter set gets almost no benefits or is negatively affected by the strategies considered for architecture enhancement. Still, the strategies taken into account combined with expanded training data sizes can be improved, though very modestly.

We present the rank of ThermalFeet models in two scenarios: with and without data augmentation. Figure 4-5 displays the rank without data augmentation. Upon initial inspection, we observe that FCN models consistently rank last, aligning with previous experiments' findings. Conversely, ResUNet models with the CRFFg layer perform the best across most cases. U-Net models are comparable to ResUNet models, but the highest-performing Ones did not include a CRFFg layer.

In the scenario with data augmentation, Figure 4-7 shows the rank. Here is not evident which architecture performs better. However, we can see that the best models tend to be the ones that have a layer at the skip connections, and the influence of CRFFg is not clear.

In reference to the Oxford-IIIT Pet dataset, the metrics results are illustrated in Figure 3-17. There is no significant variation in performance between the models. This can be attributed to the large size of the dataset, which leads to the models behaving similarly.

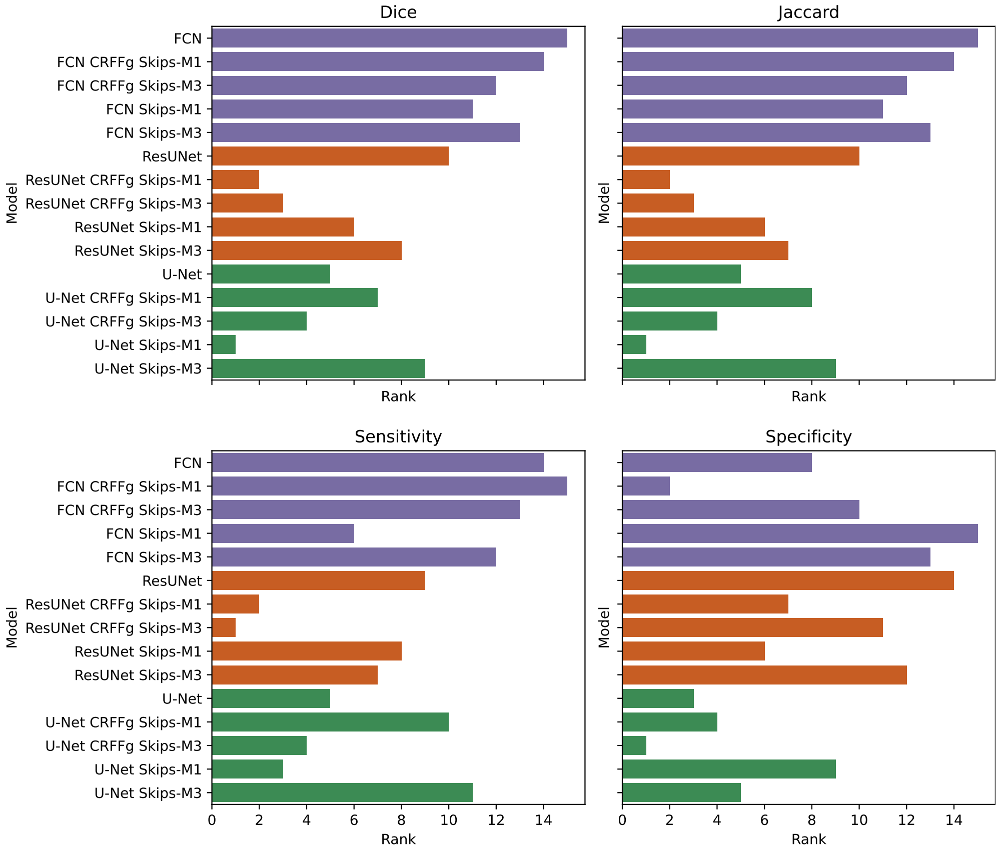


Figure **3-15**. Performance Rank Over ThermalFeet without Data Augmentation.

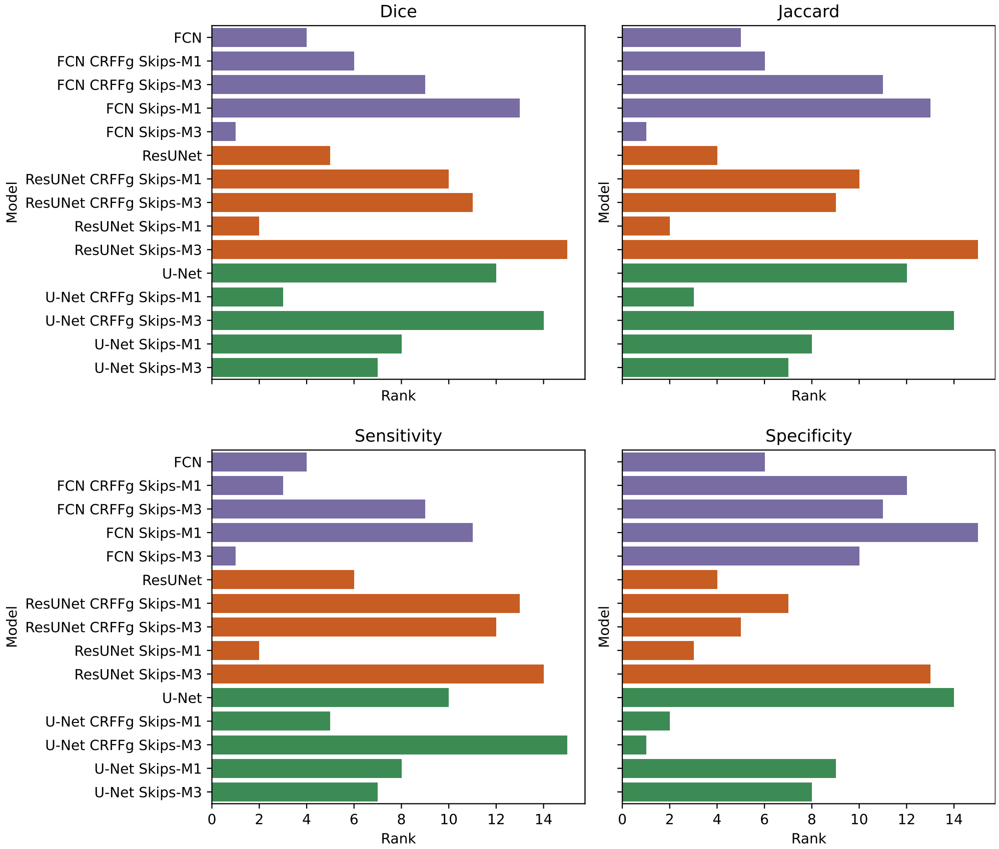


Figure **3-16**. Performance Rank Over ThermalFeet with Data Augmentation.

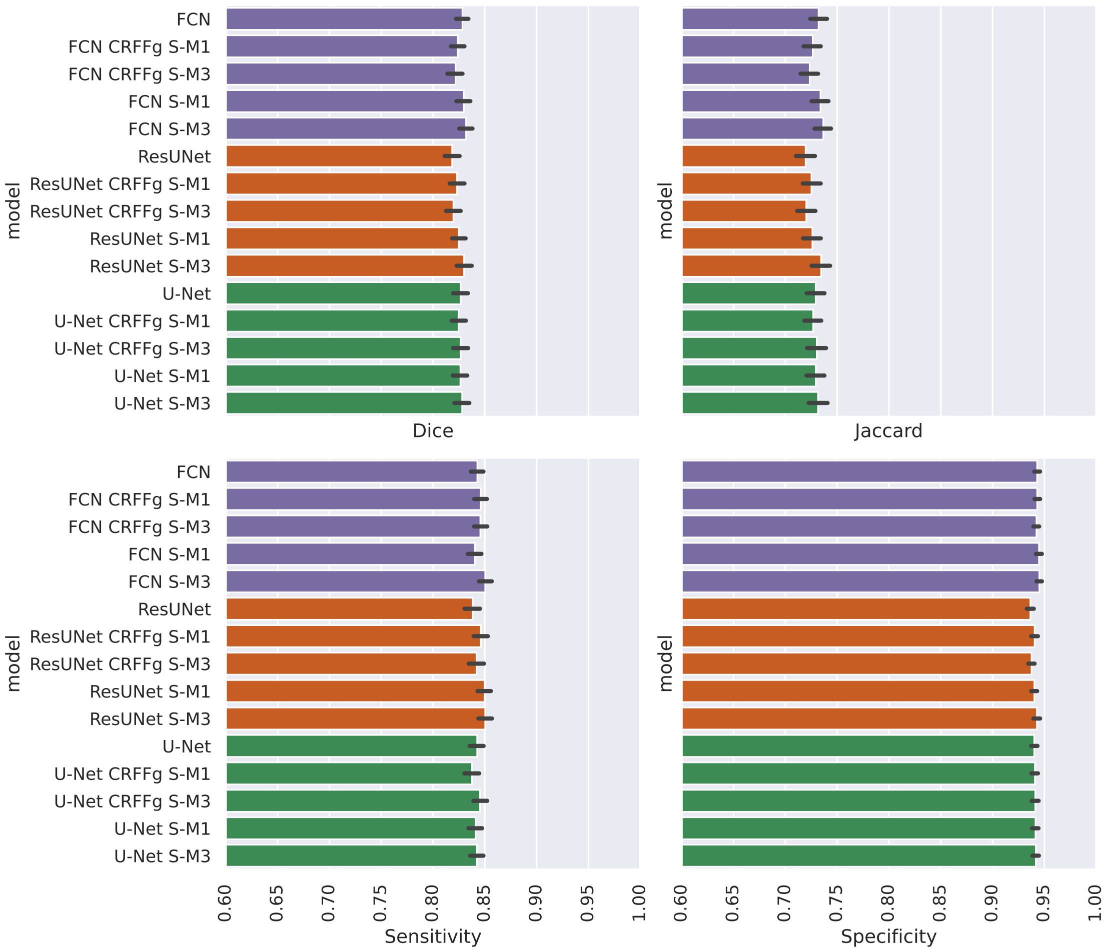


Figure **3-17**. Performance Metrics Results on Oxford-IIIT Pet. The three types of architecture used in this study (FCN, U-Net, ResUNet) are differentiated by color.

For further experimentation, we present Figures 3-18, 3-19, and 3-20, which depict the performance metrics as we vary the size of the training dataset. Upon initial examination, in general, there are no significant disparities between the different models. This can be attributed to the high variability inherent in the dataset, which prevents any single model from gaining a decisive advantage over the others. However, when focusing on the FCN architecture, an interesting trend emerges. It is apparent that models incorporating CRFFg face challenges when provided with less than 50% of the training data. This phenomenon may be attributed to the influence of the CRFFg activation function, particularly in the latter stages of the model. With a reduced amount of training data, the remaining layers struggle to adapt to the sinusoidal activation pattern, leading to a decline in overall performance. Moreover, future improvement can be training the scale parameter σ and using a full encoder-decoder architecture with CRFFg.

3.5 Summary

Our research introduced the CRFFg layer to enhance the representation of low-level semantic features in popular encoder-decoder architectures for semantic segmentation. Our study involved evaluating the effectiveness of CRFFg on two datasets: ThermalFeet (a smaller dataset) and Oxford Iiit Pet (a larger dataset). We specifically focused on three well-known architectures - FCN, U-Net, and ResNet - to test the impact of CRFFg on skip connections. Our findings indicated first that data augmentation has a big impact on increasing the performance of the models. Second, the CRFFg layer had a more pronounced effect on improving performance in the ThermalFeet dataset without data augmentation. Finally, regarding the Oxford Iiit Pet, we showed that our approach is competitive with the standard approach.

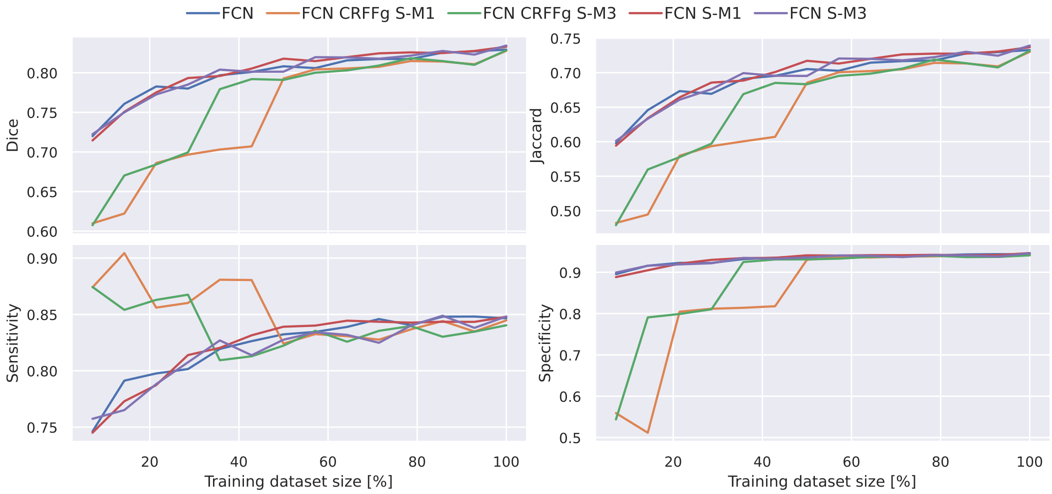


Figure 3-18. Performance metrics at different training dataset sizes of FCN models.

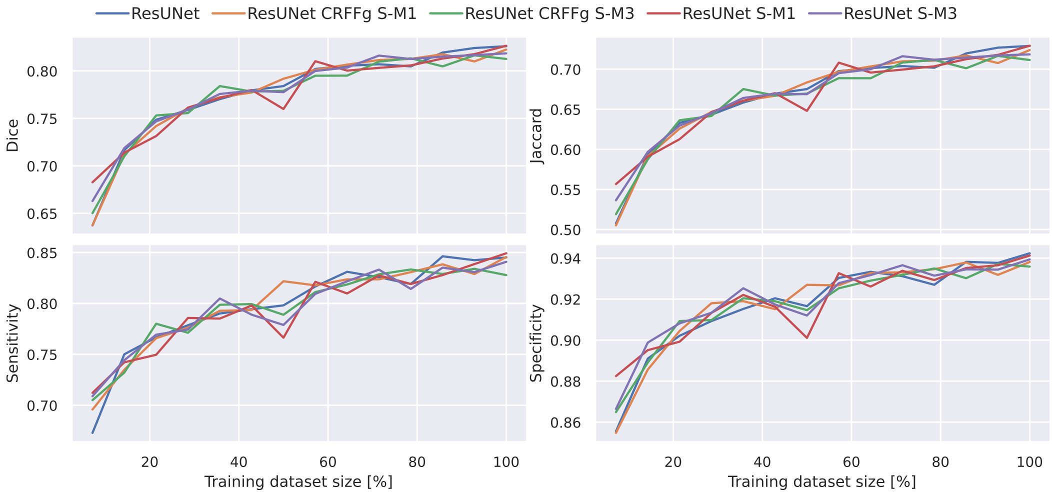


Figure 3-19. Performance metrics at different training dataset sizes of ResUNet models.

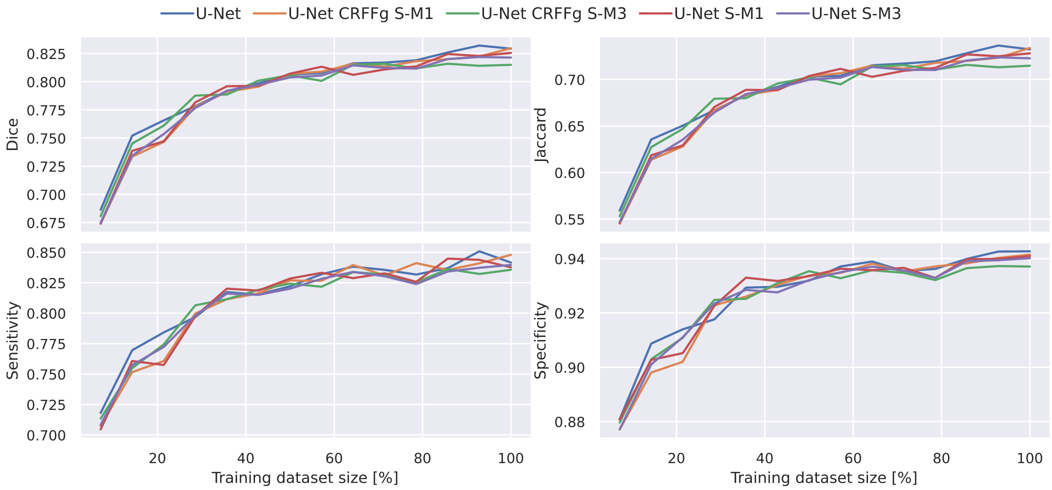


Figure 3-20. Performance metrics at different training dataset sizes of U-Net models.

CHAPTER

FOUR

INTERPRETABILITY OF DEEP LEARNING SEMANTIC SEGMENTATION MODELS

This chapter presents comprehensive semantic segmentation interpretability measures designed specifically for semantic segmentation models based on Convolutional neural networks. The proposed measures offer a systematic approach to understanding the decision-making process of these complex models.

4.1 Semantic Segmentation Interpretability

We first describe Class Activation Maps (CAMs), a method that allows the interpretability of Convolutional Neural Networks. To further deepen our understanding, we also introduce a measurements that provides additional insight into the workings of CAMs in semantic segmentation models based on convolutional networks.

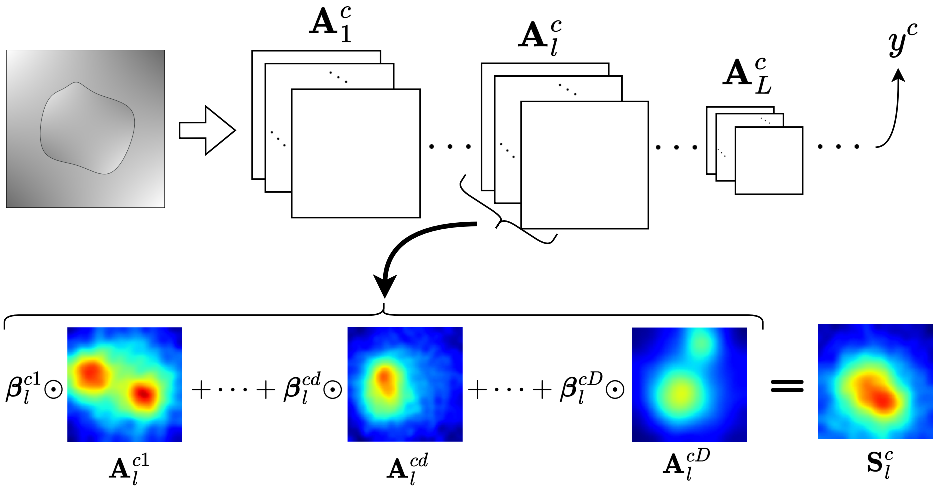


Figure **4-1**. Class Activation Maps Representation.

4.1.1 Class Activation Maps Class Activation Maps (CAMs)

CAMs [Zhou et al., 2015b] is an interpretability technique for highlighting the image regions that contribute the most to the predicted output model given a input. The intuition behind CAMs is to construct the highlight regions for a specific class c using the linear combination of activations or feature maps from the specific layer l of the convolutional neural network as depicted in Equation 4-1 and Figure **4-1**

$$\mathbf{S}_l^c = \Lambda\Big(\sum_{d \in D_l} \boldsymbol{\beta}_l^{cd} \odot \mathbf{A}_l^{cd}\Big) \quad \in \mathbb{R}^{H \times W} \tag{4-1}$$

where $\varLambda : \mathbb{R}^{H_l \times W_l} \to \mathbb{R}^{H \times W}$ is the up-sampling operator, $\mathbf{A}_l^{ck} \in \mathbb{R}^{H_l \times W_l}$ represents the activation map of the l^{th} layer for the d^{th} filter, β_l^{cd} represents the weight matrix associated with the d^{th} filter that depends on $y^c \in \mathbb{R}$, the score returned by the model at class c, and $\odot$ represents the element-wise product operation. Then, the task of CAMs methods is to find the $\boldsymbol{\beta}_l^{cd}$, and that is the difference between the different CAMs methods [Zhou et al., 2015a, Selvaraju et al., 2016, Chattopadhyay et al., 2017, Wang et al., 2019, Jiang et al., 2021].

1. Grad-CAM [Selvaraju et al., 2016]

$$[\boldsymbol{\beta}_l^{cd}]_{i,j} = \frac{1}{H_l W_l} \sum_{n \in H_j, m \in W_l} \frac{\partial y^c}{\partial \mathbf{A}_{nm}^{cd}} \ \forall \ i, j$$

2. Grad-CAM ++ [Chattopadhyay et al., 2017]

$$[\boldsymbol{\beta}_l^{cd}]_{i,j} = \sum_{n \in H_j, m \in W_l} \boldsymbol{\alpha}_{nm}^{cd} ReLU(\frac{\partial y^c}{\partial \mathbf{A}_{nm}^{d}}) \ \forall \ i, j$$

$$\boldsymbol{\alpha}_{nm}^{cd} = \frac{\frac{\partial^2 y^c}{(\partial \mathbf{A}_{nm}^d)^2}}{2\frac{\partial^2 y^c}{(\partial \mathbf{A}_{nm}^d)^2} + \sum_a \sum_b \mathbf{A}_{ab}^d \frac{\partial^3 y^c}{(\partial \mathbf{A}_{nm}^d)^3}}$$

3. Score-CAM [Wang et al., 2019]

$$[\boldsymbol{\beta}_l^{cd}]_{i,j} = \frac{\exp(\xi_d^c)}{\sum_n \exp(\xi_n^c)} \ \forall \ i, j$$

$$\xi_d^c = f^c(\mathbf{I} \circ \mathbf{A}^d) - f^c(\mathbf{I})$$

Where f^c is a function that returns y^c score of the model in the class c, and $\mathbf{I}$ is the input of the model. Additionally, Score-CAM applies a $ReLU$ function to the calculated class activation map (CAM) before using an upsampling operator.

4. Layer-CAM [Jiang et al., 2021]

$$[\boldsymbol{\beta}_l^{cd}]_{i,j} = ReLU(\frac{\partial y^c}{\partial \mathbf{A}_{ij}^d})$$

Layer-CAM also applies a $ReLU$ function to the calculated class activation map before using an upsampling operator.

As presented, CAMs methods are constructed to work with classification models. To work in Semantic segmentation models, we follow a similar approach as [Vinogradova et al., 2020], using the Equation 4-2.

$$y^c = \frac{\mathbf{1}^\top(\hat{\mathbf{M}}^c \odot \mathbf{M}^c)\mathbf{1}}{\mathbf{1}^\top \mathbf{M}^c \mathbf{1}} \tag{4-2}$$

Particularly, we are concerned with spacial information, and for that reason, we choose Layer-CAM [Jiang et al., 2021], which allows for preserving more spacial information since it does not use pooling over the derivatives.

4.1.2 Interpretability Measures Proposal

While CAMs can identify the most discriminatory regions in an input medical image, it was originally designed for classification models and does not utilize the location information available in the mask segmentation. However, this information can be valuable in understanding how CAMs operate in regions of interest and can provide insights for medical image segmentation tasks.

For instance, in medical image segmentation, it is crucial to identify the relevant structures of interest accurately. Therefore, measuring how the CAMs is distributed within the ROI can help determine whether the areas of high discriminative value are located inside or outside the structure of interest. This can be useful in identifying false positives or false negatives and can guide improvements in the segmentation model. This is demonstrated in Figure 4-2.

Additionally, knowing which class the models tend to have a high activation of relevance can be crucial in medical image segmentation. This is because medical image segmentation models often deal with multiple classes, such as different types of tissue or organs. Understanding which classes the model is biased

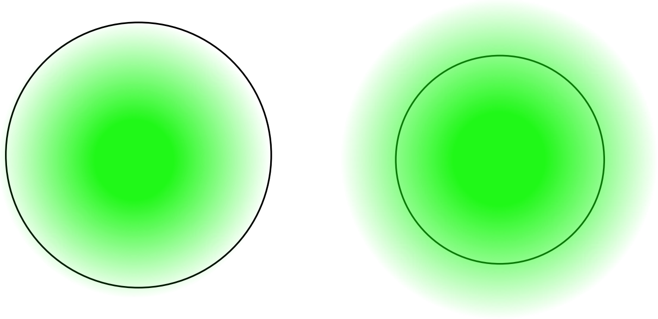


Figure **4-2**. CAMs exemplification of behavior desirable to capture, distribution of the relevance.

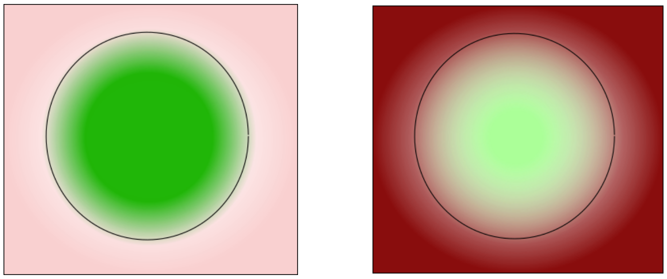


Figure **4-3**. CAMs exemplification of behavior desirable to capture, more activation in one class than the other.

towards can help identify potential limitations or biases in the model and guide further improvements. Figure 4-3 shows one example.

Finally, evaluating the overall homogeneity of CAMs across the entire ROI can provide an understanding of how consistently the model assigns relevance to different regions within the ROI. This can be useful in assessing the robustness and reliability of the segmentation model. For example, suppose the CAMs are highly concentrated in certain regions and have very low activations in others. In that case, it may suggest that the model is not generalizing well to variations in the input data. This information can guide improvements to the model architecture or training data. This situation is exemplified in Figure 4-4.

For those reasons, we proposed the following measures to quantify above mentioned concerns. As previously mentioned, using CAM-based representations enhances the explainability of deep learning models for segmentation tasks. To

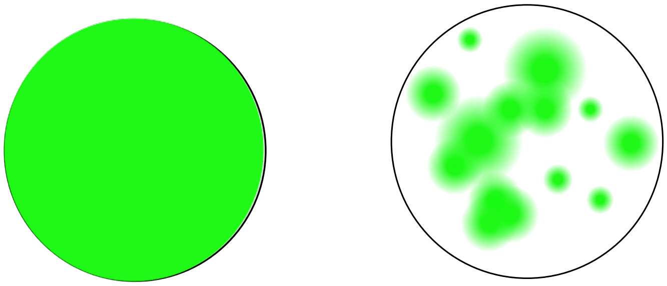


Figure **4-4**. CAMs exemplification of behavior desirable to capture, homogeneity of the relevance through all the ROI.

evaluate the interpretability of CAMs for a given model, we propose the following semantic segmentation measures, where higher scores indicate better interpretability:

- **CAM-based Cumulative Relevance** (ρ_c): It involves computing the cumulative contribution from each CAM representation to detect class c within the segmented region of interest. This can be expressed as follows:

$$\rho_r = \mathbb{E}_l\left\{\mathbb{E}_n\left\{\frac{\mathbf{1}^\top(\tilde{\boldsymbol{M}}_n^c \odot \boldsymbol{S}_{nl}^c)\mathbf{1}}{\mathbf{1}^\top\boldsymbol{S}_{nl}^c\mathbf{1}} : \forall n \in N\right\}\forall l \in L\right\}, \quad \rho_c \in [0,1] \tag{4-3}$$

$$\approx \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N}\sum_{n=1}^{N}\frac{\mathbf{1}^\top(\tilde{\boldsymbol{M}}_n^c \odot \boldsymbol{S}_{nl}^c)\mathbf{1}}{\mathbf{1}^\top\boldsymbol{S}_{nl}^c\mathbf{1}}, \quad \rho_c \in [0,1] \tag{4-4}$$

where $\boldsymbol{S}_{nl}^c$ holds the Layer-CAM for image n with respect to layer l (see Eq. 4-1). Additionally, $\tilde{\boldsymbol{M}}_n^c \in \{0,1\}^{H\times\tilde{W}}$ collects a binary mask that identifies the pixel locations associated with the class c

- **Mask-based Cumulative Relevance** (ϱ_c): It assesses the relevance averaged across the class pixel set related to the target mask of interest. Then, each class-based cumulative relevance is computed as follows:

$$\varrho_{cl}^{'} = \mathbb{E}_n\left\{\frac{\mathbf{1}^\top(\tilde{\boldsymbol{M}}_n^c \odot \boldsymbol{S}_{nl}^c)\mathbf{1}}{\mathbf{1}^\top\tilde{\boldsymbol{M}}_n^c\mathbf{1}} : \forall n \in N\right\}, \quad \varrho_c^{'} \in \mathbb{R}^+ \tag{4-5}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N}\frac{\mathbf{1}^\top(\tilde{\boldsymbol{M}}_n^c \odot \boldsymbol{S}_{nl}^c)\mathbf{1}}{\mathbf{1}^\top\tilde{\boldsymbol{M}}_n^c\mathbf{1}}, \quad \varrho_c^{'} \in \mathbb{R}^+ \tag{4-6}$$

The normalized Mask-based Cumulative Relevance can be computed as:

$$\varrho_c = \mathbb{E}_l\left\{\frac{{\varrho'}_{cl}}{\max\limits_{r\in\{0,1\}} \varrho_{rl}^{'}} : \forall l \in L\right\}, \quad \varrho_c \in [0,1] \tag{4-7}$$

$$\approx \frac{1}{L}\sum_{l=1}^{L}\frac{{\varrho'}_{cl}}{\max\limits_{r\in\{0,1\}} \varrho_{rl}^{'}}, \quad \varrho_c \in [0,1] \tag{4-8}$$

- **CAM-Dice** ($D^{'}$): a version of the Dice measure that quantifies mask thickness and how the extracted CAM is densely filled:

$$D_c^{'} = \mathbb{E}_l\left\{\mathbb{E}_n\left\{2\frac{\mathbf{1}^\top\left(\tilde{\boldsymbol{M}}_n^c \odot \boldsymbol{S}_{nl}^c\right)\mathbf{1}}{\mathbf{1}^\top\tilde{\boldsymbol{M}}_n^c\mathbf{1} + \mathbf{1}^\top\boldsymbol{S}_{nl}^c\mathbf{1}} : \forall n \in N\right\} : \forall l \in L\right\}, \quad D_c^{'} \in [0,1], \tag{4-9}$$

$$\approx \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N}\sum_{n=1}^{N}2\frac{\mathbf{1}^\top\left(\tilde{\boldsymbol{M}}_n^c \odot \boldsymbol{S}_{nl}^c\right)\mathbf{1}}{\mathbf{1}^\top\tilde{\boldsymbol{M}}_n^c\mathbf{1} + \mathbf{1}^\top\boldsymbol{S}_{nl}^c\mathbf{1}}, \quad D_c^{'} \in [0,1] \tag{4-10}$$

4.2 Experimental Set-Up

To demonstrate the practical application of the described metrics, we utilized the FCN, U-Net, and ResUNet models previously trained in Chapter 3 on the

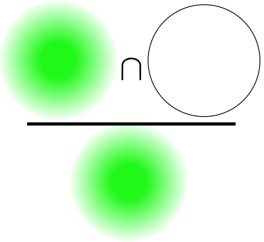


Figure **4-5**. Graphic depiction of Measure ρ_c.

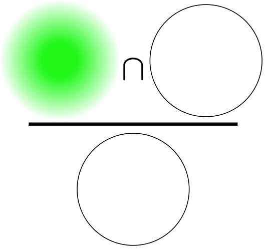


Figure **4-6**. Graphic depiction of Measure ϱ_c'.

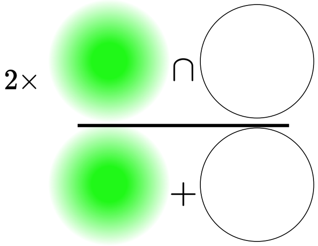


Figure **4-7**. Graphic depiction of Measure D_c'.

ThermalFeet partition and Oxford datasets. We focused solely on calculating the CAMs for the output blocks illustrated in Figures 3-3, 3-4, and 3-5. Since the metrics rely on a set of images, we used the training partition without any modification or data augmentation.

4.3 Results and Discussion

We aim to evaluate the effectiveness of CAM-based representations in enhancing interpretability of tested DL models. To achieve this goal, we analyze the relationship between essential explanation elements, i.e., background and foreground, and the metrics proposed for assessing the CAM-based relevance of image segmentation masks in ThermalFeet and Oxford IIIt Pet datasets.

4.3.1 ThermalFeet

Figure 4-8 displays the scatter plots obtained by each segmentation learner. CAMs extracted by the learner contribute more to the interpretability of regions of interest if the measure value tends toward the top-right corner. Moreover, we focus on the contribution of CAM representations to segmenting between background and foreground, utilizing the patient's feet as critical identification features. The findings from the modified CAM-Dice results can be split into two groups (refer to the left plot in Figure 4-8). One group involves ResUnet and U-Net architectures, and the other showcases better performance, featuring FCN architectures. It is also important to mention that the data augmentation strategy does not significantly boost interpretability as much as it enhances segmentation performance measures. Looking at the CAM-based Cumulative Relevance (refer to the middle plot in Figure 4-8), it is apparent that models with refined representations at skip connections surpass the baseline models. Even though there is no substantial difference between models with these enhancements, most models are situated in the top-right corner. This position suggests that the primary relevance is focused on the area of interest. Significantly, relevance seems to accumulate more in the background than in the foreground, which is logical, considering the relative sizes of both areas. In Figure 4-8, the Mask-based Cumulative Relevance plot on the right side demonstrates that most models tend to exhibit high-foreground-low-background relevance. This pattern leads to a bias favoring the foreground class, as reflected in the more robust activation of CAMs for the foreground class. However, it is interesting that models employing CRFFg perform better in separating classes situated towards the top-right corner, suggesting superior capabilities in differentiating foreground and background classes, as we can see in the mean performance in each metric.

Figures 4-9, 4-10, and 4-11 display examples of CAMs extracted by the best models per architecture under the metric ϱ_c for feet (colored in green) and background (red color), respectively. As seen, the higher weight is located at the last part of the decoder, where the higher values of semantic information are found. Besides, the weights for the background class are also less than for the foreground class, showing that the models emphasize the latter while preserving the relevance weights for the former.

In particular, FCN CRFFg S-M3 is the best FCN architecture, as shown in Figure 4-9, and extracts most of the weights in three layers (i.e., l3, l4, and l5), meaning that other layers do not contribute to the class foreground. On the other hand, this

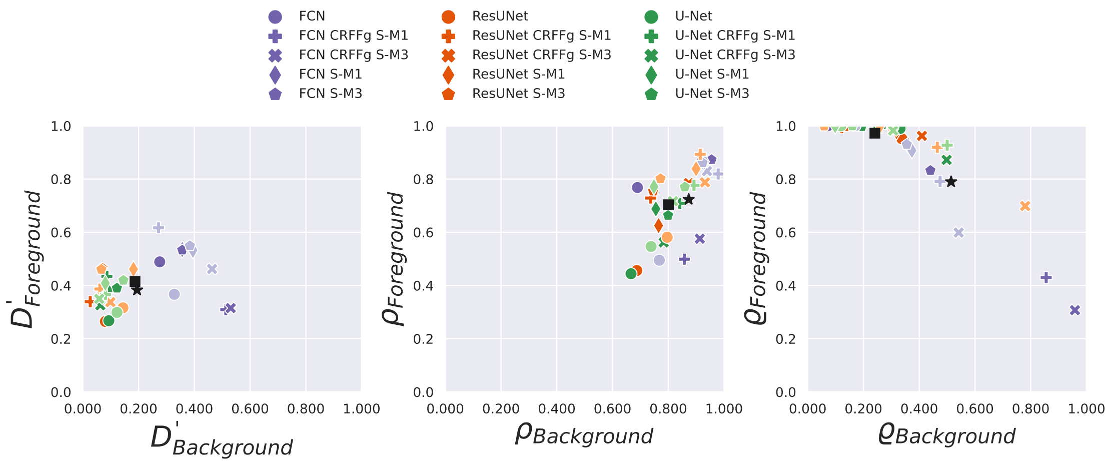


Figure **4-8**. Results of interpretability measures on ThermalFeet. The black markers symbolize the mean of CRFFg model, the start symbol, and the mean of models without CRFFg, the square symbol.

architecture leads to CAMs with lower values for background class (see examples on the right). This behavior can be explained because the FCN architecture holds an extensive receptive field. Hence, the FCN CRFFg S-M3 model enables capturing more global information crucial for segmentation and concentrating weights in a few layers.

In the case of ResUNet, ResUNet CRFFg S-M3 performs the most efficiently, as shown in Figure 4-11. Since the receptive field decreases, the ResUNet architecture distributes the contribution more evenly among the extracted CAM representations. However, the more significant values remain in the l3, l4, and l5 layers. There is also activation of weights for the background class that can be explained, firstly, since the CRFFg configuration helps capture complex non-linear dependencies. Secondly, the local receptive field allows class separation.

Lastly, the CRFFg S-M3 model is the most effective for the U-Net architecture, with a performance similar to the outperforming ResUNet architecture, as shown in Figure 4-10. However, several differences in the Fusion CAMs extracted by U-Net CRFFg S-M3 show high activation within the feet, suggesting that this model is not only sensitive to the foreground class. Also, it captures more global features from feet.

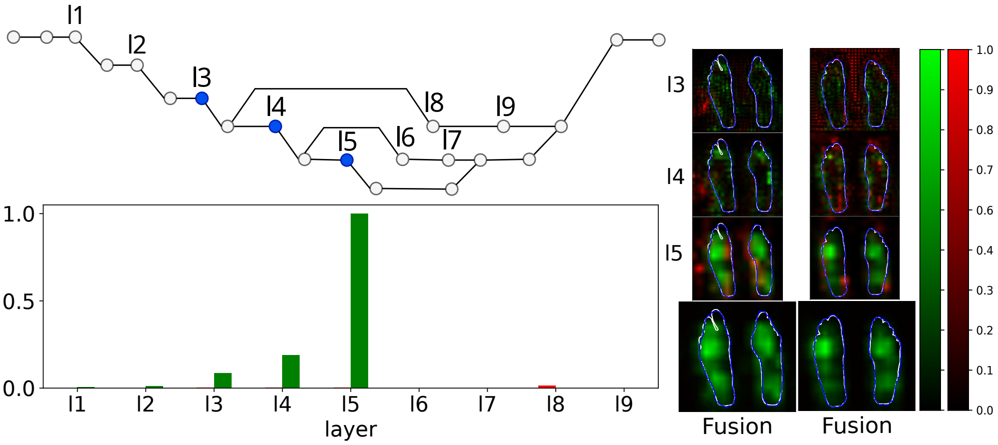


Figure **4-9**. FCN CRFFg S-M3 without data Augmentation.

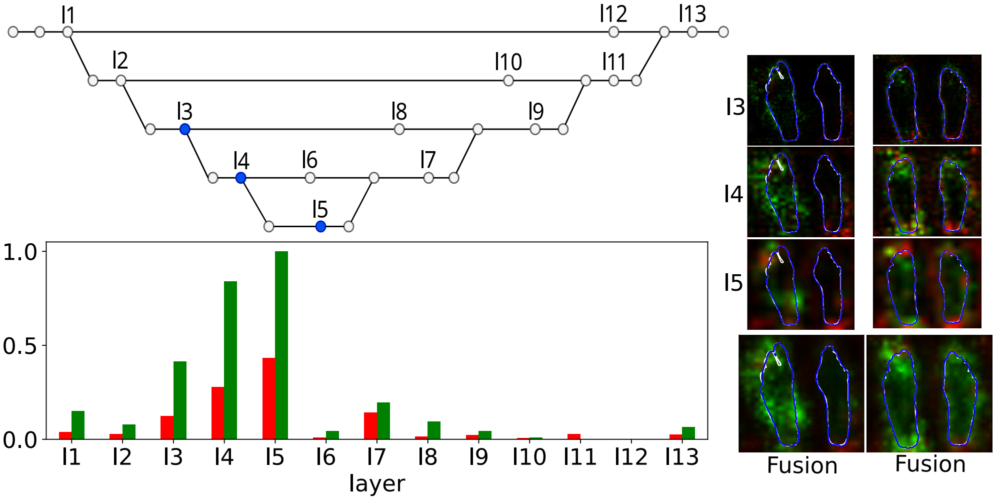


Figure **4-10**. ResUNet CRFFg S-M3 without data Augmentation.

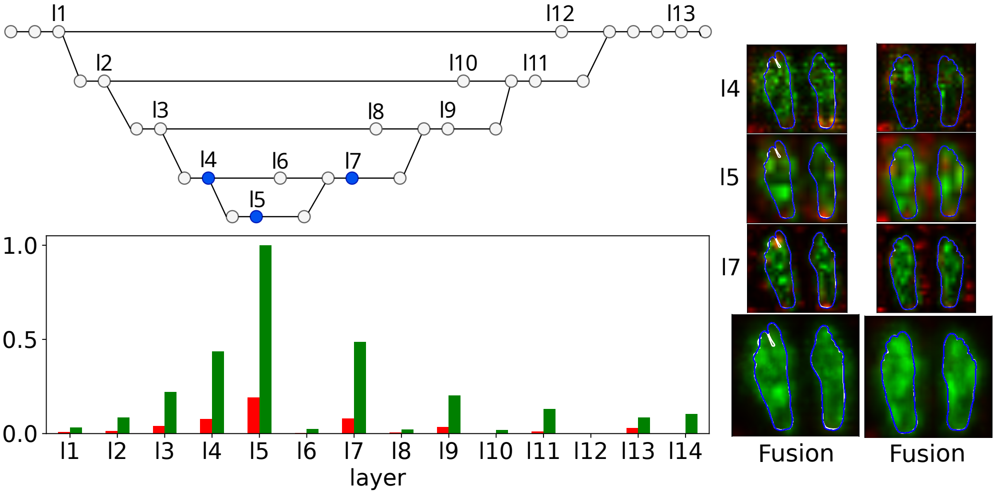


Figure **4-11**. U-Net CRFFg S-M3 with data Augmentation.

4.3.2 Oxford-IIIt Pet

In a similar manner, we present the results obtained from the Oxford-IIIT Pet dataset. Figure 4-12 illustrates the metrics described in Section 4.1.2 and their corresponding results. Upon initial observation, two distinct groups of models can be discerned: the U-Net and ResUNet models exhibiting inferior performance, and the FCN models demonstrating relatively better performance. This pattern mirrors the findings from the ThermalFeet dataset, but is more pronounced in this particular case.

The FCN architecture possesses a larger receptive field, enabling it to capture intricate relationships within heterogeneous structures. Consequently, the CAMs produced by the FCN models exhibit greater homogeneity (as depicted in the left plot of Figure 4-12 for CAM-Dice), with relevance concentrated in the regions of interest (as shown in the middle plot of Figure 4-12 for CAM-based Cumulative Relevance). Moreover, in this dataset as well, the FCN architecture demonstrates the ability to consider both the foreground and background regions (as indicated in the right plot of Figure 4-12 for CAM-based Cumulative Relevance). Additionally, we observe a similar trend to that observed in the ThermalFeet dataset, where models exhibit a tendency towards the top-left corner, albeit to a lesser extent (as seen in the right plot of Figure 4-12 for CAM-based Cumulative Relevance). This deviation may be attributed to the larger size of the dataset and the variations within the regions of interest, as it appears that models tend to diminish the bias towards the foreground class.

Furthermore, it is noteworthy that in most cases, the baseline models demonstrate poorer performance, while models with enhanced representation achieve superior results. However, no apparent distinctions can be observed among these enhanced models.

We present the superior models for each architecture in terms of Mask-based Cumulative Relevance (MCR) as depicted in Figures 4-13, 4-15, and 4-14 for FCN,

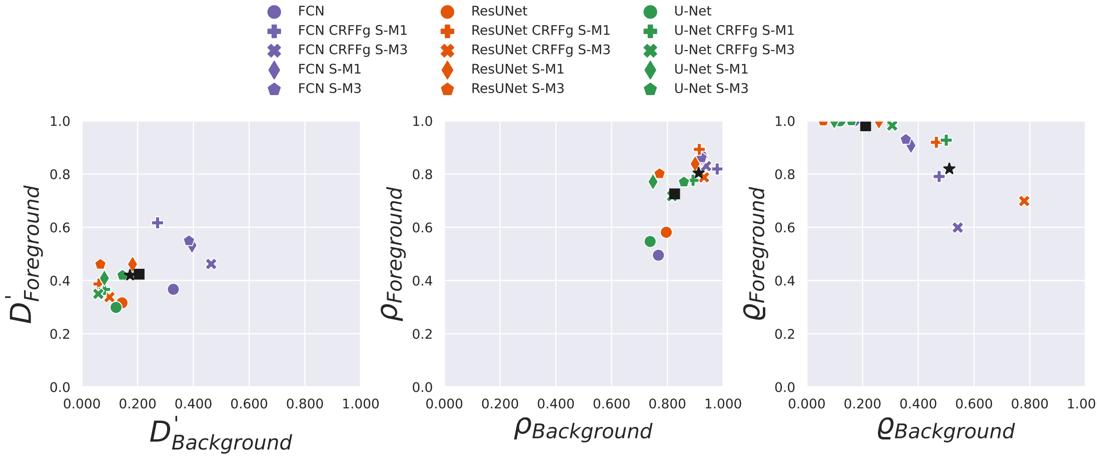


Figure **4-12**. Results of interpretability measures on Oxford IIIt Pet. The black markers symbolize the mean of CRFFg model, the start symbol, and the mean of models without CRFFg, the square symbol.

ResUNet, and U-Net, respectively. These findings are consistent with the observed patterns in the ThermalFeet dataset, where higher weights are concentrated near the top of the encoder, while the background weight is relatively lower compared to the foreground class. Notably, the best models, despite including layers in the skip connection, do not incorporate a CRFFg layer.

Specifically, the most optimal model within the FCN architecture is the S-M1 (Figure 4-13). In contrast to the ThermalFeet scenario, the l3 layer has been removed, and the l6 layer has emerged as one of the top three layers. This architectural shift emphasizes the deeper layers in the encoder, increasing their relevance. Furthermore, it is important to highlight that the l6 layer assigns similar weights to both the background and foreground classes. This is evident from the two examples of CAMs displayed in Figure 4-13, where both classes exhibit comparable levels of relevance. Finally, it is worth noting that CAMs for each class are homogenous and contained within the region of interest, as we can prove in the CAMs examples and the metrics depicted in 4-12.

Regarding the ResUNet architecture, the most optimal model is ResUNet S-M3, which is depicted in Figure 4-13. Interestingly, the weight distribution follows a similar pattern as observed in the ThermalFeet dataset. Most weights are concentrated in layers l3, l4, and l5. It is worth noting, however, that the CAMs) appear to be noisier compared to those shown in the ThermalFeet dataset. Nevertheless, the noise is mainly confined within the region of interest, especially for the foreground class, which tends to be situated at the boundary of regions with high activation, as shown in the CAMs displayed in Figure 4-13.

Lastly, as the best model for U-Net architecture, we have the U-Net S-M1 model. Here, we can observe that layer l5 disappears as one of the top three layers. Instead, the l6 layer takes its position. Moreover, we can appreciate how the CAMs in this scenario are better contained within the region of interest, particularly for the foreground class. Finally, it is worth noting that in all those models compared with the model from ThermalFeet, the CAMs tend to be within the region of interest.

4.4 Summary

We have proposed new measures for interpretability in semantic segmentation models with convolutional neural networks. These measures allow us to measure three different aspects of models using CAMs first, where the relevance for specific interest regions is located through CAM-based Cumulative Relevance. Second, how the sensibility is dispersed through the multiple regions of interest, Mask-based Cumulative Relevance, and finally, how homogenous is relevant in the interest region using CAM-Dice. To prove these proposed measures' consistency and utility, we have analyzed the models trained in Section 3 under the three scenarios of data ThermalFeet, ThermalFeet with data augmentation, and Oxford IIIt Pet. The results show that these metrics allow a better understanding of semantic segmentation models.

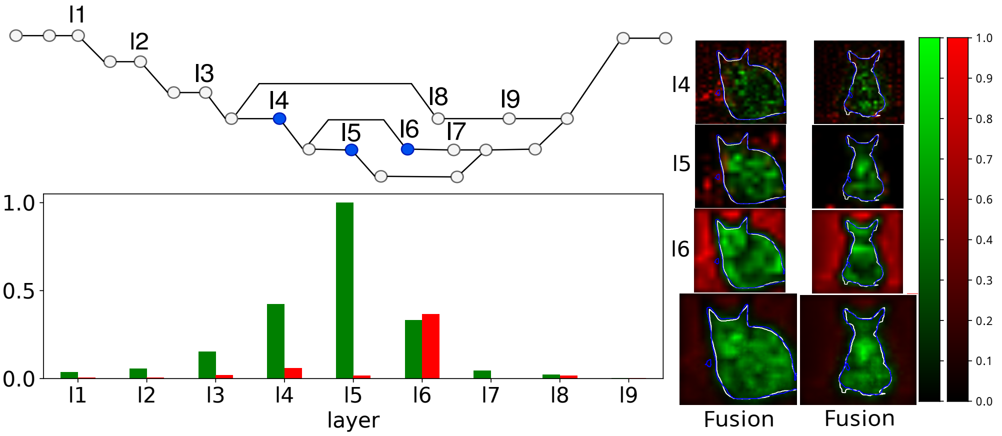


Figure **4-13**. FCN S-M1.

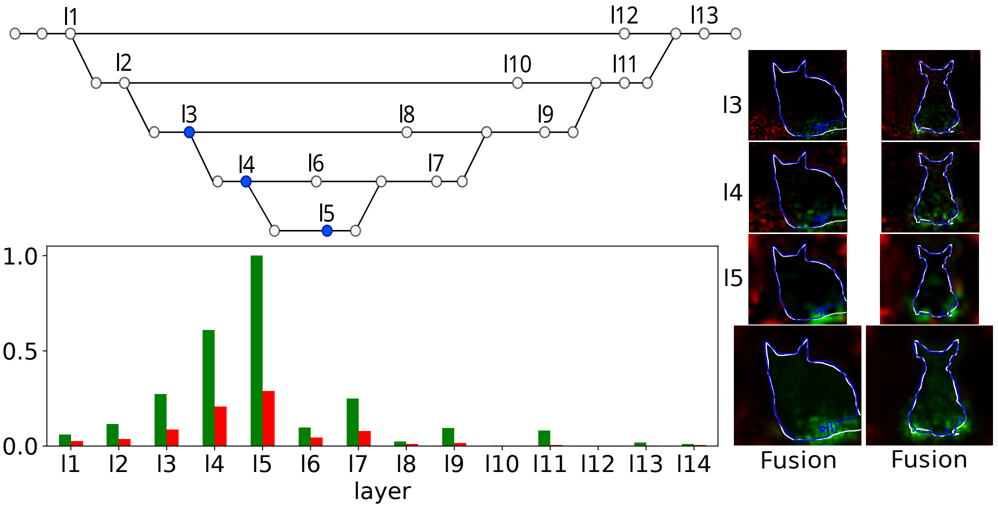


Figure **4-14**. ResUNet S-M3.

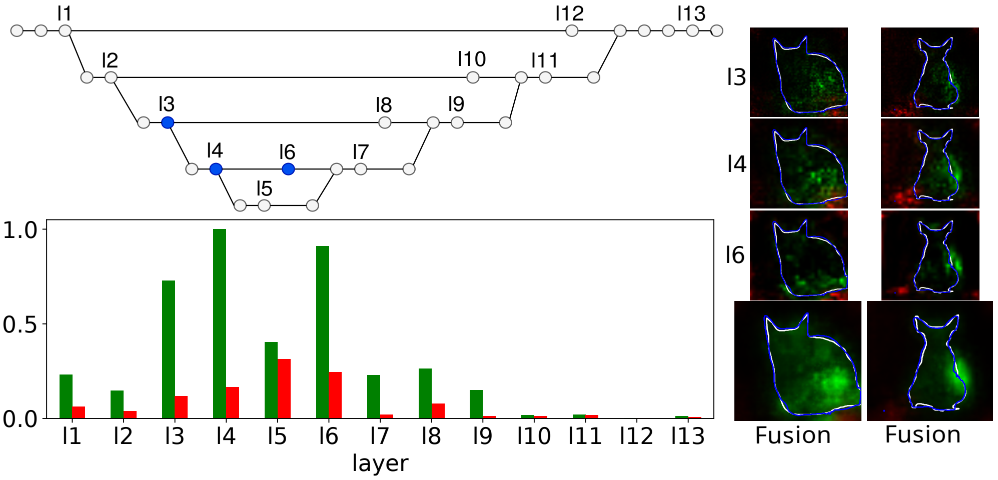


Figure **4-15**. U-Net S-M1.

CHAPTER

FIVE

FINAL REMARKS

5.1 Conclusions

The thesis presented in this work addresses the challenges associated with medical image segmentation, particularly in obstetric environments where data acquisition is limited. Our first contribution is the optimization of Random Fourier Features for spatial data through gradient descent, named CRFFg. This approach incorporates the generalization properties of kernel methods and enhances data efficiency for spatial data derived from convolutions to CRFFg. The implementation of the proposed techniques is available in a GitHub repository, providing a valuable resource for further research and development in this area.

Another key contribution of this thesis is enhancing shallow encoder-decoder models for semantic segmentation. We have recognized the challenges posed by shape, size, and texture variability across patients and imaging protocols. In the context of segmenting feet in thermal images, we introduce a novel approach to improve the representation at the skip connection by incorporating a CRFFg layer. This enhancement aims to capture low-level features more effectively from the encoder and improve their fusion in the decoder, thereby addressing the challenges of ROI variability in semantic segmentation.

Interpretability is a crucial aspect of semantic segmentation models in the medical field, but the complex nature of deep learning models presents challenges in achieving it. The thesis proposes novel quantitative measures to enhance interpretability in semantic segmentation models. These measures assess different aspects of the models, including the location of relevance in specific regions of interest (CAM-based Cumulative Relevance), sensibility across multiple regions of interest (Mask-based Cumulative Relevance), and the homogeneity of relevance in interest regions (CAM-Dice). These measures provide objective and comprehensive evaluations of interpretability, going beyond visual inspection and qualitative analysis.

Overall, this thesis makes significant contributions to the field of medical image segmentation and semantic segmentation. The optimization of Random Fourier Features for spatial data and the enhancement of shallow encoder-decoder models address the challenges specific to the domain, such as limited data availability and ROI variability. Additionally, the proposed quantitative measures for interpretability in semantic segmentation models fill a gap in the existing methods by objectively evaluating model performance in terms of interpretability. The availability of the implementation code in a GitHub repository further facilitates the research community's adoption and further development of these techniques.

5.2 Future Work

To advance our research and build upon our current findings, we have identified several directions for future work. These avenues will enhance our understanding of the topic at hand and contribute to our approach's overall effectiveness and reliability. The following are the next possible continuing paths we envision:

- Analyzing the Spectral Representation of the CRFFg Layer: By examining the spectral characteristics, we can gain valuable insights into the underlying patterns and structures within the layer. This analysis can reveal hidden information and further unravel the inner workings of the CRFFg layer, leading to a deeper understanding of its behavior and potential improvements [Zhang et al., 2020].

- Incorporating Bayesian Approximation: An exciting prospect for improving the understanding of the CRFFg layer lies in incorporating Bayesian approximation techniques. Bayesian methods allow for probabilistic reasoning, enabling us to quantify uncertainties and make more informed decisions. By applying Bayesian approximations to the CRFFg layer, we can better understand its inherent uncertainty, model the relationships between variables more accurately, and potentially uncover new strategies for optimization or interpretation [Miller and Reich, 2022].

- Employing Regularization Techniques: Building upon the measures proposed in Chapter 4, another fruitful path is the utilization of them for regularization purposes. Regularization serves as a means to mitigate overfitting, improve generalization, and control the complexity of the model. By incorporating regularization methods, we can address potential issues such as parameter redundancy, enhance our model's robustness, and achieve better performance on unseen data with a focused desirable behavior of the discriminative regions [Chang et al., 2020, Lin et al., 2021].

- Furthermore, our analysis has revealed several potential enhancements in our approach. Firstly, we could explore alternative mappings for Random Fourier Features, such as combining trigonometric functions like cosines and sines [sut, 2015], the training of the scale parameter $sigma$, and the use of a full encode-decoder architecture with CRFFg layers. Secondly, addressing the issue of class imbalance among pixels may warrant the adoption of different loss functions [Yeung et al., 2022]. Lastly, considering the substitution of convolutions with transformers as a fundamental operation could be advantageous to better capture large-range context [Azad et al., 2023].

In conclusion, the identified avenues for future work are promising in advancing our research. Analyzing the spectral representation of the CRFFg layer will unveil hidden patterns and improve our understanding. Incorporating Bayesian approximation techniques can enhance decision-making and optimization strategies. Employing regularization techniques based on the proposed measures will mitigate overfitting and enhance the model's performance with a focused desirable behavior of the discriminative regions. Pursuing these paths will contribute to our approach's overall effectiveness and reliability, pushing the boundaries of knowledge in this field.

5.3 Academic Products

5.3.1 Academic Discussion

- Aguirre-Arango, J.C.; Álvarez-Meza, A.M.; Castellanos-Dominguez, G. Feet Segmentation for Regional Analgesia Monitoring Using Convolutional RFF and Layer-Wise Weighted CAM Interpretability. Computation 2023, 11, 113. https://doi.org/10.3390/computation11060113

- Mejia-Zuluaga, Rafael, Juan Carlos Aguirre-Arango, Diego Collazos-Huertas, Jessica Daza-Castillo, Néstor Valencia-Marulanda, Mauricio Calderón-Marulanda, Óscar Aguirre-Ospina, Andrés Alvarez-Meza, and Germán Castellanos-Dominguez. ”Deep Learning Semantic Segmentation of Feet Using Infrared Thermal Images.” In Advances in Artificial Intelligence–IBERAMIA 2022: 17th Ibero-American Conference on AI, Cartagena de Indias, Colombia, November 23–25, 2022, Proceedings, pp. 342-352. Cham: Springer International Publishing, 2023.

- A.D. Tobar, J.C. Aguirre, D.A. Cardenas-Pena, A.M. Alvarez-Meza, and C.G. Castellanos-Dominguez, ”Hippocampus Segmentation using Patch-based Representation and ROC Label Enhancement,” Engineering Letters, vol. 31, no. 2, pp504-510, 2023

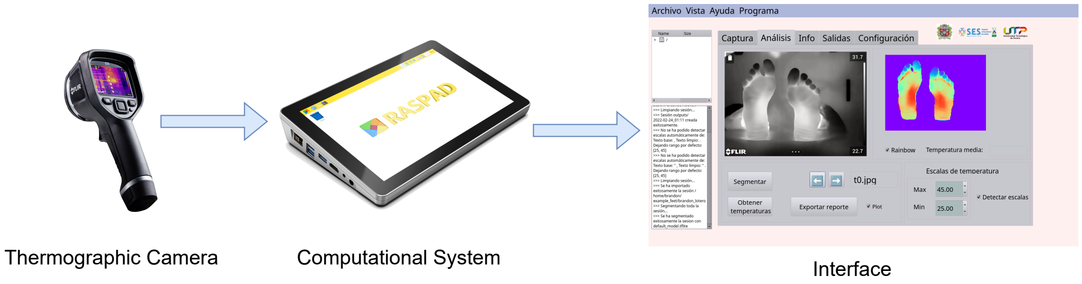
5.3.2 Software and Repositories

- Image Segmentation: Baseline library for semantic segmentation experiments.

- CRFFg: Implementation of the CRFFg layer.

- FEET-GUI: a python application for characterizing epidural anesthesia performance during birth through the segmentation of feet soles and series analysis of temperatures. Figure **5-1** shows an overview of the system, and **5-2** shows a report that the medical personnel can obtain with this tool.

- Software registration Trade-Net: people detection and tracking software accompanied by distance measurements between objects of interest on video sequences in industrial environments (indoor scenarios). Its objective is to monitor personnel in internal environments using the videos captured by security cameras as inputs. This software uses deep learning models to process video in real-time, which can come from surveillance cameras, or already stored videos.



```
>>> Limpiando sesión...
>>> Sesión outputs/
2022-02-24_01:11 creada
exitosamente.
>>> No se ha podido detectar
escalas automáticamente de:
Texto base: , Texto limpio: .
Dejando rango por defecto:
[25, 45]
>>> No se ha podido detectar
escalas automáticamente de:
Texto base: '' , Texto limpio: '' .
Dejando rango por defecto:
[25, 45]
>>> Limpiando sesión...
>>> Se ha importado
exitosamente la sesión /
home/brandon/
example_feet/brandon_lotero
>>> Segmentando toda la
sesión...
>>> Se ha segmentado
exitosamente la sesion con
default_model.tflite
```

☑ Rainbow

☑ Plot

Max 45.00

Min 25.00

☑ Detectar escalas

Figure **5-1**. Illustration of FEET-GUI system.

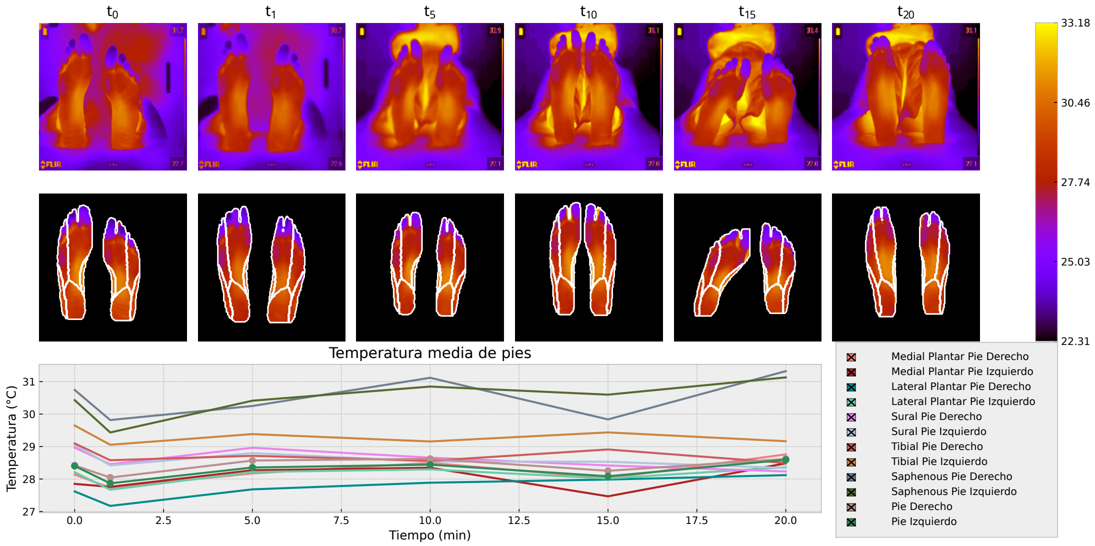


Figure **5-2**. Example of the report generated by the FEET-GUI system.

APPENDIX

ONE

LEARNING CURVES

A.1 Infrared Thermal Images

A.1.1 Without Data Augmentation

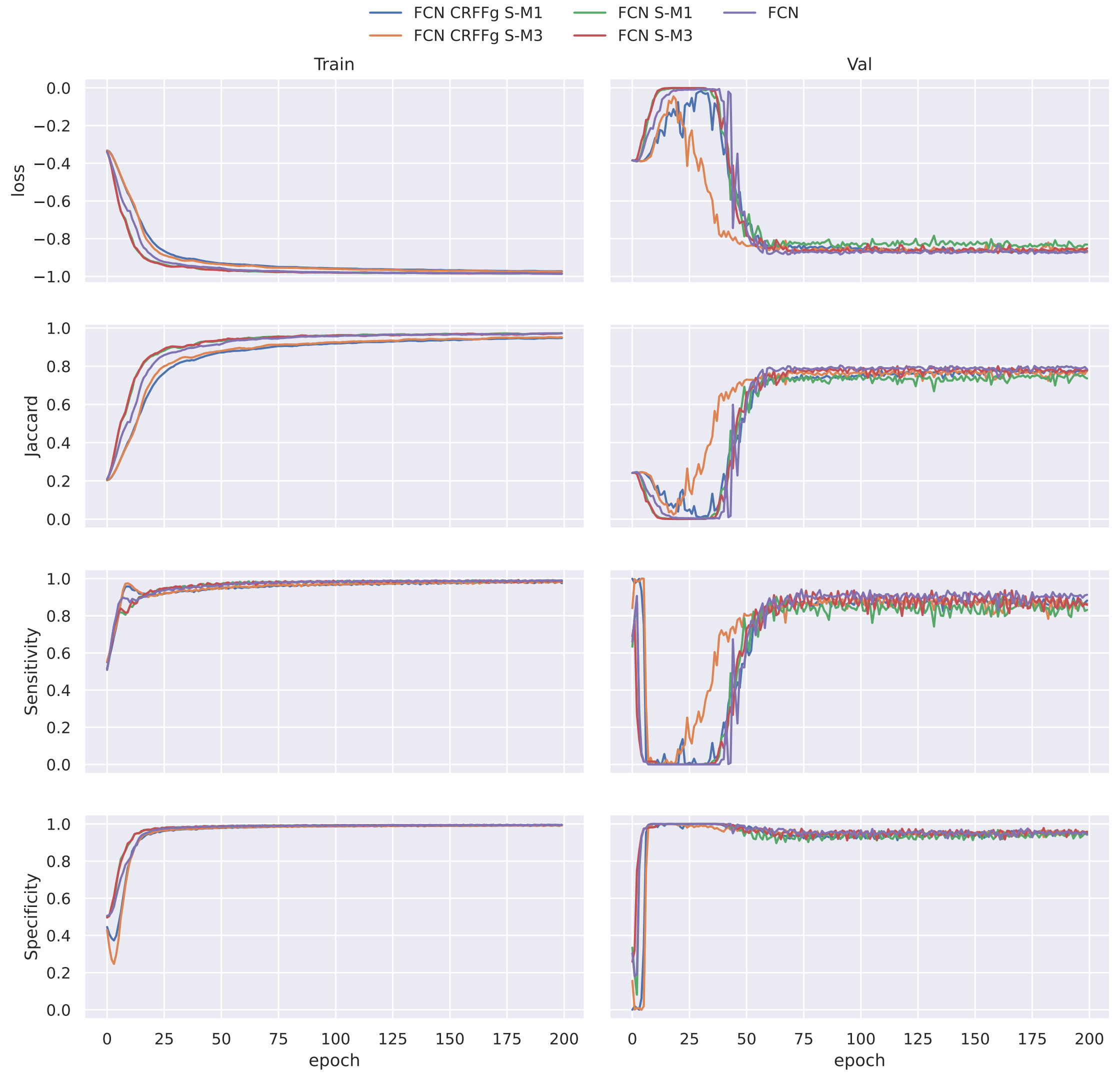


Figure **A-1**. Learning curves of ThermalFeet without data augmentation FCN models

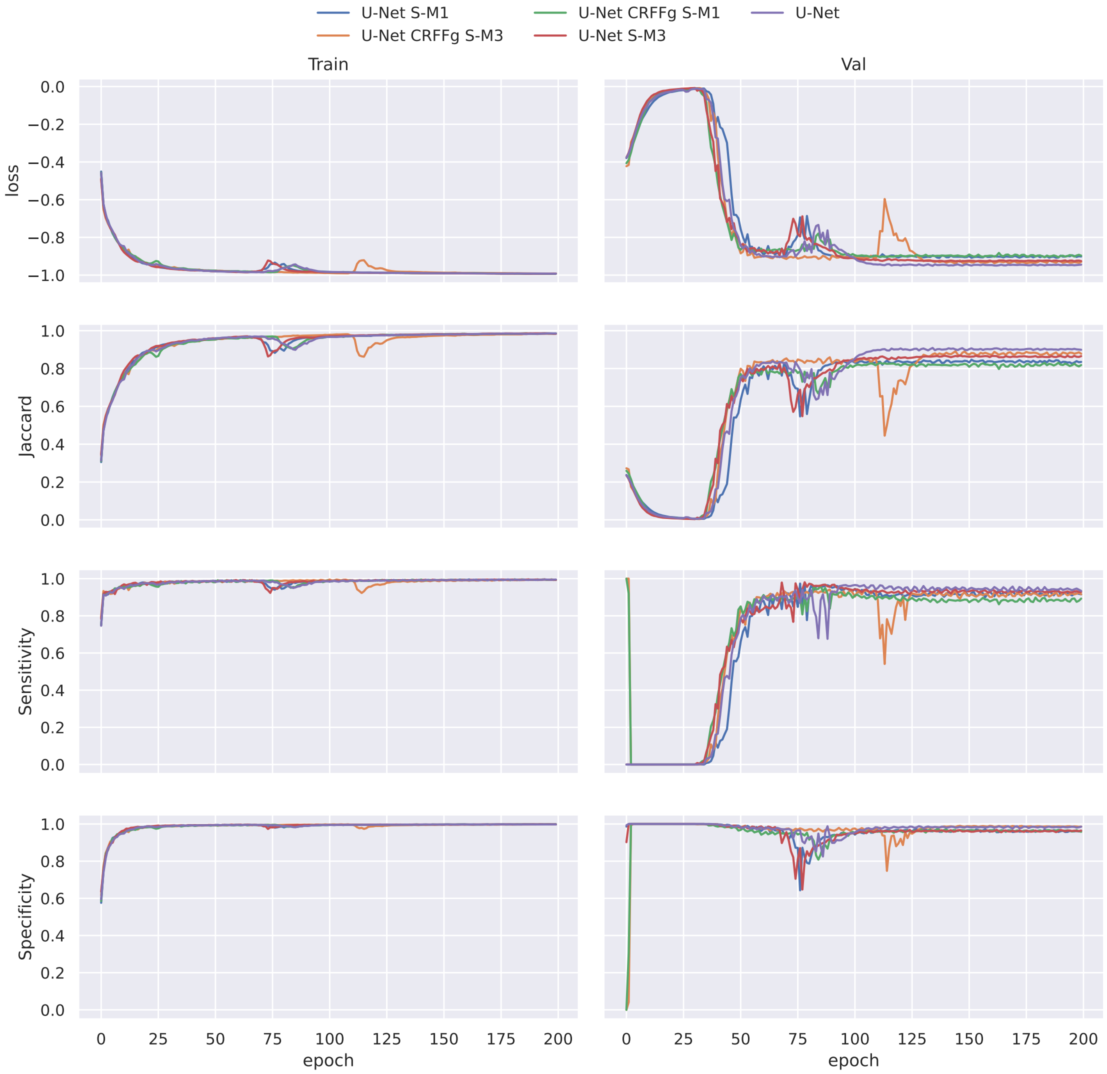


Figure **A-2**. Learning curves of ThermalFeet without data augmentation U-Net models

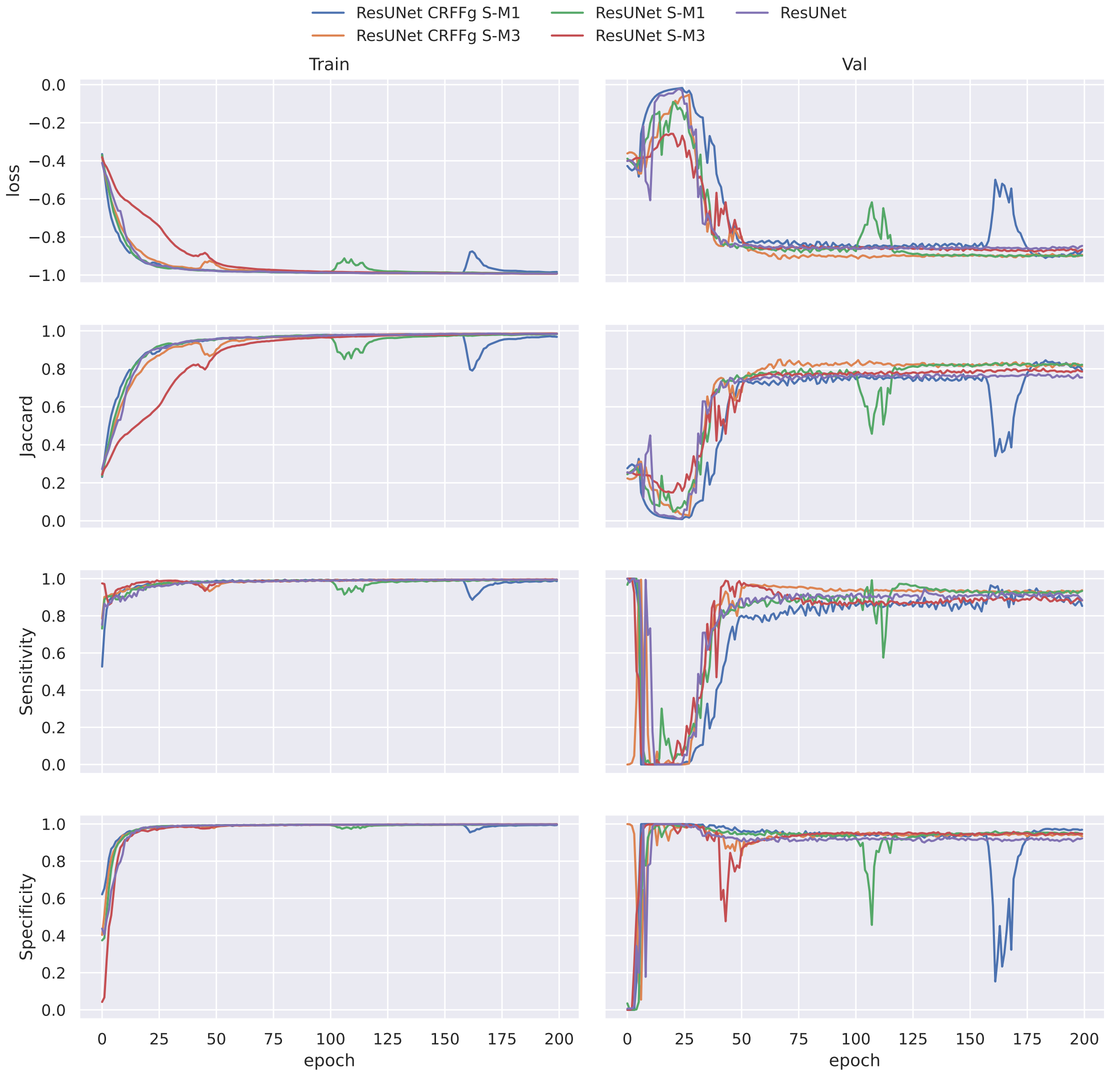


Figure **A-3**. Learning curves of ThermalFeet without data augmentation ResUnet models

A.1.2 With Data Augmentation

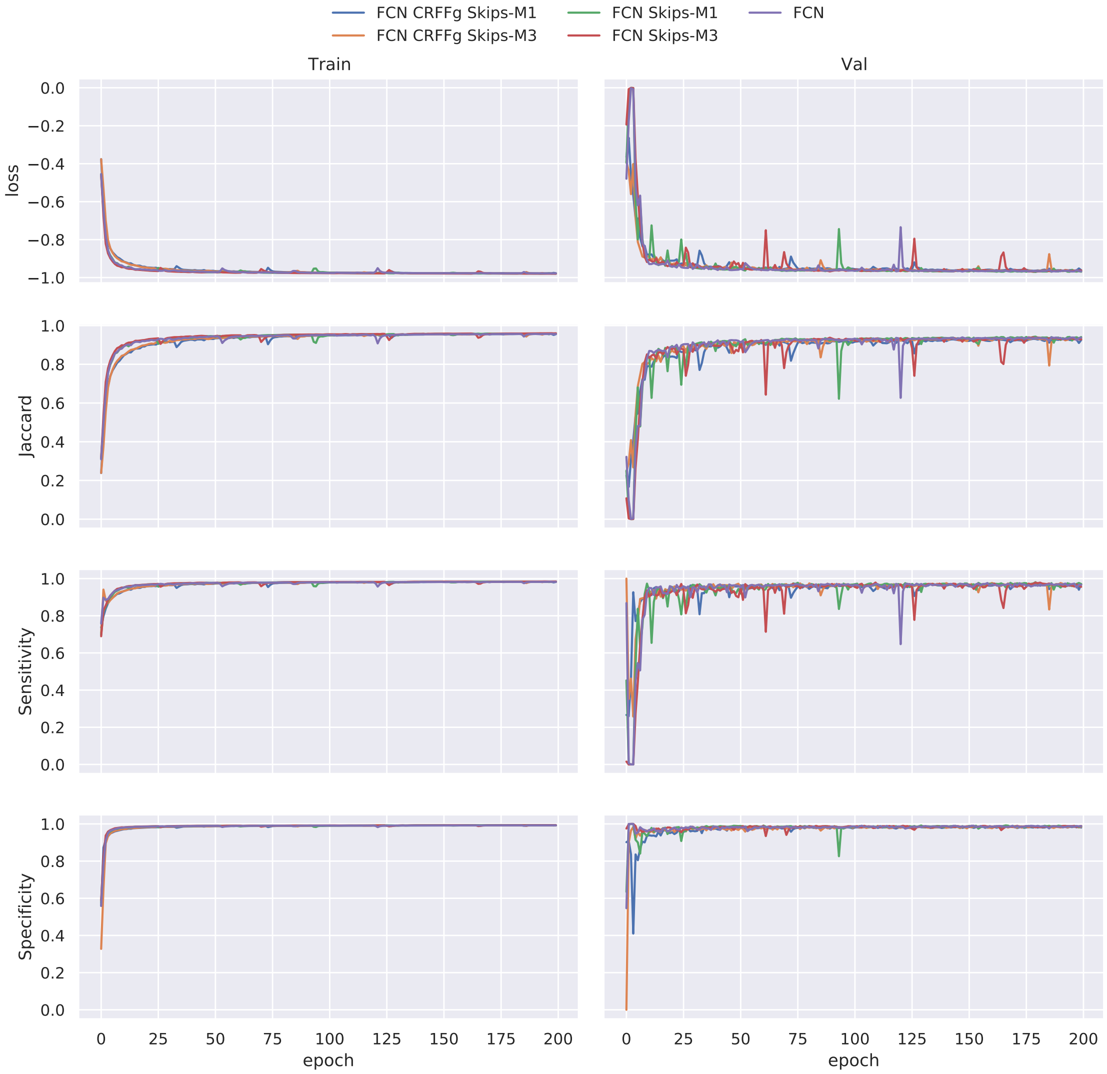


Figure A-4. Learning curves of ThermalFeet with data augmentation FCN models

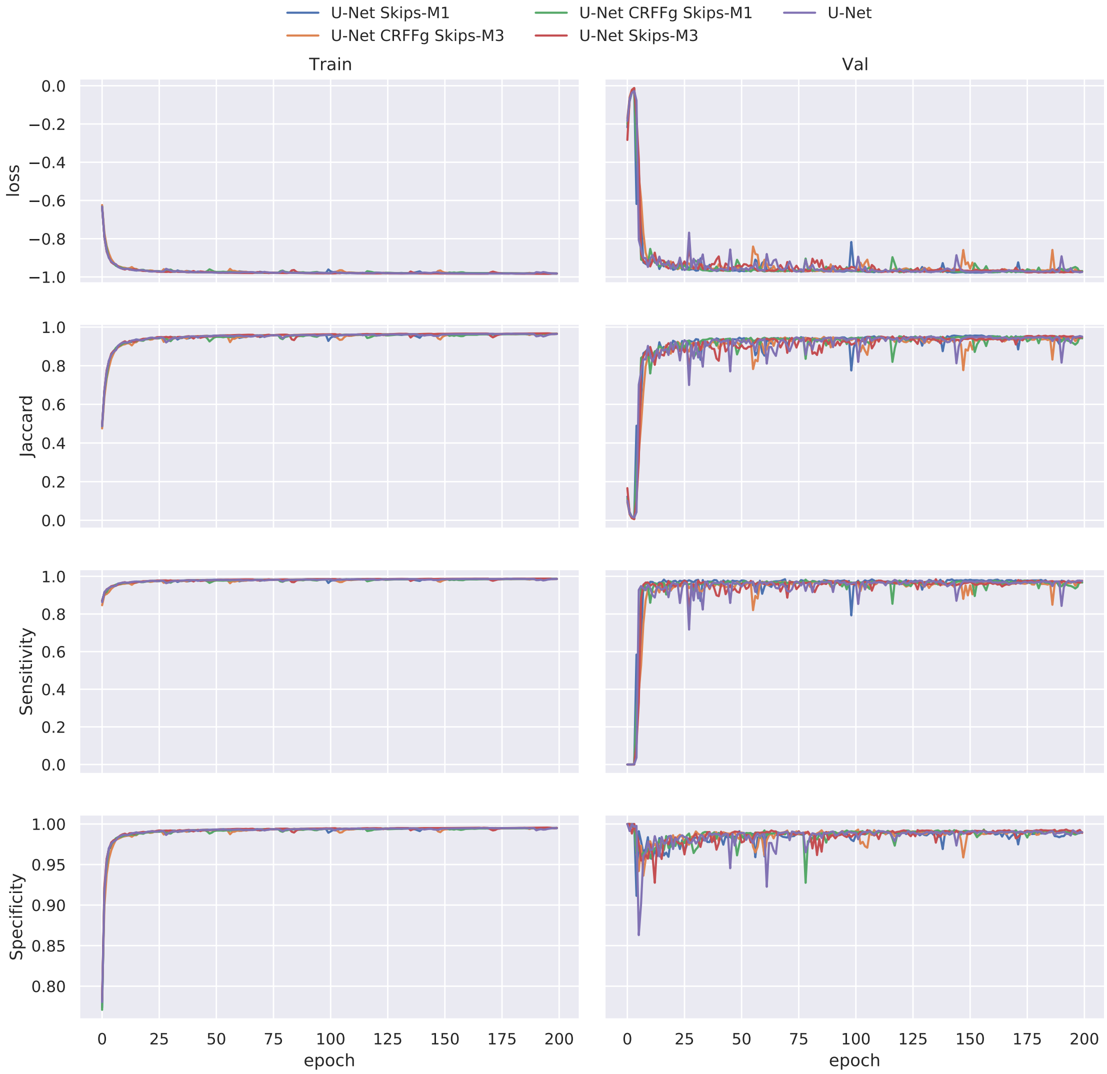


Figure **A-5**. Learning curves of ThermalFeet with data augmentation U-Net models

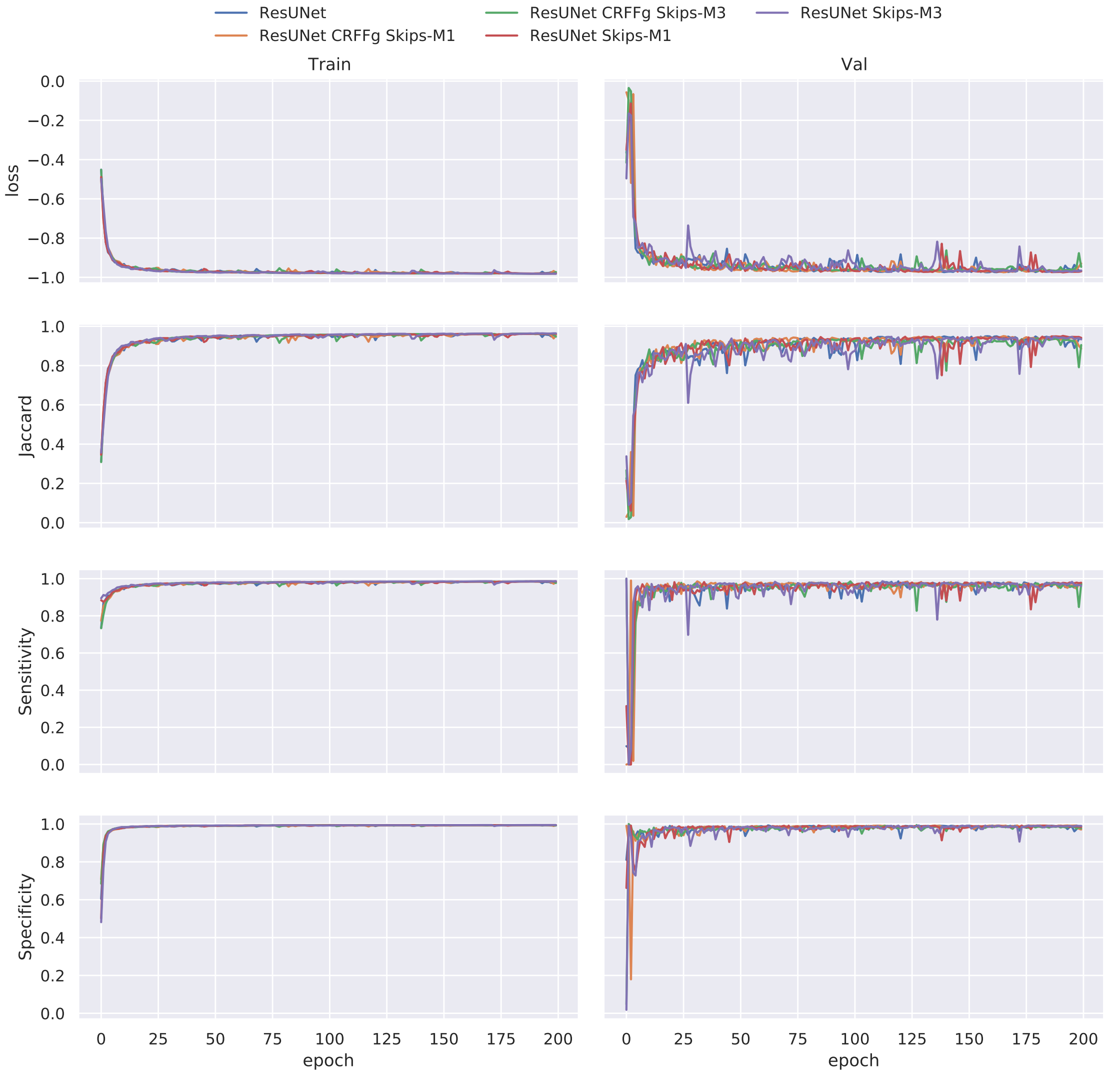


Figure **A-6**. Learning curves of ThermalFeet with data augmentation ResUNet models

A.2 Oxford IIIt Pet

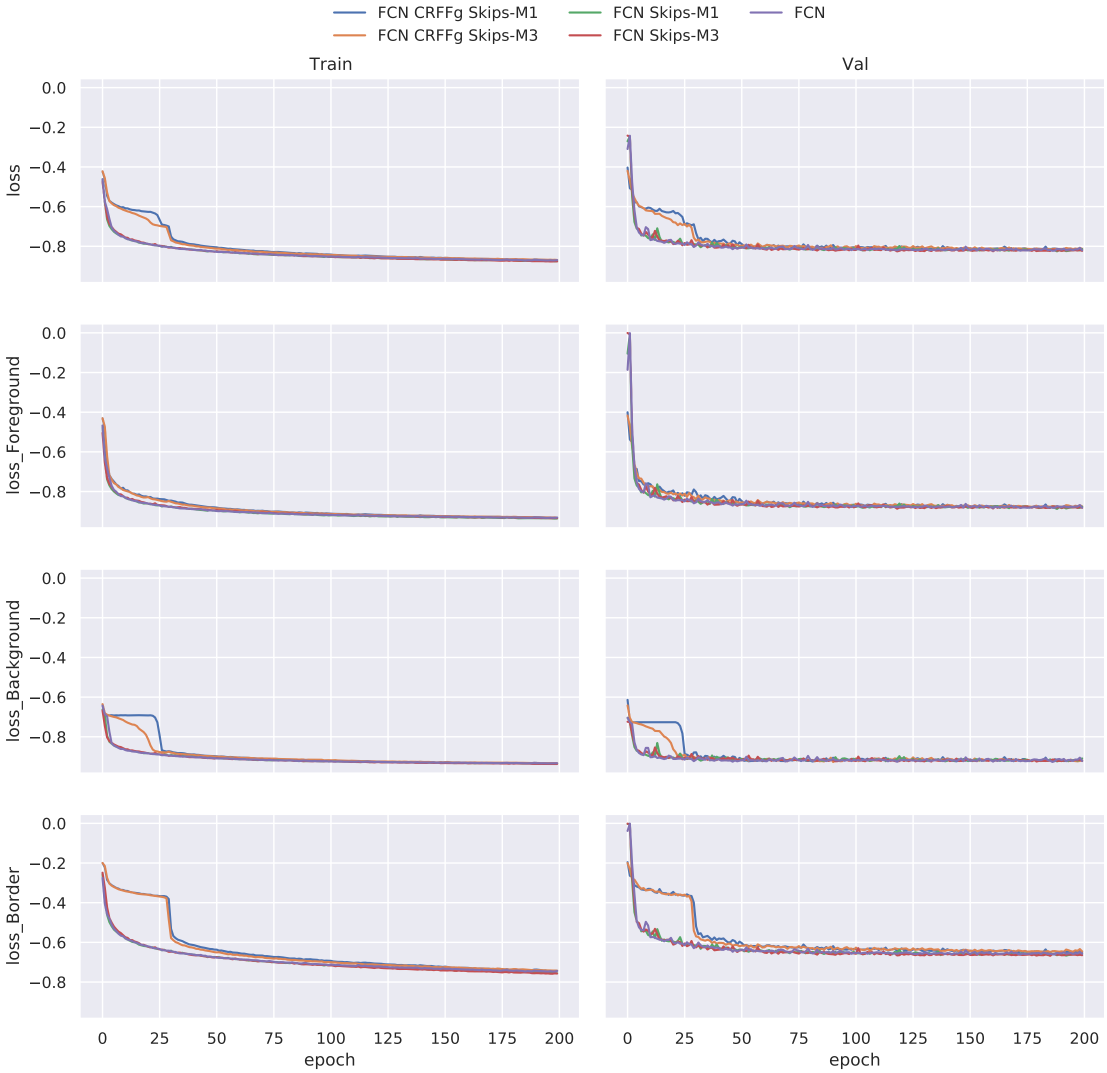


Figure **A-7**. Train Loss Oxford IIIt Pet FCN models

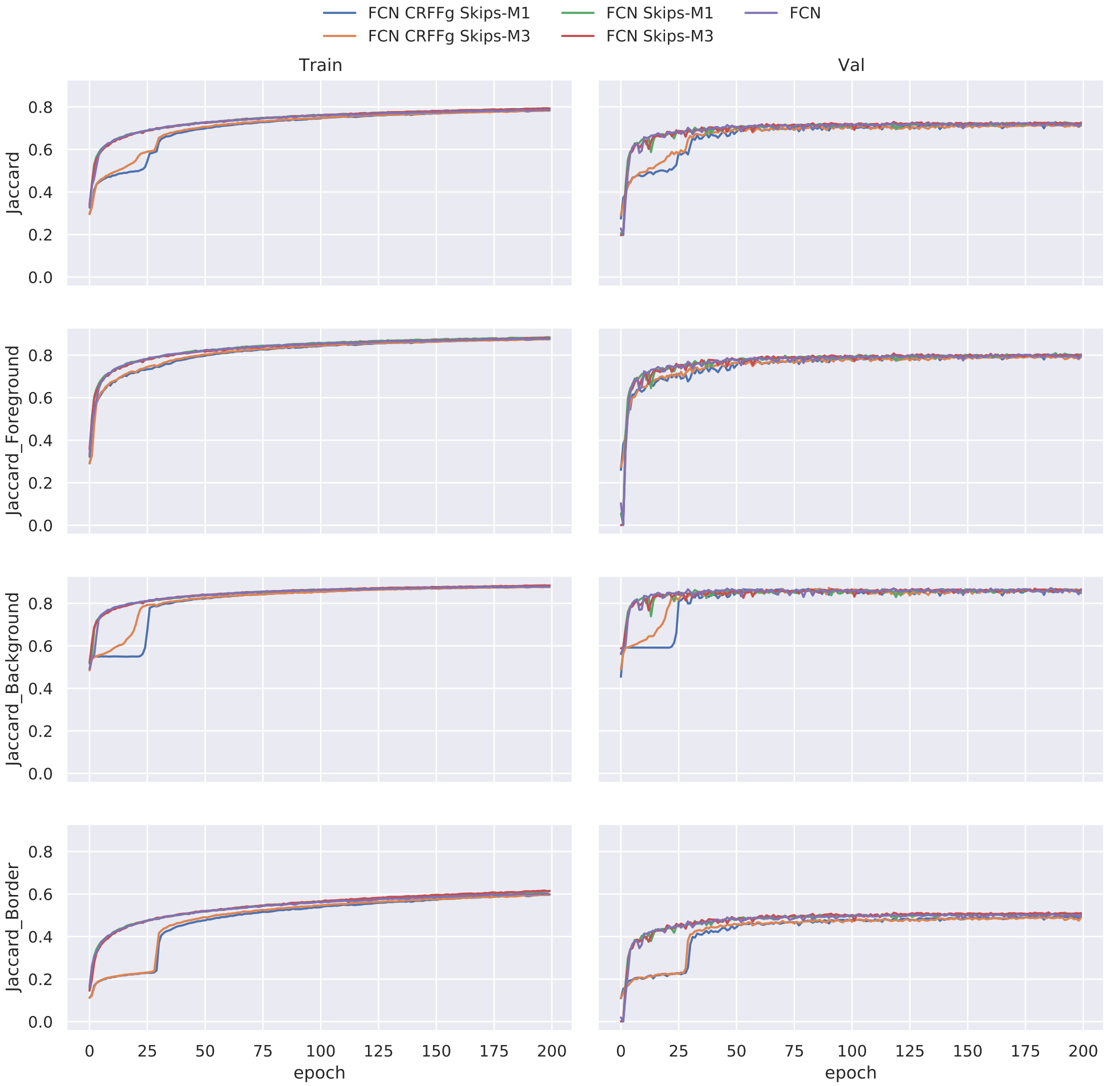


Figure **A-8**. Train Jaccard Oxford IIIt Pet FCN models

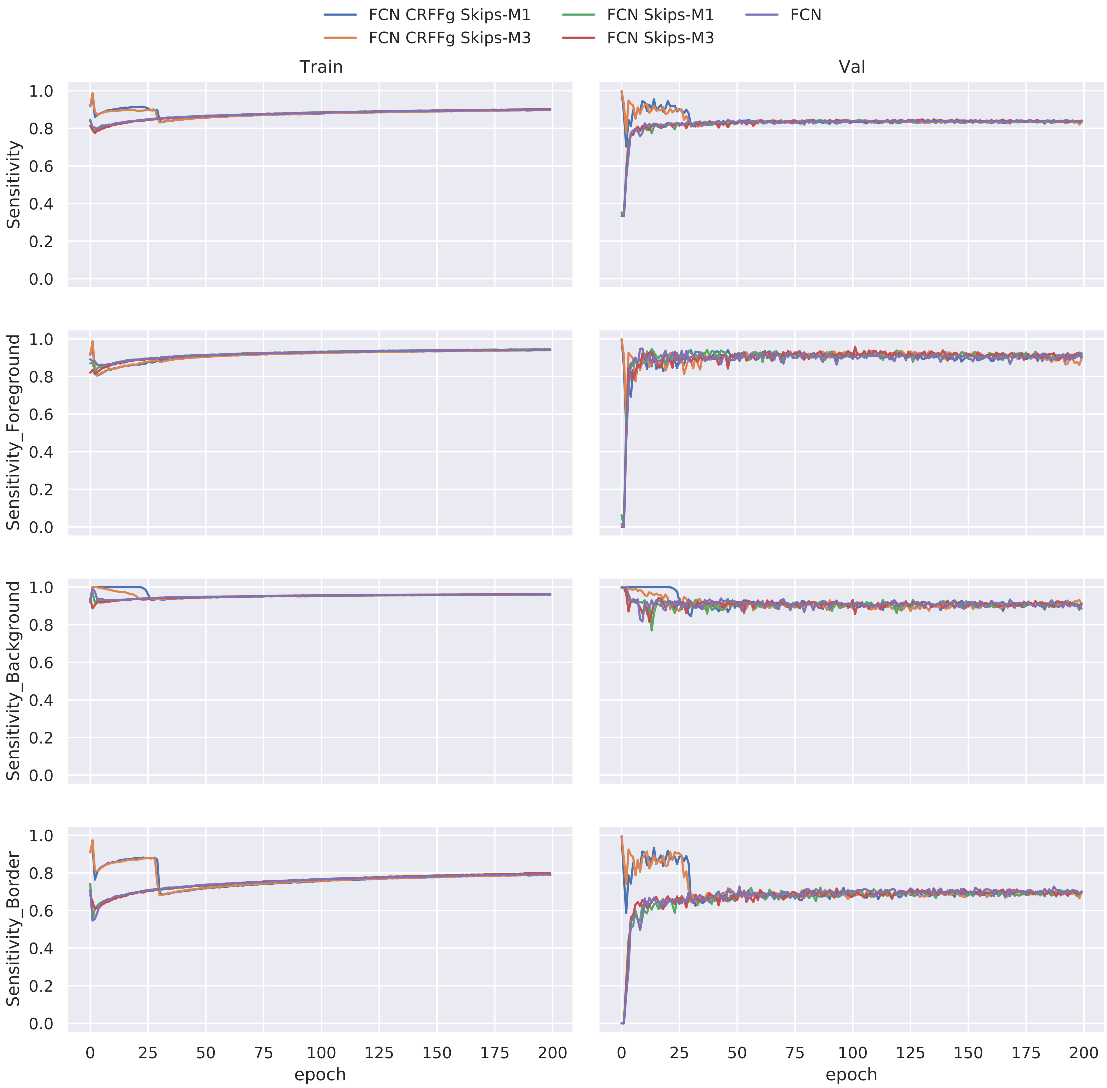


Figure **A-9**. Train Sensitivity Oxford IIIt Pet FCN models

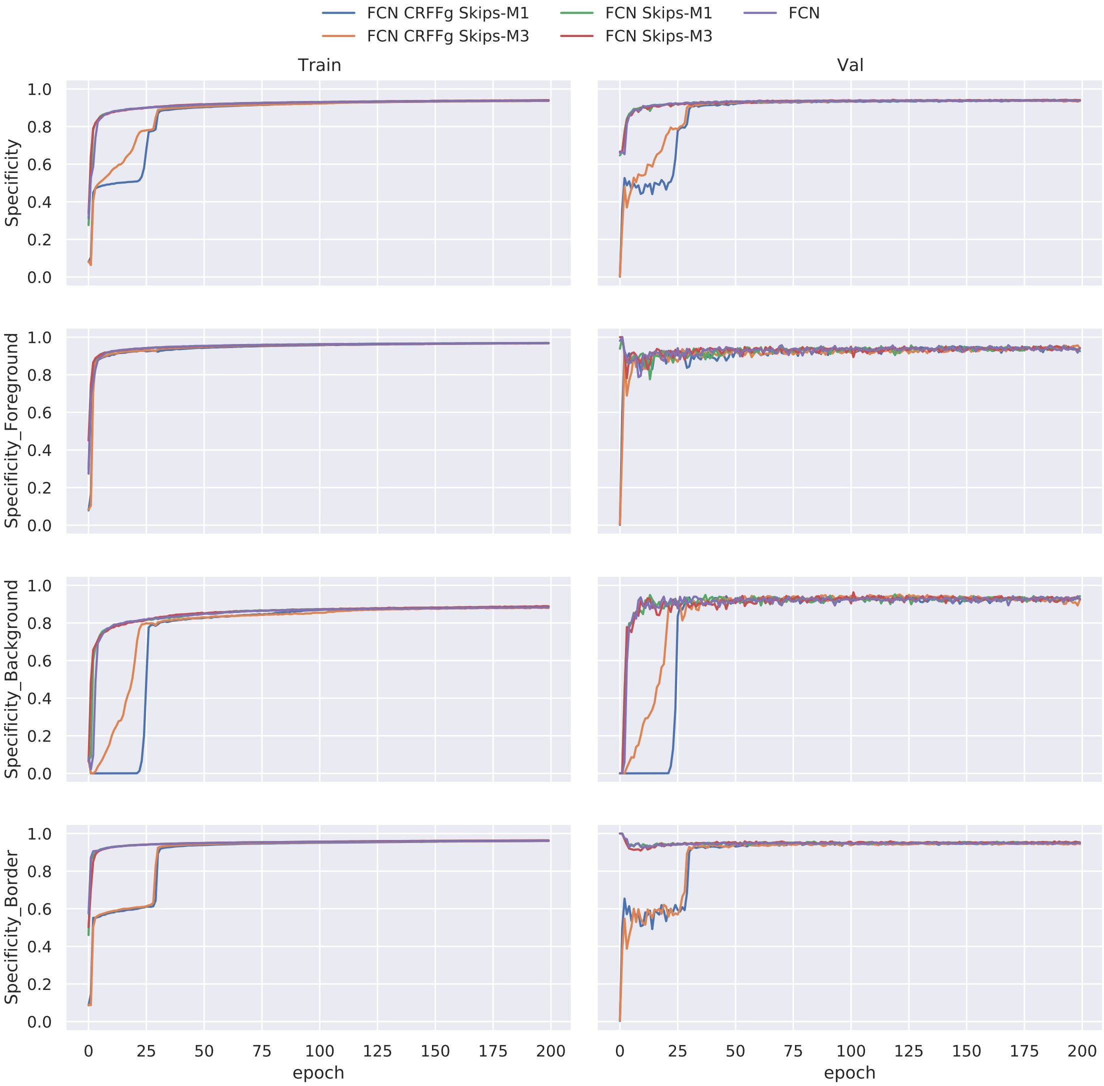


Figure **A-10**. Train Specificity Oxford IIIt Pet FCN models

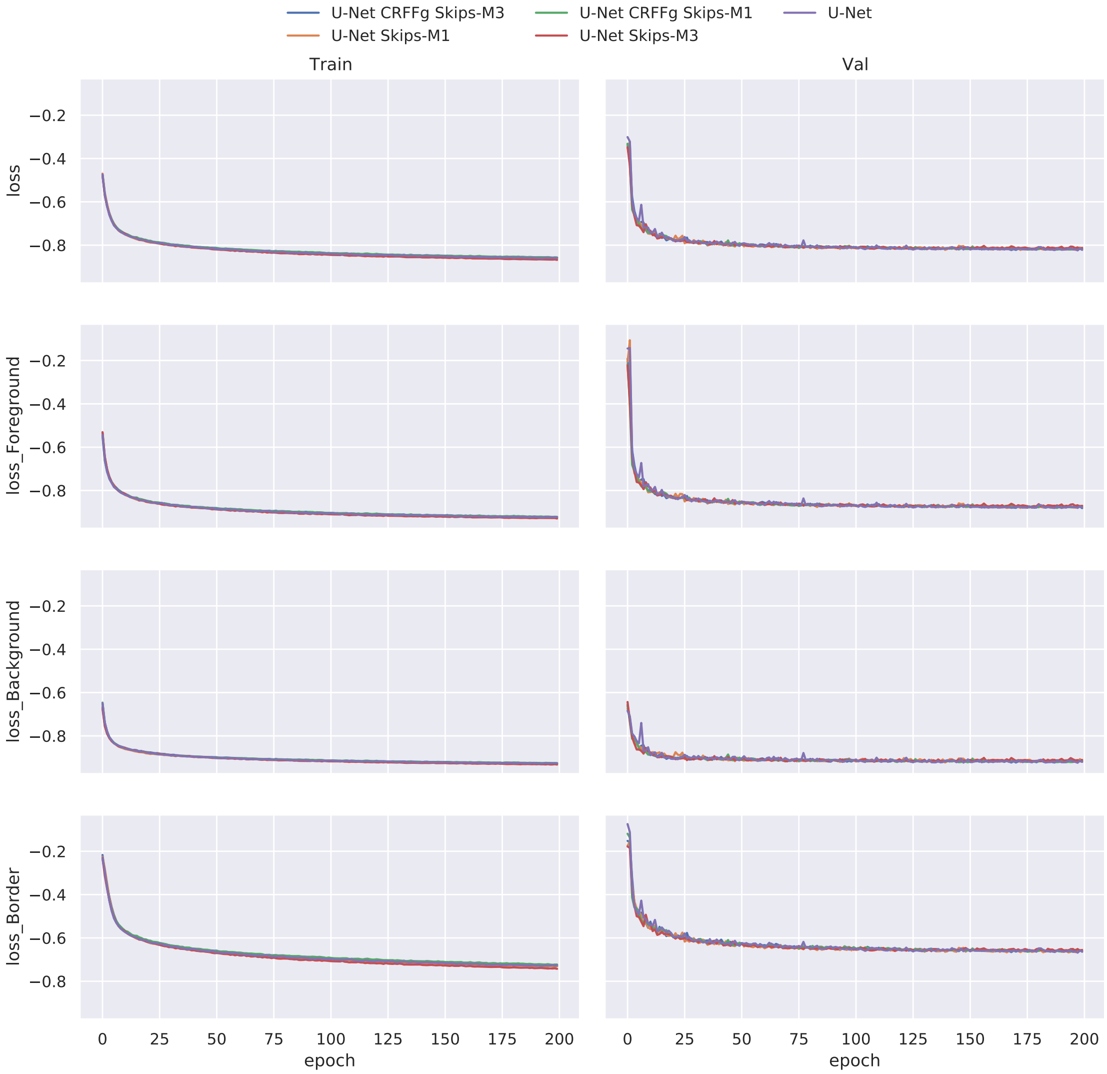


Figure **A-11**. Train Loss Oxford IIIt Pet U-Net models

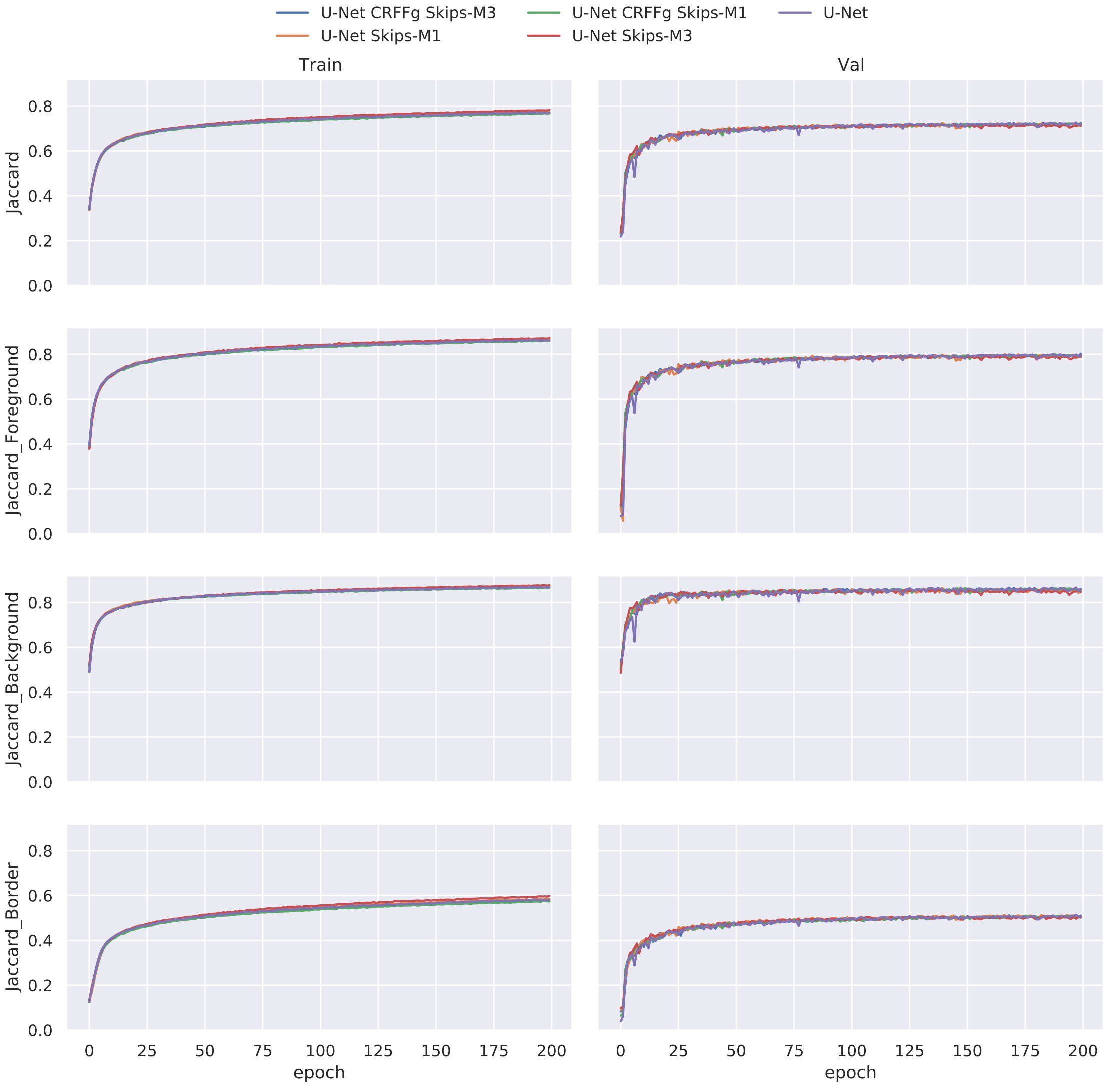


Figure **A-12**. Train Jaccard Oxford IIIt Pet U-Net models

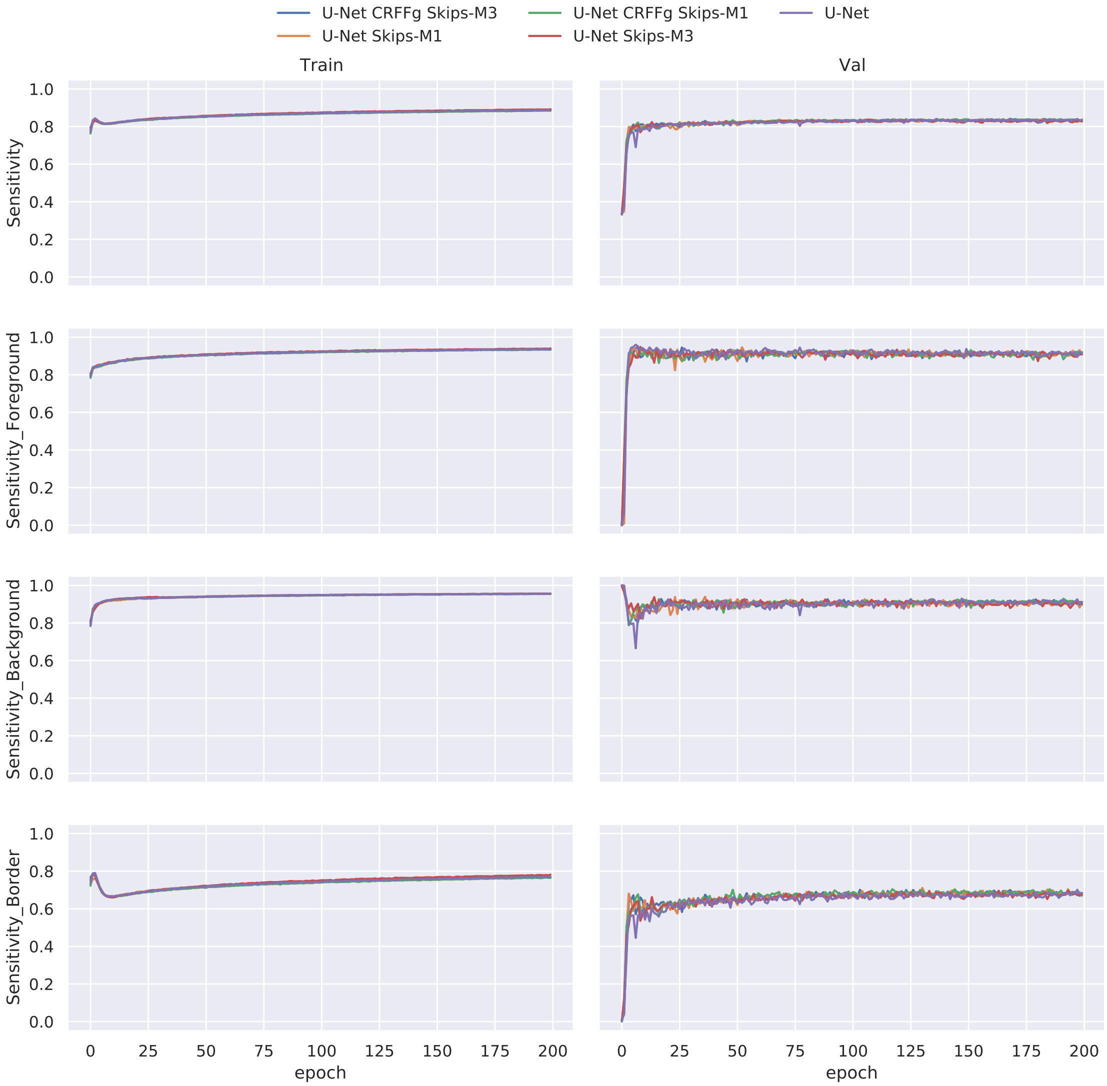


Figure **A-13**. Train Sensitivity Oxford IIIt Pet U-Net models

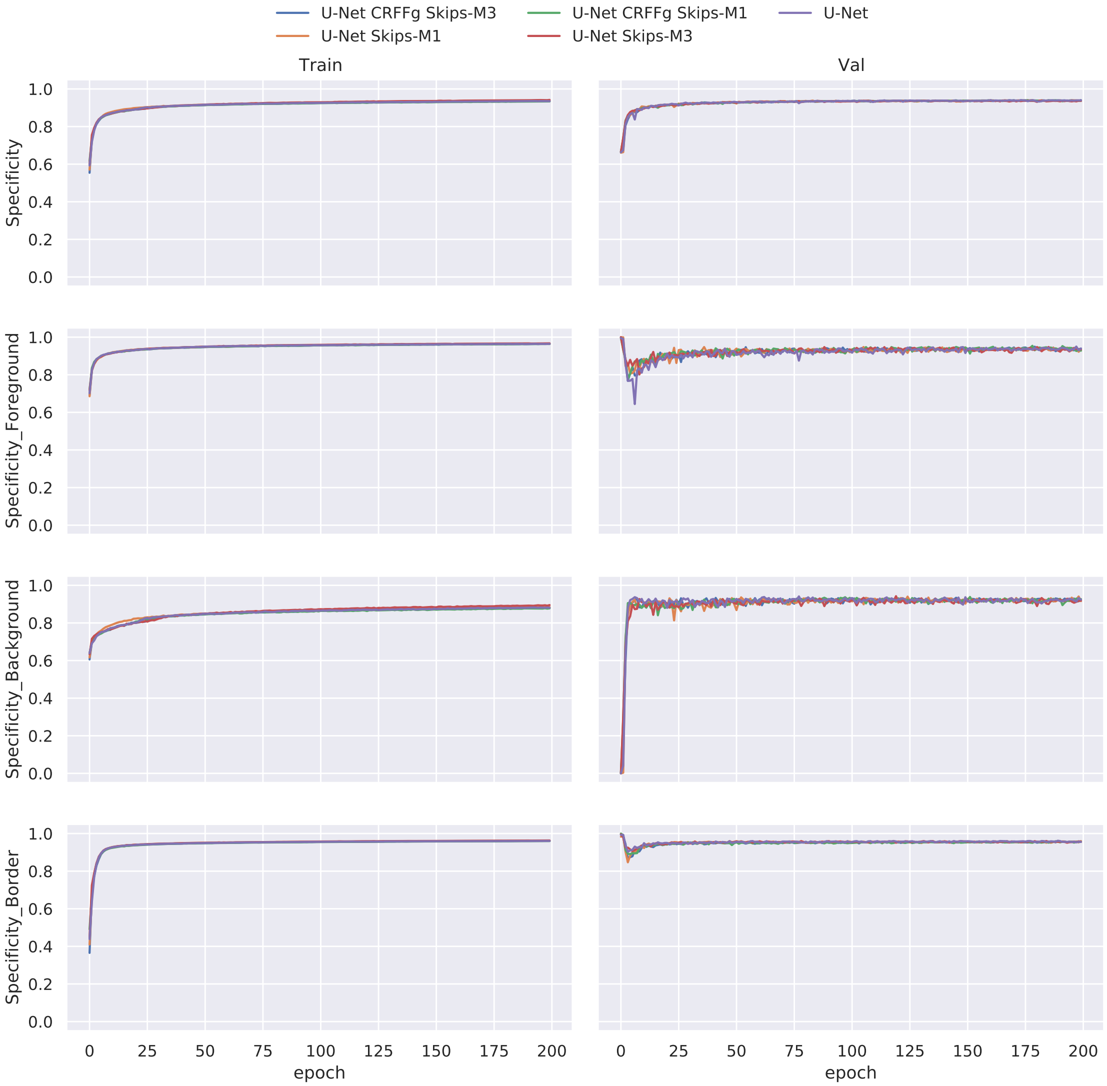


Figure **A-14**. Train Specificity Oxford IIIt Pet U-Net models

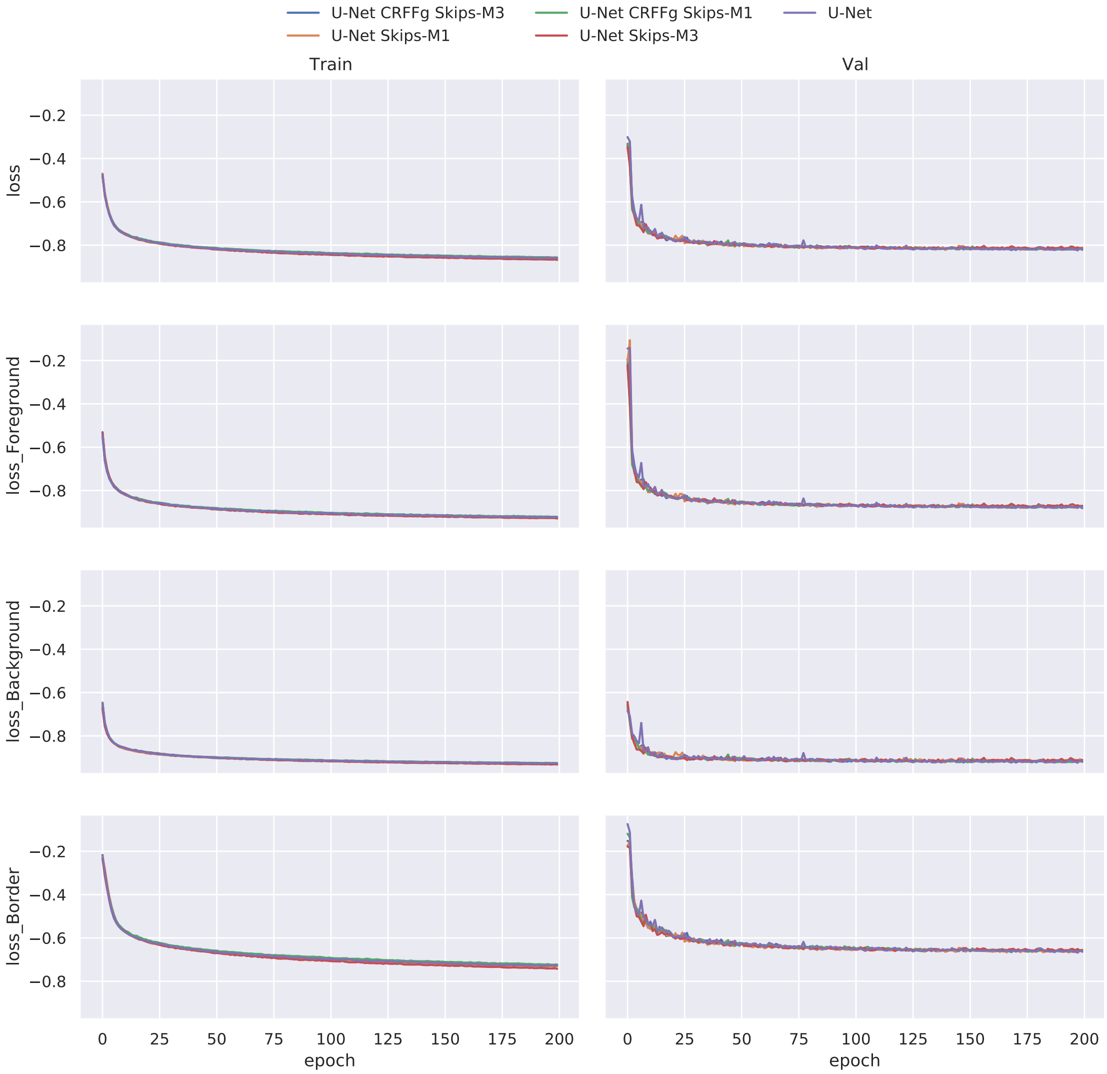


Figure **A-15**. Train Loss Oxford IIIt Pet U-Net models

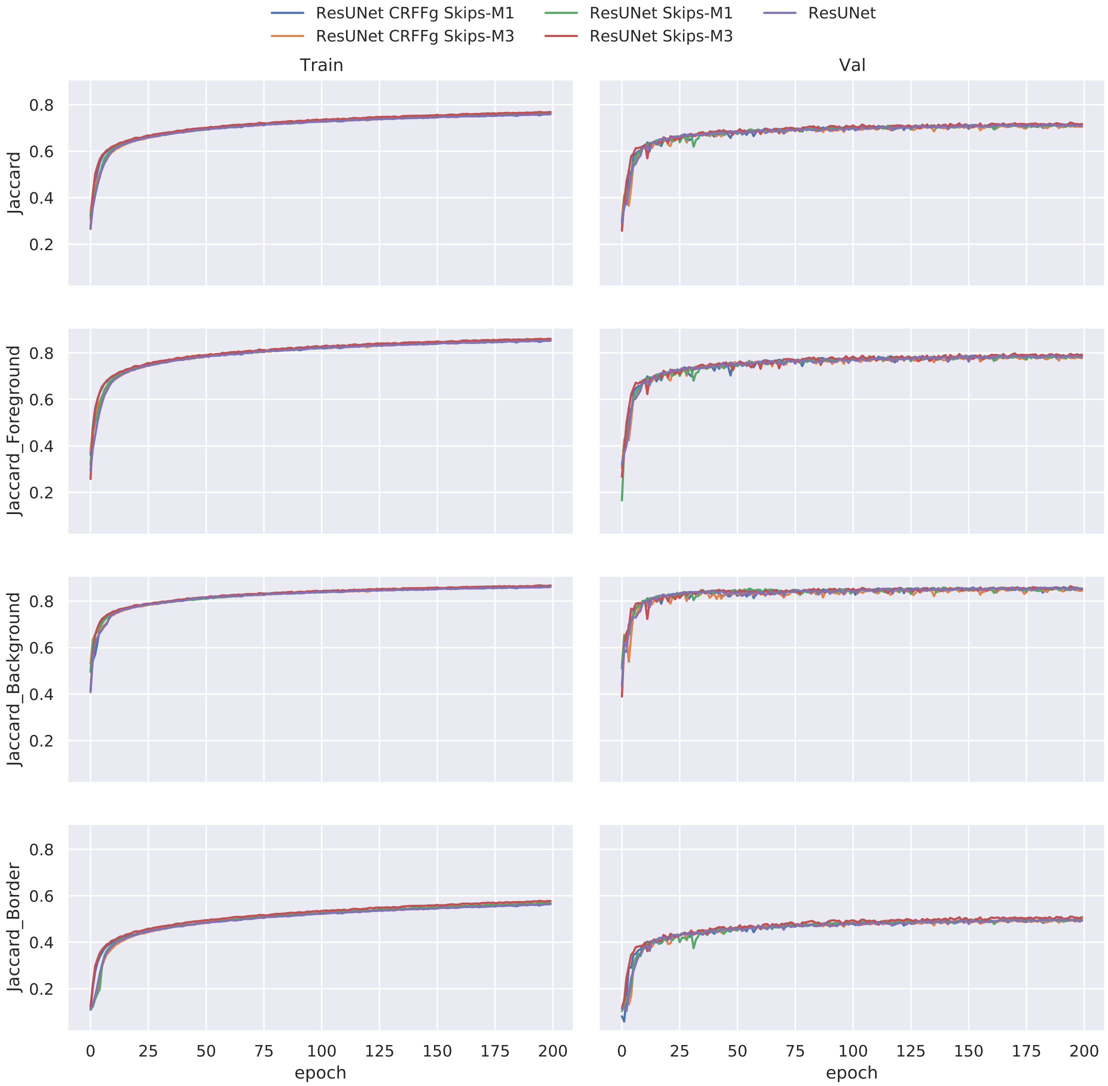


Figure **A-16**. Train Jaccard Oxford IIIt Pet ResUNet models

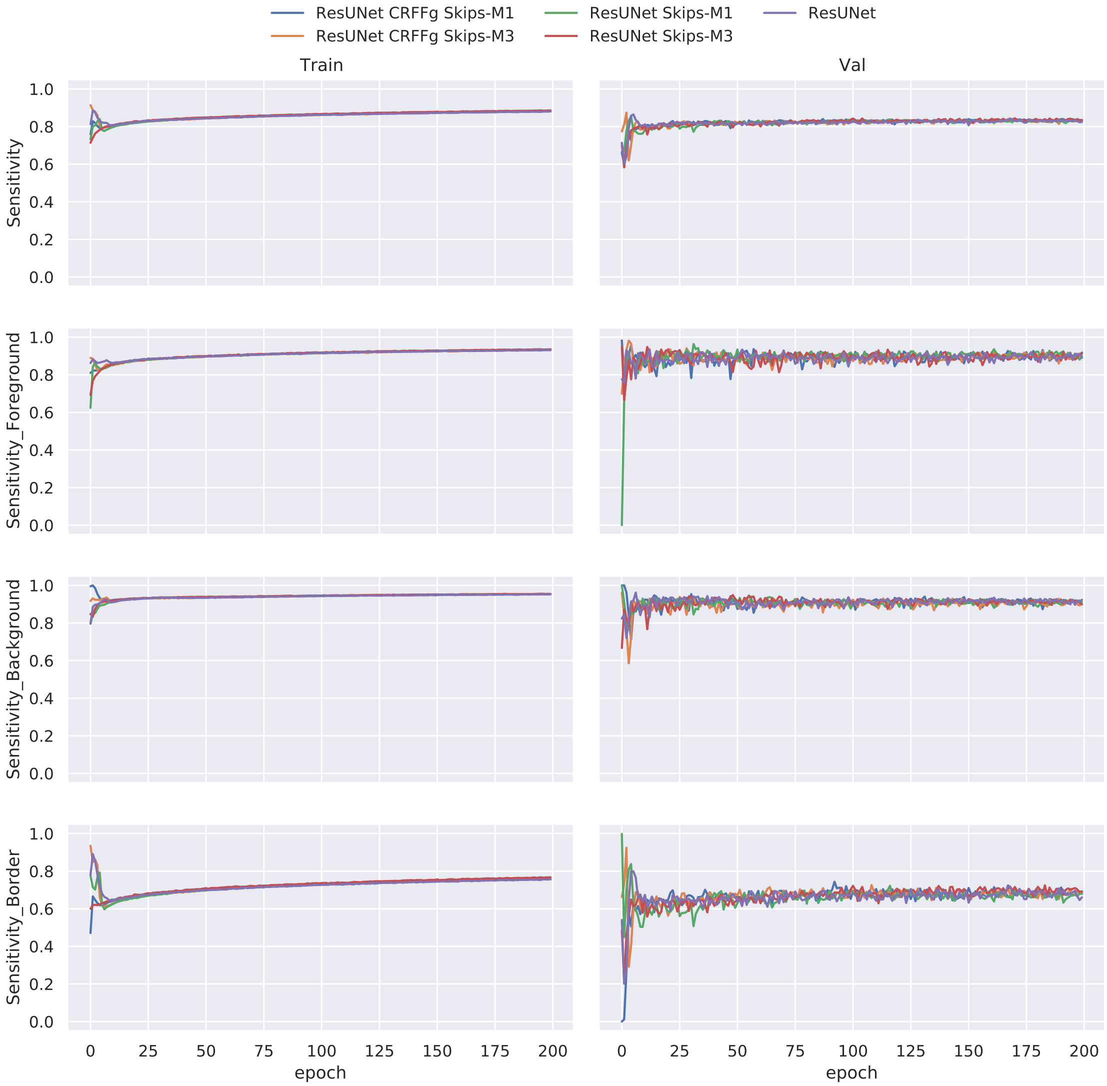


Figure **A-17**. Train Sensitivity Oxford IIIt Pet ResUNet models

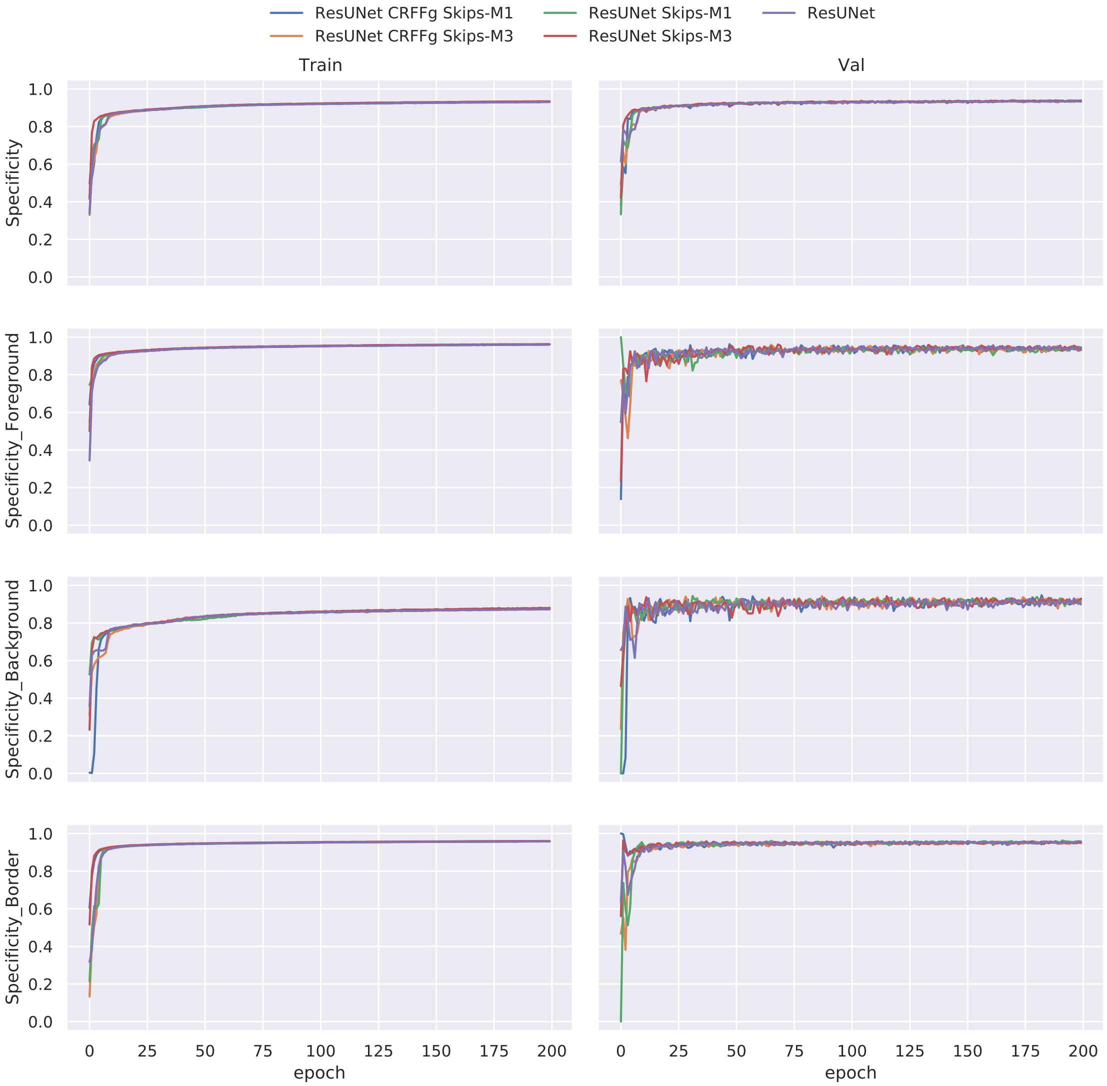


Figure **A-18**. Train Specificity Oxford IIIt Pet ResUNet models

APPENDIX

TWO

RESULT TABLES IN THE TEST PARTITION

model	metric	Without Data Augmentation	Rank	With Data Augmentation	rank
FCN	Dice	0.9527±0.0238	3.0	0.8646±0.0624	10.0
	Jaccard	0.9106±0.0424	3.0	0.7668±0.0969	10.0
	Sensitivity	0.9352±0.0482	4.0	0.8260±0.1098	6.0
	Specificity	0.9857±0.0105	7.0	0.9697±0.0186	13.0
FCN CRFFg S-M1	Dice	0.9530±0.0257	2.0	0.8510±0.0623	12.0
	Jaccard	0.9113±0.0456	2.0	0.7456±0.0913	12.0
	Sensitivity	0.9424±0.0526	3.0	0.8016±0.0999	13.0
	Specificity	0.9810±0.0158	12.0	0.9697±0.0233	14.0
FCN CRFFg S-M3	Dice	0.9480±0.0224	5.0	0.8346±0.0916	15.0
	Jaccard	0.9021±0.0403	5.0	0.7262±0.1284	15.0
	Sensitivity	0.9340±0.0423	6.0	0.7771±0.1325	15.0
	Specificity	0.9804±0.0168	13.0	0.9714±0.0246	10.0
FCN S-M1	Dice	0.9469±0.0273	6.0	0.8421±0.0870	14.0
	Jaccard	0.9003±0.0486	6.0	0.7367±0.1254	14.0
	Sensitivity	0.9286±0.0518	7.0	0.7867±0.1422	14.0
	Specificity	0.9843±0.0109	9.0	0.9714±0.0207	9.0
FCN S-M3	Dice	0.9519±0.0281	4.0	0.8470±0.0737	13.0
	Jaccard	0.9096±0.0499	4.0	0.7414±0.1070	13.0
	Sensitivity	0.9341±0.0543	5.0	0.8160±0.1152	9.0
	Specificity	0.9865±0.0107	6.0	0.9604±0.0300	15.0
ResUNet	Dice	0.9348±0.0502	11.0	0.8569±0.0779	11.0
	Jaccard	0.8816±0.0868	11.0	0.7575±0.1152	11.0
	Sensitivity	0.9029±0.0825	12.0	0.8152±0.1316	11.0
	Specificity	0.9896±0.0067	2.0	0.9712±0.0180	12.0
ResUNet CRFFg S-M1	Dice	0.9456±0.0317	7.0	0.8851±0.0449	4.0
	Jaccard	0.8984±0.0560	7.0	0.7968±0.0709	4.0
	Sensitivity	**0.9472±0.0540**	**1.0**	0.8283±0.0853	5.0
	Specificity	0.9725±0.0230	14.0	0.9841±0.0123	3.0
ResUNet CRFFg S-M3	Dice	0.9111±0.0602	15.0	**0.8969±0.0444**	**1.0**
	Jaccard	0.8420±0.0951	15.0	**0.8160±0.0737**	**1.0**
	Sensitivity	0.9075±0.0607	11.0	**0.8675±0.0803**	**1.0**
	Specificity	0.9663±0.0346	15.0	0.9712±0.0244	11.0
ResUNet S-M1	Dice	**0.9558±0.0279**	**1.0**	0.8865±0.0676	3.0
	Jaccard	**0.9167±0.0498**	**1.0**	0.8026±0.1061	3.0
	Sensitivity	0.9459±0.0482	2.0	0.8403±0.1123	2.0
	Specificity	0.9831±0.0152	10.0	0.9750±0.0287	8.0
ResUNet S-M3	Dice	0.9237±0.0411	14.0	0.8677±0.0894	9.0
	Jaccard	0.8610±0.0713	14.0	0.7763±0.1281	9.0
	Sensitivity	0.8875±0.0756	14.0	0.8179±0.1333	8.0
	Specificity	0.9846±0.0128	8.0	0.9755±0.0217	7.0
U-Net	Dice	0.9371±0.0312	10.0	0.8713±0.0756	8.0
	Jaccard	0.8832±0.0551	10.0	0.7796±0.1145	8.0
	Sensitivity	0.9120±0.0571	10.0	0.8107±0.1248	12.0
	Specificity	0.9811±0.0199	11.0	0.9847±0.0130	2.0
U-Net CRFFg S-M1	Dice	0.9448±0.0297	8.0	0.8827±0.0617	5.0
	Jaccard	0.8969±0.0528	8.0	0.7954±0.0965	5.0
	Sensitivity	0.9160±0.0561	9.0	0.8383±0.1062	4.0
	Specificity	**0.9902±0.0057**	**1.0**	0.9780±0.0124	5.0
U-Net CRFFg S-M3	Dice	0.9252±0.0404	13.0	0.8821±0.0645	6.0
	Jaccard	0.8634±0.0694	13.0	0.7948±0.1004	6.0
	Sensitivity	0.8831±0.0730	15.0	0.8231±0.1110	7.0
	Specificity	0.9893±0.0066	3.0	**0.9873±0.0088**	**1.0**
U-Net S-M1	Dice	0.9400±0.0364	9.0	0.8898±0.0536	2.0
	Jaccard	0.8890±0.0635	9.0	0.8056±0.0861	2.0
	Sensitivity	0.9162±0.0619	8.0	0.8384±0.0904	3.0
	Specificity	0.9866±0.0086	5.0	0.9777±0.0208	6.0
U-Net S-M3	Dice	0.9293±0.0419	12.0	0.8767±0.0772	7.0
	Jaccard	0.8707±0.0728	12.0	0.7883±0.1152	7.0
	Sensitivity	0.8934±0.0792	13.0	0.8152±0.1181	10.0
	Specificity	0.9878±0.0098	4.0	0.9805±0.0189	4.0

Table **B-1**. Infrared Thermal Images results on test partition

		Foreground	Background	Border	Average	rank
model	metric					
FCN	Dice	0.8240±0.0832	**0.9336±0.0587**	0.6514±0.1175	0.8869±0.1134	3.5
	Jaccard	0.7281±0.0982	**0.8806±0.0923**	0.4933±0.1185	0.8103±0.1369	3.75
	Sensitivity	0.8432±0.0796	0.9263±0.0755	0.7060±0.1321	0.8974±0.1189	5.25
	Specificity	0.9430±0.0364	0.9261±0.0963	0.9492±0.0228	0.9538±0.0533	6.5
FCN CRFFg S-M1	Dice	0.8166±0.0841	0.9297±0.0673	0.6393±0.1176	0.8807±0.1107	10.0
	Jaccard	0.7184±0.1009	0.8749±0.1003	0.4801±0.1197	0.8002±0.1373	10.25
	Sensitivity	0.8401±0.0803	0.9203±0.0801	**0.7112±0.1292**	0.8888±0.1166	7.0
	Specificity	0.9421±0.0362	0.9300±0.0892	0.9432±0.0251	0.9532±0.0549	7.75
FCN CRFFg S-M3	Dice	0.8164±0.0795	0.9282±0.0625	0.6388±0.1131	0.8824±0.1033	11.0
	Jaccard	0.7171±0.0957	0.8716±0.0966	0.4788±0.1154	0.8010±0.1275	11.5
	Sensitivity	0.8398±0.0714	0.9180±0.0803	0.7091±0.1198	0.8922±0.1025	7.5
	Specificity	0.9420±0.0323	0.9304±0.0758	0.9426±0.0284	0.9531±0.0501	8.25
FCN S-M1	Dice	0.8195±0.0830	0.9250±0.0686	0.6493±0.1148	0.8841±0.1116	9.5
	Jaccard	0.7211±0.1002	0.8668±0.1009	0.4908±0.1189	0.8057±0.1375	9.5
	Sensitivity	0.8410±0.0763	0.9030±0.0893	0.7046±0.1269	**0.9156±0.1099**	6.75
	Specificity	0.9454±0.0350	**0.9435±0.0858**	0.9484±0.0248	0.9443±0.0632	7.25
FCN S-M3	Dice	**0.8265±0.0766**	0.9330±0.0581	0.6568±0.1110	**0.8897±0.1009**	**1.5**
	Jaccard	**0.7300±0.0922**	0.8793±0.0908	0.4984±0.1144	0.8124±0.1247	**2.0**
	Sensitivity	**0.8458±0.0741**	0.9233±0.0696	0.7035±0.1240	0.9107±0.1067	**3.75**
	Specificity	**0.9457±0.0344**	0.9341±0.0894	0.9516±0.0231	0.9514±0.0476	5.5
ResUNet	Dice	0.8173±0.0875	0.9271±0.0629	0.6476±0.1235	0.8773±0.1190	11.25
	Jaccard	0.7186±0.1028	0.8698±0.0964	0.4901±0.1249	0.7960±0.1416	11.75
	Sensitivity	0.8320±0.0888	**0.9320±0.0541**	0.6779±0.1463	0.8860±0.1354	11.25
	Specificity	0.9391±0.0418	0.9049±0.1201	0.9555±0.0197	**0.9569±0.0393**	8.0
ResUNet CRFFg S-M1	Dice	0.8215±0.0771	0.9276±0.0611	0.6500±0.1159	0.8868±0.0982	7.0
	Jaccard	0.7232±0.0957	0.8703±0.0946	0.4919±0.1220	0.8075±0.1255	7.5
	Sensitivity	0.8426±0.0760	0.9148±0.0740	0.7005±0.1380	0.9124±0.1000	6.75
	Specificity	0.9446±0.0333	0.9349±0.0849	0.9507±0.0212	0.9481±0.0554	7.0
ResUNet CRFFg S-M3	Dice	0.8142±0.0880	0.9219±0.0722	0.6444±0.1224	0.8763±0.1134	14.5
	Jaccard	0.7141±0.1053	0.8623±0.1074	0.4865±0.1241	0.7936±0.1392	14.5
	Sensitivity	0.8369±0.0866	0.9167±0.0873	0.7087±0.1435	0.8853±0.1175	10.0
	Specificity	0.9379±0.0401	0.9152±0.1022	0.9465±0.0232	0.9520±0.0594	12.5
ResUNet S-M1	Dice	0.8210±0.0828	0.9290±0.0629	0.6511±0.1205	0.8829±0.1105	6.75
	Jaccard	0.7233±0.0993	0.8730±0.0967	0.4937±0.1245	0.8033±0.1342	6.5
	Sensitivity	0.8392±0.0845	0.9269±0.0694	0.6990±0.1403	0.8917±0.1244	8.0
	Specificity	0.9408±0.0391	0.9164±0.1084	0.9513±0.0218	0.9548±0.0475	9.0
ResUNet S-M3	Dice	0.8258±0.0855	0.9284±0.0702	**0.6595±0.1219**	0.8895±0.1072	2.5
	Jaccard	0.7299±0.1022	0.8731±0.1041	**0.5033±0.1258**	**0.8134±0.1313**	**2.0**
	Sensitivity	0.8437±0.0829	0.9196±0.0848	0.7078±0.1395	0.9037±0.1136	5.0
	Specificity	0.9443±0.0376	0.9275±0.0993	0.9525±0.0216	0.9528±0.0605	**5.25**
U-Net	Dice	0.8207±0.0840	0.9271±0.0652	0.6542±0.1171	0.8806±0.1117	8.75
	Jaccard	0.7224±0.1032	0.8701±0.0999	0.4968±0.1231	0.8003±0.1396	8.5
	Sensitivity	0.8395±0.0805	0.9211±0.0758	0.6966±0.1362	0.9009±0.1085	8.25
	Specificity	0.9423±0.0366	0.9230±0.0930	0.9525±0.0226	0.9515±0.0520	7.75
U-Net CRFFg S-M1	Dice	0.8158±0.0880	0.9255±0.0666	0.6449±0.1212	0.8770±0.1213	13.0
	Jaccard	0.7169±0.1052	0.8677±0.1018	0.4869±0.1242	0.7962±0.1454	12.75
	Sensitivity	0.8354±0.0851	0.9151±0.0858	0.6864±0.1422	0.9049±0.1188	10.5
	Specificity	0.9398±0.0396	0.9238±0.1040	0.9522±0.0214	0.9434±0.0647	10.75
U-Net CRFFg S-M3	Dice	0.8217±0.0867	0.9280±0.0659	0.6538±0.1240	0.8833±0.1140	6.25
	Jaccard	0.7247±0.1039	0.8717±0.0989	0.4974±0.1284	0.8051±0.1402	5.75
	Sensitivity	0.8381±0.0843	0.9227±0.0711	0.6906±0.1431	0.9009±0.1215	8.25
	Specificity	0.9427±0.0389	0.9210±0.1074	0.9544±0.0216	0.9527±0.0509	6.75
U-Net S-M1	Dice	0.8235±0.0846	0.9282±0.0648	0.6557±0.1207	0.8867±0.1115	4.5
	Jaccard	0.7269±0.1020	0.8719±0.0974	0.4991±0.1266	0.8097±0.1362	4.0
	Sensitivity	0.8355±0.0850	0.9226±0.0739	0.6824±0.1435	0.9016±0.1217	9.5
	Specificity	0.9417±0.0400	0.9167±0.1094	**0.9573±0.0197**	0.9512±0.0570	8.5
U-Net S-M3	Dice	0.8190±0.0895	0.9234±0.0686	0.6519±0.1245	0.8817±0.1169	10.0
	Jaccard	0.7209±0.1074	0.8643±0.1030	0.4953±0.1277	0.8032±0.1447	9.75
	Sensitivity	0.8342±0.0862	0.9129±0.0850	0.6888±0.1411	0.9009±0.1190	12.25
	Specificity	0.9417±0.0400	0.9214±0.1025	0.9541±0.0214	0.9497±0.0613	9.25

Table **B-2**. Oxford IIIt Pet results on test partition

		rank
model	metric	
FCN	Dice	8.67
	Jaccard	9.00
	Sensitivity	10.33
	Specificity	5.00
FCN CRFFg S-M1	Dice	8.33
	Jaccard	8.67
	Sensitivity	7.33
	Specificity	9.00
FCN CRFFg S-M3	Dice	10.33
	Jaccard	11.00
	Sensitivity	11.33
	Specificity	8.33
FCN S-M1	Dice	8.67
	Jaccard	9.00
	Sensitivity	6.67
	Specificity	12.33
FCN S-M3	Dice	5.00
	Jaccard	5.00
	Sensitivity	**4.67**
	Specificity	12.67
ResUNet	Dice	9.67
	Jaccard	9.33
	Sensitivity	8.00
	Specificity	10.67
ResUNet CRFFg S-M1	Dice	6.67
	Jaccard	6.67
	Sensitivity	7.00
	Specificity	7.67
ResUNet CRFFg S-M3	Dice	9.67
	Jaccard	9.00
	Sensitivity	9.33
	Specificity	7.33
ResUNet S-M1	Dice	5.67
	Jaccard	6.33
	Sensitivity	7.00
	Specificity	4.67
ResUNet S-M3	Dice	8.67
	Jaccard	8.00
	Sensitivity	9.67
	Specificity	9.00
U-Net	Dice	9.33
	Jaccard	9.67
	Sensitivity	9.67
	Specificity	6.67
U-Net CRFFg S-M1	Dice	7.67
	Jaccard	8.00
	Sensitivity	7.33
	Specificity	5.67
U-Net CRFFg S-M3	Dice	10.00
	Jaccard	9.00
	Sensitivity	9.67
	Specificity	**3.33**
U-Net S-M1	Dice	**4.33**
	Jaccard	**4.33**
	Sensitivity	5.33
	Specificity	9.33
U-Net S-M3	Dice	7.33
	Jaccard	7.00
	Sensitivity	6.67
	Specificity	8.33

Table **B-3**. All models results on all test partition

BIBLIOGRAPHY

[sut, 2015] (2015). On the error of random fourier features. (page 97)

[Alebiosu et al., 2023] Alebiosu, D. O., Dharmaratne, A., and Lim, C. H. (2023). Improving tuberculosis severity assessment in computed tomography images using novel davou-net segmentation and deep learning framework. *Expert Systems with Applications*, 213:119287. (page 2)

[Aljabri and AlGhamdi, 2022] Aljabri, M. and AlGhamdi, M. (2022). A review on the use of deep learning for medical images segmentation. *Neurocomputing*, 506:311–335. (page 2)

[Altameem et al., 2022] Altameem, A., Mahanty, C., Poonia, R. C., Saudagar, A. K. J., and Kumar, R. (2022). Breast cancer detection in mammography images using deep convolutional neural networks and fuzzy ensemble modeling techniques. *Diagnostics*, 12(8). (page 2)

[Álvarez-Meza et al., 2014] Álvarez-Meza, A. M., Cárdenas-Peña, D., and Castellanos-Dominguez, G. (2014). Unsupervised kernel function building using maximization of information potential variability. In Bayro-Corrochano, E. and Hancock, E., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 335–342, Cham. Springer International Publishing. (page 41)

[Alzubaidi et al., 2020] Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., and R. Oleiwi, S. (2020). Towards a better understanding of transfer learning for medical imaging: a case study. *Applied Sciences*, 10(13):4523. (page 14)

[Amann et al., 2020] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., and Consortium, P. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9. (page 7)

[Anand et al., 2021] Anand, T., Sinha, S., Mandal, M., Chamola, V., and Yu, F. R. (2021). Agrisegnet: Deep aerial semantic segmentation framework for iot-assisted precision agriculture. *IEEE Sensors Journal*, 21(16):17581–17590. (page 2)

[Anas et al., 2017a] Anas, E. M. A., Nouranian, S., Mahdavi, S. S., Spadinger, I., Morris, W. J., Salcudean, S. E., Mousavi, P., and Abolmaesumi, P. (2017a). Clinical target-volume delineation in prostate brachytherapy using residual neural networks. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 365–373. Springer. (page 12)

[Anas et al., 2017b] Anas, E. M. A., Nouranian, S., Mahdavi, S. S., Spadinger, I., Morris, W. J., Salcudean, S. E., Mousavi, P., and Abolmaesumi, P. (2017b). Clinical target-volume delineation in prostate brachytherapy using residual neural networks. In Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., and Duchesne, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pages 365–373, Cham. Springer International Publishing. (page 61)

[Antonelli et al., 2022] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. (2022). The medical segmentation decathlon. *Nature communications*, 13(1):4128. (page 2)

[Arteaga-Marrero et al., 2021] Arteaga-Marrero, N., Hernández, A., Villa, E., González-Pérez, S., Luque, C., and Ruiz-Alzola, J. (2021). Segmentation approaches for diabetic foot disorders. *Sensors*, 21(3):934. (page 9)

[Ataloglou et al., 2019] Ataloglou, D., Dimou, A., Zarpalas, D., and Daras, P. (2019). Fast and precise hippocampus segmentation through deep convolutional neural network ensembles and transfer learning. *Neuroinformatics*, 17(4):563—582. (page 2)

[Azad et al., 2023] Azad, R., Kazerouni, A., Heidari, M., Aghdam, E. K., Molaei, A., Jia, Y., Jose, A., Roy, R., and Merhof, D. (2023). Advances in medical image analysis with vision transformers: A comprehensive review. *arXiv preprint arXiv:2301.03505*. (pages 13 and 98)

[Badrinarayanan et al., 2016] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2016). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. (page 12)

[Bengio and Lecun, 2017] Bengio, Y. and Lecun, Y. (2017). Scaling learning algorithms towards ai to appear in “large-scale kernel machines”. *George Mason University: Fairfax, VA, USA*. (page 40)

[Bennetot et al., 2022] Bennetot, A., Donadello, I., Qadi, A. E., Dragoni, M., Frossard, T., Wagner, B., Saranti, A., Tulli, S., Trocan, M., Chatila, R., Holzinger, A., d’Avila Garcez, A., and Díaz-Rodríguez, N. (2022). A practical guide on explainable ai techniques applied on biomedical use case applications. (page 10)

[Bi et al., 2017] Bi, L., Kim, J., Kumar, A., Fulham, M. J., and Feng, D. (2017). Stacked fully convolutional networks with multi-channel learning: application to medical image segmentation. *The Visual Computer*, 33:1061 – 1071. (page 12)

[Bougrine et al., 2022] Bougrine, A., Harba, R., Canals, R., Ledee, R., Jabloun, M., and Villeneuve, A. (2022). Segmentation of plantar foot thermal images using prior information. *Sensors*, 22(10):3835. (page 9)

[Bouvet et al., 2020] Bouvet, L., Roukhomovsky, M., Desgranges, F.-P., Allaouchiche, B., and Chassard, D. (2020). Infrared thermography to assess dermatomal levels of labor epidural analgesia with 1 mg/ml ropivacaine plus 0.5 µg/ml sufentanil: a prospective cohort study. *International Journal of Obstetric Anesthesia*, 41:53–58. (page 2)

[Brocki and Chung, 2022] Brocki, L. and Chung, N. C. (2022). Evaluation of interpretability methods and perturbation artifacts in deep neural networks. *arXiv preprint arXiv:2203.02928*. (page 21)

[Bron et al., 2015] Bron, E. E., Smits, M., van der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J. M., Steketee, R. M., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J. R., Garrett, C., Bastos-Leite, A. J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., Álvarez Meza, A. M., Dolph, C. V., Iftekharuddin, K. M., Eskildsen, S. F., Coupé, P., Fonov, V. S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K. R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G. M., Stoyanov, Z. V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J. C., Niessen, W. J., and Klein, S. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The caddementia challenge. *NeuroImage*, 111:562–579. (page 5)

[Bronstein et al., 2021] Bronstein, M. M., Bruna, J., Cohen, T., and Velickovic, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478. (pages 12 and 44)

[Brown et al., 1980] Brown, D. T., Wildsmith, J. A. W., Covino, B. G., and Scott, D. B. (1980). Effect of Baricity on Spinal Anaesthesia with Amethocaine. *BJA: British Journal of Anaesthesia*, 52(6):589–596. (page 3)

[Bruins et al., 2018a] Bruins, A., Kistemaker, K., Boom, A., Klaessens, J., Verdaasdonk, R., and Boer, C. (2018a). Thermographic skin temperature measurement compared with cold sensation in predicting the efficacy and distribution of epidural anesthesia. *Journal of clinical monitoring and computing*, 32(2):335–341. (page 4)

[Bruins et al., 2018b] Bruins, A. A., Kistemaker, K. R. J., Boom, A., Klaessens, J., Verdaasdonk, R., and Boer, C. (2018b). Thermographic skin temperature measurement compared with cold sensation in predicting the efficacy and distribution of epidural anesthesia. *Journal of Clinical Monitoring and Computing*, 32:335 – 341. (page 4)

[Burel et al., 2022] Burel, S., Evans, A., and Anghel, L. (2022). Improving dnn fault tolerance in semantic segmentation applications. *2022 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, pages 1–6. (page 1)

[Cakir et al., 2022] Cakir, S., Gauß, M., Häppeler, K., Ounajjar, Y., Heinle, F., and Marchthaler, R. (2022). Semantic segmentation for autonomous driving: Model evaluation, dataset generation, perspective comparison, and real-time capability. (page 1)

[Cao et al., 2021] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. pages 205–218. (page 13)

[Carmo et al., 2021] Carmo, D., Silva, B., Yasuda, C., Rittner, L., and Lotufo, R. (2021). Hippocampus segmentation on epilepsy and alzheimer's disease studies with multiple convolutional neural networks. *Heliyon*, 7(2):e06226. (page 2)

[Cavazos-Rehg et al., 2015] Cavazos-Rehg, P. A., Krauss, M. J., Spitznagel, E. L., Bommarito, K., Madden, T., Olsen, M. A., Subramaniam, H., Peipert, J. F., and Bierut, L. J. (2015). Maternal age and risk of labor and delivery complications. *Maternal and child health journal*, 19:1202–1211. (page 3)

[Chae et al., 2022] Chae, Y., Park, H.-J., and Lee, I.-S. (2022). Pain modalities in the body and brain: Current knowledge and future perspectives. *Neuroscience & Biobehavioral Reviews*, 139:104744. (page 4)

[Chang et al., 2020] Chang, Y.-T., Wang, Q., Hung, W.-C., Piramuthu, R., Tsai, Y.-H., and Yang, M.-H. (2020). Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization. (page 97)

[Chattopadhyay et al., 2017] Chattopadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2017). Grad-cam++: Improved visual explanations for deep convolutional networks. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, 2018-January:839–847. (pages 11, 22, and 81)

[Chen et al., 2021] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. (page 13)

[Chen et al., 2020a] Chen, Y., Ma, B., and Xia, Y. (2020a). α-unet++: A data-driven neural network architecture for medical image segmentation. In *DART/DCL@MICCAI*. (page 18)

[Chen et al., 2020b] Chen, Z., Niu, J., Liu, X., and Tang, S. (2020b). Selectscale: Mining more patterns from images via selective and soft dropout. (page 14)

[Chughtai et al., 2018] Chughtai, A. R., Navaee, M., Alijanvand, M. H., and Yaghoubinia, F. (2018). Comparing the effect of aromatherapy with essential oils of rosa damascena and lavender alone and in combination on severity of pain in the first phase of labor in primiparous women. (page 3)

[Conhaim and Girnius, 2023] Conhaim, J. and Girnius, A. (2023). 17 - anesthesia for cardiac patients during labor and delivery. In Wilson, J. H., Schnettler, W. T., Lubert, A. M., and Girnius, A., editors, *Maternal Cardiac Care*, pages 112–119. Elsevier, New Delhi. (page 4)

[Curatolo et al., 2005] Curatolo, M., Petersen-Felix, S., and Arendt-Nielsen, L. (2005). Assessment of regional analgesia in clinical practice and research. *British Medical Bulletin*, 71(1):61–76. (page 4)

[Dardouillet et al., 2022] Dardouillet, P., Benoit, A., Amri, E., Bolon, P., Dubucq, D., and Crédoz, A. (2022). Explainability of image semantic segmentation through shap values. In *ICPR-XAIE*. (page 23)

[De Sousa Ribeiro et al., 2020] De Sousa Ribeiro, F., Leontidis, G., and Kollias, S. (2020). Introducing routing uncertainty in capsule networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6490–6502. Curran Associates, Inc. (page 17)

[Dong et al., 2023] Dong, J., Chen, J., Xie, X., Lai, J., and Chen, H. (2023). Adversarial attack and defense for medical image analysis: Methods and applications. *arXiv preprint arXiv:2303.14133*. (page 7)

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929. (page 13)

[Doumard et al., 2022] Doumard, E., Aligon, J., Escriva, E., Excoffier, J.-B., Monsarrat, P., and Soulé-Dupuy, C. (2022). A comparative study of additive local explanation methods based on feature influences. In *24th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data ((DOLAP 2022)*, volume 3130, pages 31–40. CEUR-WS. org. (page 21)

[Dushatskiy et al., 2022] Dushatskiy, A., Lowe, G., Bosman, P. A. N., and Alderliesten, T. (2022). Data variation-aware medical image segmentation. (page 17)

[Fazilov et al., 2022] Fazilov, S. K., Yusupov, O., and Abdiyeva, K. S. (2022). Mammography image segmentation in breast cancer identification using the otsu method. *Web of Scientist: International Scientific Research Journal*, 3(8):196–205. (page 2)

[Galati et al., 2022] Galati, F., Ourselin, S., and Zuluaga, M. A. (2022). From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: A review. *Applied Sciences*. (page 8)

[Ghosh et al., 2022] Ghosh, K., Halder, T., Roy, M., Biswas, C., Gayen, R. K., and Chakravarty, D. (2022). A survey on medical image diagnosis systems: Problems and prospects. In *Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing: IEM-ICDC 2021*, pages 243–252. Springer. (page 7)

[Ghosh and Gupta, 2019] Ghosh, R. and Gupta, A. K. (2019). Scale steerable filters for locally scale-invariant convolutional neural networks. (page 16)

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org. (pages 14 and 44)

[Gottlieb et al., 2022] Gottlieb, M., Penington, A., and Schraft, E. (2022). Digital nerve blocks: A comprehensive review of techniques. *The Journal of Emergency Medicine*. (page 4)

[Guan and Liu, 2022] Guan, H. and Liu, M. (2022). Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185. (page 14)

[Gundersen, 2019] Gundersen, G. (2019). Random fourier features. (page 42)

[Haren et al., 2013] Haren, F., Kadic, L., and Driessen, J. (2013). Skin temperature measured by infrared thermography after ultrasound-guided blockade of the sciatic nerve. *Acta anaesthesiologica Scandinavica*, 57. (page 4)

[He et al., 2018] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2018). Mask r-cnn. (page 12)

[Holzinger et al., 2019] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1312. (pages 19 and 20)

[Hong et al., 2022] Hong, T.-P., Hu, M.-J., Yin, T.-K., and Wang, S.-L. (2022). A multi-scale convolutional neural network for rotation-invariant recognition. *Electronics*, 11(4):661. (page 16)

[Hoyle and Yentis, 2015] Hoyle, J. and Yentis, S. M. (2015). Assessing the height of block for caesarean section over the past three decades: trends from the literature. *Anaesthesia*, 70(4):421–428. (page 4)

[Huang et al., 2020] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J. (2020). Unet 3+: A full-scale connected unet for medical image segmentation. (pages 18 and 59)

[Ibtehaz and Rahman, 2020a] Ibtehaz, N. and Rahman, M. S. (2020a). Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87. (page 19)

[Ibtehaz and Rahman, 2020b] Ibtehaz, N. and Rahman, M. S. (2020b). Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87. (page 59)

[Ivanovs et al., 2021] Ivanovs, M., Kadikis, R., and Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234. (page 21)

[Jafari et al., 2020] Jafari, M., Auer, D., Francis, S., Garibaldi, J., and Chen, X. (2020). Dru-net: An efficient deep convolutional neural network for medical image segmentation. (pages 19 and 59)

[Jahwar and Abdulazeez, 2022] Jahwar, A. F. and Abdulazeez, A. M. (2022). Segmentation and classification for breast cancer ultrasound images using deep learning techniques: A review. In *2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA)*, pages 225–230. IEEE. (page 2)

[Jain et al., 2020] Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S. C., Mujumdar, S., Afzal, S., Mittal, R. S., and Munigala, V. (2020). Overview and importance of data quality for machine learning tasks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. (page 8)

[Jiang et al., 2021] Jiang, P. T., Zhang, C. B., Hou, Q., Cheng, M. M., and Wei, Y. (2021). Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888. (pages 11, 22, 81, and 82)

[Jimenez et al., 2018] Jimenez, D. A., García, H. F., Álvarez, A. M., and Orozco, Á. A. (2018). 3d probabilistic morphable models for brain tumor segmentation. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 22nd Iberoamerican Congress, CIARP 2017, Valparaíso, Chile, November 7–10, 2017, Proceedings 22*, pages 314–322. Springer. (page 5)

[Jimenez-Castaño et al., 2021] Jimenez-Castaño, C. A., Álvarez Meza, A. M., Aguirre-Ospina, O. D., Cárdenas-Peña, D. A., and Álvaro Angel Orozco-Gutiérrez (2021). Random fourier features-based deep learning improvement with class activation interpretability for nerve structure segmentation. *Sensors 2021, Vol. 21, Page 7741*, 21:7741. (pages 5 and 15)

[Joensuu et al., 2022] Joensuu, J., Saarijärvi, H., Rouhe, H., Gissler, M., Ulander, V.-M., Heinonen, S., Torkki, P., and Mikkola, T. (2022). Maternal childbirth experience and pain relief methods: a retrospective 7-year cohort study of 85 488 parturients in finland. *BMJ Open*, 12(5). (page 3)

[Jonaitytė and Petkevičius, 2021] Jonaitytė, I. and Petkevičius, L. (2021). Analysis of information compression of medical images for survival models. *2021 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, pages 1–6. (pages 7 and 8)

[Jung et al., 2021] Jung, Y.-J., Han, S.-H., and Choi, H.-J. (2021). Explaining cnn and rnn using selective layer-wise relevance propagation. *IEEE Access*, 9:18670–18681. (page 22)

[Kang, 2020] Kang, S. (2020). Rotation-invariant wafer map pattern classification with convolutional neural networks. *IEEE Access*, 8:170650–170658. (page 17)

[Khan et al., 2021] Khan, M. Z., Gajendran, M. K., Lee, Y., and Khan, M. A. (2021). Deep neural architectures for medical image semantic segmentation: Review. *IEEE ACCESS*, 9:83002–83024. (page 1)

[Khan et al., 2022] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41. (page 13)

[Kim et al., 2020] Kim, J., Jung, W., Kim, H., and Lee, J. (2020). Cycnn: A rotation invariant cnn using polar mapping and cylindrical convolution layers. *arXiv preprint arXiv:2007.10588*. (page 17)

[Kim and Byun, 2020] Kim, M. and Byun, H. (2020). Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (page 17)

[Kirillov et al., 2023] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment anything. (page 13)

[Kolyshkina and Simoff, 2021] Kolyshkina, I. and Simoff, S. (2021). Interpretability of machine learning solutions in public healthcare: The crisp-ml approach. *Frontiers in big Data*, 4:660206. (page 10)

[Koszewicz et al., 2021] Koszewicz, M., Szydlo, M., Gosk, J., Wieczorek, M., Slotwinski, K., and Budrewicz, S. (2021). The use of quantitative sensation testing to identify the physiological differences between the median and ulnar nerves. *Frontiers in Human Neuroscience*, 15:601322. (page 4)

[Koyucu and Karaca, 2022] Koyucu, R. G. and Karaca, P. P. (2022). Effects of low back pain during the first stage of labor on maternal birth satisfaction: A cross sectional study. *Acıbadem Üniversitesi Sağlık Bilimleri Dergisi*, 13(2):247–255. (page 3)

[Kulathunga et al., 2020] Kulathunga, N., Ranasinghe, N. R., Vrinceanu, D., Kinsman, Z., Huang, L., and Wang, Y. (2020). Effects of the nonlinearity in activation functions on the performance of deep learning models. (page 10)

[Kumar, 2021] Kumar, D. K. N. (2021). Survey of machine learning applications of convolutional neural networks to medical image analysis. *International Journal for Research in Applied Science and Engineering Technology*. (page 7)

[Kumar et al., 2018] Kumar, V., Webb, J. M., Gregory, A., Denis, M., Meixner, D. D., Bayat, M., Whaley, D. H., Fatemi, M., and Alizad, A. (2018). Automated and real-time segmentation of suspicious breast masses using convolutional neural network. *PloS one*, 13(5):e0195816. (page 12)

[Kutateladze, 2022] Kutateladze, V. (2022). The kernel trick for nonlinear factor modeling. *International Journal of Forecasting*, 38(1):165–177. (page 39)

[Kütük and Algan, 2022] Kütük, Z. and Algan, G. (2022). Semantic segmentation for thermal images: A comparative survey. (page 9)

[LaLonde et al., 2021] LaLonde, R., Xu, Z., Irmakci, I., Jain, S., and Bagci, U. (2021). Capsules for biomedical image segmentation. *Medical image analysis*, 68:101889. (page 17)

[Lee et al., 2022] Lee, C.-H., Lin, M.-H., Lin, Y.-T., Hsu, C.-C., Lin, C.-H., Chen, S.-H., and Huang, R.-W. (2022). Comparison of the effectiveness of local anesthesia for the digital block between single-volar subcutaneous and double-dorsal finger injections: a systematic review and meta-analysis of randomized control trials. *Journal of Plastic Surgery and Hand Surgery*, pages 1–14. (page 4)

[Li et al., 2021a] Li, J., Zhu, G., Hua, C., Feng, M., Li, P., Lu, X., Song, J., Shen, P., Xu, X., Mei, L., et al. (2021a). A systematic collection of medical image datasets for deep learning. *arXiv preprint arXiv:2106.12864*. (pages 2, 3, 8, and 9)

[Li et al., 2021b] Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., and Goh, R. (2021b). Medical image segmentation using squeeze-and-expansion transformers. *IJCAI International Joint Conference on Artificial Intelligence*, pages 807–815. (page 13)

[Lilay and Taye, 2023] Lilay, M. Y. and Taye, G. D. (2023). Semantic segmentation model for land cover classification from satellite images in gambella national park, ethiopia. *SN Applied Sciences*, 5(3):76. (page 2)

[Lin et al., 2021] Lin, D., Li, Y., Prasad, S., Nwe, T. L., Dong, S., and Oo, Z. M. (2021). Cam-guided u-net with adversarial regularization for defect segmentation. In *2021 IEEE International Conference on Image Processing* (ICIP), pages 1054–1058. (page 97)

[Linardatos et al., 2020] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18. (pages 10, 20, and 22)

[Liu et al., 2021] Liu, F., Huang, X., Chen, Y., and Suykens, J. A. K. (2021). Random features for kernel approximation: A survey on algorithms, theory, and beyond. (pages 14 and 38)

[Liu et al., 2020] Liu, M., Li, F., Yan, H., Wang, K., Ma, Y., Shen, L., and Xu, M. (2020). A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease. *NeuroImage*, 208:116459. (page 2)

[Loizidou et al., 2022] Loizidou, K., Skouroumouni, G., Nikolaou, C., and Pitris, C. (2022). A review of computer-aided breast cancer diagnosis using sequential mammograms. *Tomography*, 8(6):2874–2892. (page 7)

[Long et al., 2014] Long, J., Shelhamer, E., and Darrell, T. (2014). Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038. (pages 12, 18, 59, and 61)

[Lundberg and Lee, 2017] Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. (page 21)

[Luo et al., 2022] Luo, X., Hu, M., Song, T., Wang, G., and Zhang, S. (2022). Semi-supervised medical image segmentation via cross teaching between cnn and transformer. pages 820–833. (page 13)

[Lv et al., 2020] Lv, F., Liang, T., Chen, X., and Lin, G. (2020). Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. (page 14)

[Maier-Hein et al., 2018] Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O., Maier-Hein, K. H., Menze, B. H., Müller, H., Neher, P. F., Niessen, W. J., Rajpoot, N. M., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., van der Sommen, F., Wang, C., Weber, M., Zheng, G., Jannin, P., and Kopp-Schneider, A. (2018). Is the winner really the best? A critical analysis of common research practice in biomedical image analysis competitions. *CoRR*, abs/1806.02051. (page 2)

[Mairal et al., 2014] Mairal, J., Koniusz, P., Harchaoui, Z., and Schmid, C. (2014). Convolutional kernel networks. (page 15)

[Mairal and Vert, 2018] Mairal, J. and Vert, J.-P. (2018). Machine learning with kernel methods. *Lecture Notes, January*, 10:52. (pages 39 and 40)

[Maldonado et al., 2020] Maldonado, H., Bayareh, R., Torres, I., Vera, A., Gutiérrez, J., and Leija, L. (2020). Automatic detection of risk zones in diabetic foot soles by processing thermographic images taken in an uncontrolled environment. *Infrared Physics & Technology*, 105:103187. (page 9)

[Markus et al., 2021] Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655. (page 7)

[McCombe and Bogod, 2021] McCombe, K. and Bogod, D. (2021). Regional anaesthesia: risk, consent and complications. *Anaesthesia*, 76(S1):18–26. (page 4)

[Mejia-Zuluaga et al., 2022] Mejia-Zuluaga, R., Aguirre-Arango, J. C., Collazos-Huertas, D., Daza-Castillo, J., Valencia-Marulanda, N., Calderón-Marulanda, M., Aguirre-Ospina, Ó., Alvarez-Meza, A., and Castellanos-Dominguez, G. (2022). Deep learning semantic segmentation of feet using infrared thermal images. In Bicharra Garcia, A. C., Ferro, M., and Rodríguez Ribón, J. C., editors, *Advances in Artificial Intelligence – IBERAMIA 2022*, pages 342–352, Cham. Springer International Publishing. (page 32)

[Melesse et al., 2022] Melesse, A. S., Bayable, S. D., Simegn, G. D., Ashebir, Y. G., Ayenew, N. T., and Fetene, M. B. (2022). Survey on knowledge, attitude and practice of labor analgesia among health care providers at debre markos comprehensive specialized hospital, ethiopia 2021. a cross-sectional study. *ANNALS OF MEDICINE AND SURGERY*, 74. (pages 3, 5, and 8)

[Miller and Reich, 2022] Miller, M. J. and Reich, B. J. (2022). Bayesian spatial modeling using random fourier frequencies. *Spatial Statistics*, 48:100598. (page 97)

[Minaee et al., 2022] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. (2022). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542. (page 2)

[Ministerio de Salud y Protección Social de Colombia, 2023] Ministerio de Salud y Protección Social de Colombia (2023). Salud materna &nbsp; //. (page 3)

[Mo et al., 2022] Mo, Y., Wu, Y., Yang, X., Liu, F., and Liao, Y. (2022). Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646. (pages 1, 57, and 59)

[Mohammadnia-Qaraei et al., 2018] Mohammadnia-Qaraei, M. R., Monsefi, R., and Ghiasi-Shirazi, K. (2018). Convolutional kernel networks based on a convex combination of cosine kernels. *Pattern Recognition Letters*, 116:127–134. (page 15)

[Mohankumar et al., 2020] Mohankumar, A. K., Nema, P., Narasimhan, S., Khapra, M. M., Srinivasan, B. V., and Ravindran, B. (2020). Towards transparent and explainable attention models. (page 20)

[Morrow et al., 2017] Morrow, A., Shankar, V., Petersohn, D., Joseph, A., Recht, B., and Yosef, N. (2017). Convolutional kitchen sinks for transcription factor binding site prediction. (page 14)

[Mubashar et al., 2022] Mubashar, M., Ali, H., Grönlund, C., and Azmat, S. (2022). R2u++: a multiscale recurrent residual u-net with dense skip connections for medical image segmentation. *Neural Computing and Applications*, 34(20):17723–17739. (page 18)

[Omer Ibrahim Abdalla et al., 2022] Omer Ibrahim Abdalla, E., Nahid, S., Shastham Valappil, S., Gudavalli, S., Sellami, S., Korichi, N., Ahmad, S., Vicente Canizares Cespedes, V., and Gopalakrishnan, S. (2022). Impact of covid-19 status on patients receiving neuraxial analgesia during labor: A national retrospective-controlled study. *Qatar Medical Journal*, 2022(3):30. (page 3)

[Pan et al., 2022] Pan, C., Wang, J., Chai, W., Kakillioglu, B., El Masri, Y., Panagoulia, E., Bayomi, N., Chen, K., Fernandez, J. E., Rakha, T., and Velipasalar, S. (2022). Capsule network-based semantic segmentation model for thermal anomaly identification on building envelopes. *Advanced Engineering Informatics*, 54:101767. (page 17)

[Parise et al., 2021] Parise, D. C., Gilman, C., Petrilli, M. A., and Malaspina, D. (2021). Childbirth pain and post-partum depression: Does labor epidural analgesia decrease this risk? *Journal of Pain Research*, pages 1925–1933. (page 3)

[Parkhi et al., 2012] Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. (2012). Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*. (page 32)

[Peng et al., 2021] Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N. A., Kong, L., and Allen, P. G. (2021). Random feature attention. (page 15)

[Quazi and Musa, 2021] Quazi, S. and Musa, S. M. (2021). Image classification and semantic segmentation with deep learning. *2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering* (ICRAIE), 6:1–6. (page 2)

[Qureshi et al., 2023] Qureshi, I., Yan, J., Abbas, Q., Shaheed, K., Riaz, A. B., Wahid, A., Khan, M. W. J., and Szczuko, P. (2023). Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90:316–352. (pages 1 and 9)

[Rahimi and Recht, 2009] Rahimi, A. and Recht, B. (2009). Random features for large-scale kernel machines. (pages 14, 24, 37, and 40)

[Rashed and Popescu, 2022] Rashed, B. M. and Popescu, N. (2022). Critical analysis of the current medical image-based processing techniques for automatic disease evaluation: Systematic literature review. *Sensors*, 22(18):7065. (page 7)

[Ren et al., 2016] Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. (page 12)

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”why should i trust you?”: Explaining the predictions of any classifier. (page 21)

[Rizzoli et al., 2022] Rizzoli, G., Barbato, F., and Zanuttigh, P. (2022). Multimodal semantic segmentation in autonomous driving: A review of current approaches and future perspectives. *Technologies*, 10(4). (page 1)

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597. (pages 12, 18, 59, and 61)

[Roy et al., 2022] Roy, S., Meena, T., and Lim, S.-J. (2022). Demystifying supervised learning in healthcare 4.0: A new reality of transforming diagnostic medicine. *Diagnostics*, 12(10):2549. (page 7)

[Rudin, 1976] Rudin, W. .-. (1976). *Fourier analysis on groups*. Interscience tracts in pure and applied mathematics ;. Interscience, New York, New York. (page 41)

[Sacha et al., 2023] Sacha, M., Rymarczyk, D., Łukasz Struski, Tabor, J., and Zieliński, B. (2023). Protoseg: Interpretable semantic segmentation with prototypical parts. (page 23)

[Salahuddin et al., 2022] Salahuddin, Z., Woodruff, H. C., Chatterjee, A., and Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140:105111. (page 11)

[Samek et al., 2021] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278. (page 21)

[Santos and Papa, 2022] Santos, C. F. G. D. and Papa, J. P. (2022). Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Computing Surveys (CSUR)*, 54(10s):1–25. (page 8)

[Sarker, 2021] Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6):420. (pages 7 and 8)

[Scalco and Rizzo, 2017] Scalco, E. and Rizzo, G. (2017). Texture analysis of medical images for radiotherapy applications. *The British journal of radiology*, 90(1070):20160642. (page 7)

[Schorr et al., 2021] Schorr, C., Goodarzi, P., Chen, F., and Dahmen, T. (2021). Neuroscope: An explainable ai toolbox for semantic segmentation and image classification of convolutional neural nets. *Applied Sciences 2021, Vol. 11, Page 2199*, 11:2199. (page 23)

[Selvaraju et al., 2016] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2016). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359. (pages 11, 22, and 81)

[Shorten and Khoshgoftaar, 2019] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48. (page 14)

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. (page 61)

[Singh et al., 2020] Singh, A., Sengupta, S., and Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. (page 10)

[Singh et al., 2021] Singh, P., Verma, A., and Alex, J. S. R. (2021). Disease and pest infection detection in coconut tree through deep learning techniques. *Computers and Electronics in Agriculture*, 182:105986. (page 2)

[Smith et al., 2018] Smith, C. A., Levett, K. M., Collins, C. T., Armour, M., Dahlen, H. G., and Suganuma, M. (2018). Relaxation techniques for pain management in labour. *Cochrane Database of Systematic Reviews*, (3). (page 3)

[Soomro et al., 2023] Soomro, T. A., Zheng, L., Afifi, A. J., Ali, A., Soomro, S., Yin, M., and Gao, J. (2023). Image segmentation for mr brain tumor detection using machine learning: A review. *IEEE Reviews in Biomedical Engineering*, 16:70–90. (page 1)

[Speith, 2022] Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (xai) methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250. (page 20)

[Stergiou, 2021] Stergiou, A. (2021). The mind's eye: Visualizing class-agnostic features of cnns. (page 21)

[Stevens et al., 2006] Stevens, M. F., Werdehausen, R., Hermanns, H., and Lipfert, P. (2006). Skin temperature during regional anesthesia of the lower extremity. *Anesthesia and analgesia*, 102(4):1247—1251. (pages 5 and 8)

[Sun et al., 2020] Sun, J., Darbehani, F., Zaidi, M., and Wang, B. (2020). Saunet: Shape attentive u-net for interpretable medical image segmentation. (page 23)

[Taghanaki et al., 2021] Taghanaki, S. A., Abhishek, K., Cohen, J. P., Cohen-Adad, J., and Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54:137–178. Argumentan el uso de arquiteturas encoder-decoder. (page 24)

[Tancik et al., 2020] Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., and Ng, R. (2020). Fourier features let networks learn high frequency functions in low dimensional domains. (page 14)

[Teng et al., 2022] Teng, Q., Liu, Z., Song, Y., Han, K., and Lu, Y. (2022). A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, pages 1–21. (pages 10, 20, 21, and 22)

[Tsai et al., 2023] Tsai, J., Chang, C.-C., and Li, T. (2023). Autonomous driving control based on the technique of semantic segmentation. *Sensors*, 23(2). (page 1)

[Tsuneki, 2022] Tsuneki, M. (2022). Deep learning models in medical image analysis. *Journal of Oral Biosciences*, 64(3):312–320. (page 7)

[Ventura et al., 2023] Ventura, F., Greco, S., Apiletti, D., and Cerquitelli, T. (2023). Explaining deep convolutional models by measuring the influence of interpretable features in image classification. *Data Mining and Knowledge Discovery*, pages 1–58. (page 22)

[Vinogradova et al., 2020] Vinogradova, K., Dibrov, A., and Myers, G. (2020). Towards interpretable semantic segmentation via gradient-weighted class activation mapping. (pages 22 and 82)

[Wang et al., 2019] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2019). Score-cam: Score-weighted visual explanations for convolutional neural networks. (pages 11, 22, and 81)

[Wang et al., 2022a] Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., and Nandi, A. K. (2022a). Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5):1243–1267. (page 11)

[Wang and Li, 2023] Wang, S. and Li, R. (2023). Semantic-guided disentangled representation for unsupervised cross-modality medical image segmentation. (page 17)

[Wang et al., 2022b] Wang, T., Dong, B., Zhang, K., Li, J., and Xu, L. (2022b). Slim-rffnet: Slim deep convolution random fourier feature network for image classification. *Knowledge-Based Systems*, 237:107878. (page 15)

[Wang et al., 2021] Wang, X., Wang, L., Zhong, X., Bai, C., Huang, X., Zhao, R., and Xia, M. (2021). Pai-net: A modified u-net of reducing semantic gap for surgical instrument segmentation. *IET Image Processing*, 15:2959–2969. (pages 15, 18, and 59)

[Werdehausen et al., 2007] Werdehausen, R., Braun, S., Hermanns, H., Freynhagen, R., Lipfert, P., and Stevens, M. F. (2007). Uniform distribution of skin-temperature increase after different regional-anesthesia techniques of the lower extremity. *Regional Anesthesia and Pain Medicine*, 32(1):73–78. (pages 5 and 8)

[Whitburn and Jones, 2020] Whitburn, L. Y. and Jones, L. E. (2020). Looking for Meaning in Labour Pain: Are Current Pain Measurement Tools Adequate? *Pain Medicine*, 22(5):1023–1028. (page 5)

[Whittingham, 2013] Whittingham, T. (2013). Imaging and imagining the fetus – the development of obstetric ultrasound. *Ultrasound*, 21:235 – 235. (page 4)

[Wieland et al., 2023] Wieland, M., Martinis, S., Kiefl, R., and Gstaiger, V. (2023). Semantic segmentation of water bodies in very high-resolution satellite and aerial images. *Remote Sensing of Environment*, 287:113452. (page 2)

[Willemink et al., 2020] Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., and Lungren, M. P. (2020). Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15. (pages 5 and 8)

[Williams et al., 2022] Williams, E., Niehaus, S., Reinelt, J., Merola, A., Mihai, P. G., Villringer, K., Thierbach, K., Medawar, E., Lichterfeld, D., Roeder, I., Scherf, N., and del C. Valdés Hernández, M. (2022). Automatic quality control framework for more reliable integration of machine learning-based image segmentation into medical workflows. (page 9)

[World Health Organization, 2023] World Health Organization (2023). Maternal health. Accessed on May 4, 2023. (page 2)

[Wu and Tang, 2021] Wu, J. and Tang, X. (2021). Brain segmentation based on multi-atlas and diffeomorphism guided 3d fully convolutional network ensembles. *Pattern Recognition*, 115:107904. (page 2)

[Xiao et al., 2017a] Xiao, H., Rasul, K., and Vollgraf, R. (2017a). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. (pages 30 and 31)

[Xiao et al., 2017b] Xiao, H., Rasul, K., and Vollgraf, R. (2017b). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747. (page 48)

[Xie et al., 2019] Xie, J., Liu, F., Wang, K., and Huang, X. (2019). Deep kernel learning via random fourier features. (page 14)

[Xu et al., 2023] Xu, Q., Xie, W., Liao, B., Hu, C., Qin, L., Yang, Z., Xiong, H., Lyu, Y., Zhou, Y., Luo, A., et al. (2023). Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review. *Journal of Healthcare Engineering*, 2023. (page 10)

[Yang and Yu, 2021] Yang, R. and Yu, Y. (2021). Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in Oncology*, 11(1):573. (pages 7 and 8)

[Yeung et al., 2022] Yeung, M., Sala, E., Schönlieb, C.-B., and Rundo, L. (2022). Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026. (page 97)

[Yıldırım et al., 2023] Yıldırım, G., Karakaş, H. M., Özkaya, Y. A., Şener, E., Fındık, Ö., and Pulat, G. N. (2023). Development of an artificial intelligence method to detect covid-19 pneumonia in computed tomography images. *Istanbul Medical Journal*, 24(1). (page 2)

[Zhang and Dong, 2019] Zhang, H. and Dong, B. (2019). A review on deep learning in medical image reconstruction. *Journal of the Operations Research Society of China*, 8:311–340. (page 7)

[Zhang et al., 2022] Zhang, L., Nan, Q., Bian, S., Liu, T., and Xu, Z. (2022). Real-time segmentation method of billet infrared image based on multi-scale feature fusion. *Scientific Reports*, 12(1):6879. (page 9)

[Zhang et al., 2020] Zhang, X., Xu, J., Yang, J., Chen, L., Zhou, H., Liu, X., Li, H., Lin, T., and Ying, Y. (2020). Understanding the learning mechanism of convolutional neural networks in spectral analysis. *Analytica Chimica Acta*, 1119:41–51. (page 97)

[Zhang et al., 2021a] Zhang, Y., Liu, H., and Hu, Q. (2021a). Transfuse: Fusing transformers and cnns for medical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12901 LNCS:14–24. (page 13)

[Zhang et al., 2021b] Zhang, Y., Tino, P., Leonardis, A., and Tang, K. (2021b). A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742. (page 11)

[Zhao et al., 2017] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. (page 12)

[Zheng et al., 2022] Zheng, Q., Wang, Z., Zhou, J., and Lu, J. (2022). Shap-cam: Visual explanations for convolutional neural networks based on shapley value. (page 22)

[Zhou et al., 2015a] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015a). Learning deep features for discriminative localization. (pages 11, 22, and 81)

[Zhou et al., 2015b] Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. (2015b). Learning deep features for discriminative localization. *CoRR*, abs/1512.04150. (page 80)

[Zhou et al., 2018a] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018a). Unet++: A nested u-net architecture for medical image segmentation. *CoRR*, abs/1807.10165. (page 12)

[Zhou et al., 2018b] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018b). Unet++: A nested u-net architecture for medical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11045 LNCS:3–11. (page 59)

[Zhou et al., 2020] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2020). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. (pages 18 and 59)

[Zou et al., 2021] Zou, K., Chen, X., Wang, Y., Zhang, C., and Zhang, F. (2021). A modified u-net with a specific data argumentation method for semantic segmentation of weed images in the field. *Computers and Electronics in Agriculture*, 187:106242. (page 2)

[Zunair and Hamza, 2021] Zunair, H. and Hamza, A. B. (2021). Sharp u-net: Depthwise convolutional network for biomedical image segmentation. *Computers in Biology and Medicine*, 136:104699. (pages 18 and 59)